

Essays in Analytics and Econometrics

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Amil Jamal Williamson

December 2019

© by Amil Jamal Williamson, 2019
All Rights Reserved.

Dedication

To my mother...thank you.

Acknowledgments

I would firstly like to express my sincerest form of gratitude to my dissertation committee, starting first and foremost, with my primary dissertation advisor: Dr. Russell Zaretzki. Thank you for granting me the freedom of exploring my research interests with no restrictions as well as the unwavering encouragement, your intelligence, and all assistance you offered. A word of this dissertation effort could not have been written without your guidance, and I thank you deeply for that. Dr. Mary Leitnaker, in addition to being consistently inspired by your purview in statistical theory, I would like to thank you for being an unwavering source of encouragement and personal interactions that have made my tenure in this city endurable. To you (and Dr. Kelch), there are not enough words for me to truly convey the limitless appreciation I have for you both: thank you. Dr. Haileab Hilafu, to be in your presence for the short duration for which we have been acquainted has been inspirational in every, conceivable way. I have enjoyed every moment learning from you and getting a chance to know you. Dr. Luiz Lima, I cannot express how lucky I feel to have had you join my dissertation committee and to be a student in two of the greatest and dynamic courses I have had in my entire academic tenure. I wish I could have taken more. Thank you.

To my current classmates, I wish you two nothing but the absolute best in your future careers. Murat Arat and Wangcheng Yan, it has been a privilege to learn and evolve with you two. It has been a journey, but a journey I would not have preferred taking without you both. Thank you both for sharing your thoughts, experiences, and cultures with me.

I would like to thank the entire Business Analytics & Statistics department for funding my academic journey as well as offering assistance over the years with everything. To Dr. Sean Willems, for everything said and unsaid, I thank you. To Drs. Noon, Srinivasan, and Gilbert, how I also thank you all. Mrs. Melissa Morris, thank you for everything, especially the much needed laughter and levity. Dr. Bogdan Bichescu, the intelligence and grace that you exhibit are attributes I could only hope to possess one day. Drs. William Seaver, Robert Mee, Wenjun Zhou, Michel Ballings, G. Todorova, M. Frazier, and countless others, thank you for imbuing me with the academic knowledge making this effort possible as well. Finally, to the guy who explains a p-value better than anyone else in the world, Mr. Charles Cwiek, thank you very much for it all. How I hope the department continues to thrive with integrity and excellence in the future with Dr. Galbreth leading the path.

To all staff the Haslam College of Business, with a special consideration to my undergraduate advisor, Mr. Mark Willoughby, thank you.

Lastly, to all students I have had the pleasure of teaching throughout my tenure at this university, it has been and will **always** be the greatest achievement and privilege of my adult life. I learned from and with you all.

Abstract

This dissertation spans three distinct methodological contributions to the econometric and analytic modeling literature. The first chapter is motivated by the challenge of estimating the impact of a customized multi-treatment coupon campaigns on visit rates for a large panel across a two-year trial. Prior literature has highlighted the necessity for both detecting and mitigating the issue of endogeneity. In response, I first develop a novel robust Wald test for idiosyncratic endogeneity in the presence of individual heterogeneity and a time-varying endogenous regressor of restricted range for a nonlinear, unobserved effects model. Building on a prior working paper, I extend the results to tackle the identification of idiosyncratic endogeneity in the presence of time-constant endogeneity for unbalanced panels. I propose a two-stage estimation procedure that tests for positive covariance between time-varying unobservables and a time-varying, binary endogenous variable that completely controls for the latent, time-constant heterogeneity for count responses. Simulations suggest endogeneity detection for both unbalanced panels scenarios mirrors the balanced panel benchmark closely using metrics of rejection rates, nominal size, and statistical power given proper instrumental variables. A nonlinear, instrumental variables GMM procedure is proposed for parameter estimation given positive endogeneity detection. The second chapter studies the estimation of a nonlinear (generalized additive) mixed effects models for count data containing parametric factor smooths. These models are further extended to estimate Poisson responses generated from both stationary and non-stationary conditionally serially correlated AR (1) processes. Simulation studies are used to verify the accuracy of these models in a range of scenarios. The models are then applied to the problem of estimating a carryover effects of a promotional campaigns after promotions end. The final chapter seeks to extend the forecasting combination model of Granger and Ramanathan by approximating the optimal conditional mean solution with conditional quantiles. Useless forecasters at arbitrary quantiles of the response are discarded via a LASSO penalty. The final combination forecast outperforms the equal weight combination benchmark as well as a variety of other combination models in terms of out-of-sample predictive power and other significant performance metrics at the expense of marginally increased forecast variance.

Table of Contents

1	Introduction	1
1.1	Background and Motivation	2
1.2	Understanding the Impact of Loyalty Promotions on Customer Visits	3
1.2.1	A Review of Key Literature	3
1.2.2	Application I: Promotional Loyalty Marketing Program Selection-Based Endogeneity For Unbalanced Panel - Quasi-Experiment	4
1.3	Estimating Generalized Additive Mixed Models	7
1.3.1	Application II: Carryover effect estimation into the First Intermittent Cycle of a Campaign Exposures	8
1.4	Extending the Granger-Ramanathan Optimal Solution via a Penalized Approach	8
1.4.1	Application III: Forecasting Combination Problem of Monthly Annualized Inflation Gap with Penalization	10
	References	11
2	Quantifying the Exposure Effects of Customized and Targeting Promotions on Shopping Frequency in the Presence of Dual Sources of Endogeneity	13
	Abstract	14
2.1	Introduction	15
2.2	Endogeneity	21
2.2.1	Conventional Remedies for Endogeneity	23
2.3	Correlated Random Effects with a Control Function Approach - Test for Endogeneity in Nonlinear, Unbalanced Panels	25
2.3.1	Correlated Random Effects/Control Function Approach: Idiosyncratic Endogeneity Test	26
2.3.2	Idiosyncratic Test for Endogeneity - Asymptotics for the Conditional Mean Function .	33
2.3.3	Idiosyncratic Endogeneity Test: Form of Endogenous Regressor	34
2.3.4	Monte Carlo Simulation - Endogeneity Test: Data Generating Processes	37
2.3.5	Monte Carlo Simulation Results	40
2.4	Generalized Method of Moments: Construction of Three Feasibly Efficient Estimators	43
2.5	Application: Quantifying the Exposure Effects of Promotional Campaigns on In-Store Visits .	50
2.5.1	Data Description	50

2.5.2	Promotional Modeling Results: Endogeneity Test	55
2.5.3	IV-GMM Results	56
2.6	Discussion	60
	References	62
	Appendix	68
3	Estimating Conditionally Serially Correlated Generalized Additive Mixed Models with Parametric Factor Smooths for Repeated Counts	71
	Abstract	72
3.1	Introduction	73
3.2	Nonparametric Curve Fitting: A single covariate	76
3.2.1	Regression Splines	76
3.2.2	Penalized Splines	77
3.2.3	Thin Plate Regression Splines	78
3.3	Generalized Additive Mixed Model	80
3.3.1	Generalized Additive Mixed Model formulation with Parametric Factor Smooths . . .	81
3.3.2	Recasting GAMM to Penalized GLMM with parametric factor smooth	83
3.4	Penalized Quasi-Likelihood (PQL) Estimation for Re-Casted GAMM	86
3.4.1	Temporal Serial Correlation Structure	90
3.5	Conditionally Correlated AR(1) Models for Stationary and Nonstationary Count Data with random effects - Simulation Designs	91
3.5.1	Performance of Parameters	94
3.5.2	Simulation Results: Design I.	95
3.5.3	Simulation Results: Design II	96
3.6	Application: Household weekly visits during non-promotional period 2	99
3.6.1	Aggregate Descriptive Statistics in Periods of Non-Promotion	100
3.6.2	R Package: mgcv	102
3.6.3	Non-Promotional Period 2 (NP2) GAMMs	102
3.7	Discussion	111
	References	113
	Appendix	117
4	Forecasting Out-of-Sample Inflation Gaps: A Penalized Quantile Combination Approach	119
	Abstract	120
4.1	Introduction	121
4.2	The Forecast Combination Problem	124

4.3	Estimation	127
4.3.1	The L1-Penalized Quantile Forecast Combination Method	130
4.3.2	L1PQFC Forecast: Example	131
4.3.3	Combination Model Comparisons	132
4.4	The Forecasting Procedures: Forecasting the Monthly Inflation Gap	133
4.4.1	Part I. Production of the out-of-sample (OOS) forecasts	134
4.4.2	Part II: Forecasting Combination Procedures	139
4.5	Forecasting Combination Results	140
4.5.1	Additional benefits of L1PQFC models	146
4.6	Discussion	154
	References	155
	Appendix	159
5	Conclusion	164
	References	169
	Vita	170

List of Tables

1.1	Campaign Type Descriptions	6
2.1	General Advantages and Disadvantages for FE, RE, and CRE models	19
2.2	Design 1(a) - Unbalanced Panel with Selection - Monte Carlo Simulation Results	42
2.3	Design 1(b) - Unbalanced Panel without Selection - Monte Carlo Simulation Results	42
2.4	Design 1(c) - Balanced Panel - Monte Carlo Simulation Results	42
2.5	Monte Carlo Results: Mean Coefficient and F-Statistics Estimates	42
2.6	Promotional Modeling: Response and Exogenous Variables: Descriptions	50
2.7	Promotional Modeling: Endogenous Variable and Instrumental Variables - Descriptions	55
2.8	Instrument Relevance Test	55
2.9	Two Stage CRE/Control Function Test for Idiosyncratic Endogeneity	57
2.10	Nonlinear-IV GMM - Promotional Period Estimation Results	59
2.11	First Stage LPM	68
3.1	Parameters of Simulation Designs	92
3.2	Simulation Results I: Balanced Stationary Panel	95
3.3	Stationary Estimated Smooth Functions - Performance (Case 1 - Stationary)	95
3.4	Simulation Results II: Balanced Non-Stationary Panel	98
3.5	Non-stationary Estimated Smooth Functions - Performance (Case 1)	98
3.6	Average Rate Difference and Standard Error of Rate Difference in Visits per Week	101
3.7	Poisson GAMM Results	105
3.8	Negative Binomial GAMM Results	108
3.9	NP2 Factor Smooth Interactions - Estimates and Limits	111
3.10	Simulation Results I - Unbalanced Stationary Panel	118
3.11	Unbalanced Stationary Estimated Smooth Functions - Performance (Case 1)	118
3.12	Simulation Results II - Unbalanced Non-Stationary Panel	118
3.13	Unbalanced Non-Stationary Estimated Smooth Functions - Performance (Case 1)	118
4.1	Out-of-Sample Accuracy Measure: Theil's U	139
4.2	OOS Combination Forecasting Diagnostics, 2014:M1-2017:M12, $h = \{1, 2, 3, 4, 8, 12\}$	145
4.3	Mean squared prediction error (MSPE) decomposition	146

4.4	Frequency of models selected over OOS: January 2014 to December 2017	149
4.5	Out of Sample Accuracy Metrics	162

List of Figures

2.1	Instrumental Variables Causal Graph	23
2.2	Average visits per day per period	52
2.3	Average visits per day (Single Campaign Effects)	53
2.4	Average visits per day (Joint Campaign Effects)	53
3.1	Factor Smooth Plots of Mean Estimates (Case 1)	97
3.2	Average Visits Per Week given Previous Campaign Type	103
3.3	NP2: Mean and Variance Visits Per Week	106
3.4	Raw Residuals vs. Normalized Residuals	109
3.5	Partial Dependence Factor Smooth Plots - NP2	110
4.1	U.S. Annualized Monthly Inflation Gap (1998:M1 - 2018:M8)	136
4.2	OOS Forecasts, $h=1,2,3,4,8,12$	138
4.3	Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=1$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.	141
4.4	Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=2$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.	141
4.5	Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=3$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.	142

4.6	Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=4$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.	142
4.7	Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=8$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.. . . .	143
4.8	Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=12$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.	143
4.9	Bivariate display of squared bias forecast and unconditional forecast variance relative to the equal weight benchmark for horizons $h = \{1, 2\}$. Each axis represents the difference between the forecast variance [squared bias] of the proposed combination and the forecast variance [squared bias] of the equal weight benchmark. A point for our combination consideration that exist on the blue diagonal represents a combination forecast with the same MSPE as the equal weight combination.	147
4.10	Bivariate display of squared bias forecast and unconditional forecast variance relative to the equal weight benchmark for horizons $h = \{3, 4\}$. Each axis represents the difference between the forecast variance [squared bias] of the proposed combination and the forecast variance [squared bias] of the equal weight benchmark. A point for our combination consideration that exist on the blue diagonal represents a combination forecast with the same MSPE as the equal weight combination.	147
4.11	Bivariate display of squared bias forecast and unconditional forecast variance relative to the equal weight benchmark for horizons $h = \{8, 12\}$. Each axis represents the difference between the forecast variance [squared bias] of the proposed combination and the forecast variance [squared bias] of the equal weight benchmark. A point for our combination consideration that exist on the blue diagonal represents a combination forecast with the same MSPE as the equal weight combination.	148
4.12	Models Selected: $h=1$	151
4.13	Models Selected: $h=2$	151
4.14	Models Selected: $h=3$	152

4.15 Models Selected: $h=4$	152
4.16 Models Selected: $h=8$	153
4.17 Models Selected: $h=12$	153

CHAPTER 1

Introduction

The purpose of this introduction is to provide a brief background associated with each chapter as these three chapters are quite distinct. This introductory component is not meant to replace or supplant the chapter-specific abstracts nor any of the content therein; it is simply to offer background and a brief, high-level overview of the all chapters' backgrounds and general objectives. Please note that throughout the dissertation, the notation used within each chapter is unique to that chapter and that notation may change meaning across chapters. Finally, the fifth chapter provides chapter-dependent limitations and future research questions.

1.1 Background and Motivation

Over the last 100 years, the need to develop relevant statistical models for both macroeconomics and microeconomics has been the primary motivator for the development of a large family of statistical methods known as econometrics ([1], [2]). Two specific features that have shaped this development and differentiate econometrics from other statistical methods are the frequent presence of data that evolves in time (i.e. time-series) and the need to analyze observational data (i.e. data that are not collected in experimental setting). This dissertation presents several methodological innovations useful for special cases of data with these features.

Time series data are common in economic studies and have a long history of interest in the study of macroeconomic dynamics ([3]). A basic time series data set is usually characterized by a moderate collection of univariate measurements taken over a regular time grid. Longitudinal data, or equivalently, panel data, are also an extremely common and often powerful tool for a variety of methodological and modeling techniques in the realm of statistics, econometrics, and data mining. Longitudinal data are formed from the union of the cross-section of subjects measured over time. They can be viewed as a group of time series, each measured on a separate subject in the sample and brought together to analyze subject or treatment level effects in a population. Longitudinal data sets will also often contain static data such as treatment indicators. In comparison to designs containing only cross-sections (e.g. a standard regression model) or only purely temporal observations (e.g. a standard time-series model), longitudinal data designs allow the study of the temporal dynamism while simultaneously being able to account for cross-sectional differences, or heterogeneity, among the entities ([4]).

In this dissertation, we will develop statistical methods that analyze both longitudinal and standard time series data in order to address applied statistical questions in micro and macroeconomics. In particular, Chapter 2 & 3 will focus on a microeconomic study of a quasi-experimental dataset generated by a grocery store loyalty program while Chapter 4 will develop and evaluate an improved forecasting method and apply this method to a time series data set based on the inflation gap.

1.2 Understanding the Impact of Loyalty Promotions on Customer Visits

Loyalty reward programs have been of significant interest for the grocery retail industry for many years. Current retail technologies require members to scan a loyalty key card, typically containing a barcode, at the time of checkout. In return for the inconvenience, customers usually receive instant savings including direct discounts, points or other rewards ([5]). From the retailers point of view, collection of this data allows for the microscopic assessment of customer-specific transaction data over time by differentiating between personalized incentives and general discounts. Industry pioneers who believe that loyalty cards are a primary knowledge management tool for customer transaction and/or behavioral shopping trends are incentivized in tailoring their reward system continuously given the increased competition from other chains ([6]). In terms of participants in the grocery retail space, recent studies have shown that approximately 62% of consumers belong to grocery reward programs. Of these consumers, 75% of the visits to their specific grocery stores where they participate in the reward loyalty programs which indicates a high-level of customer loyalty ([7]).

1.2.1 A Review of Key Literature

The first two chapters of the current dissertation are motivated by a proprietary data source licensed by Kroger’s 48.51 subsidiary (formerly dunnhumby USA) called the *Complete Journey* and made available through an online portal. This scanner-panel dataset contains the results of 2 years of scanned charges for 2,500 households, with 1,584 of the households participating in a quasi-experiment designed to test the impact of customized coupons on customer visit frequency and retail expenditures.

Several prior works used this data to analyze a variety of marketing and microeconomic questions. We review the two most relevant works to this effort here. This first work, [8], performed the seminal study of this data. These authors sought to understand the impact of customized coupon campaigns using three outcomes (trip incidence, trip revenue, and customized coupon redemption) while controlling for a large variety of demographic and behavioral variables. The authors devise a sophisticated Bayesian model to capture the outcomes of customer i in week t using a Type III Tobit model that handles the three simultaneous regression equations corresponding to the associated responses. The results indicate, unsurprisingly, that the mere exposure to a customized coupon set contributes positively to weekly trip revenue, trip incidence, and redemption of customized coupons. These authors find that the *exposure effect* of these customized coupon campaigns contributes more positively than actual coupon redemption to these correlated responses related to household behavior. The work finds that “customer surprise” generated by irregular distribution times of personalized campaigns as well as the larger impact of discounts offered per coupon sets are positively correlated with their effectiveness.

In a separate and complementary work, [9] estimates the differential impact of campaign types (both customized and targeted) received during the test period after accounting for endogeneity for customized coupon distribution. “Spillover effects” from personalized and targeted promotions were also quantified. The author creates control functions to help mitigate the endogeneity arising from targeted coupon distribution in a specific week t to predict the likelihood a household i has received at least one coupon during the testing time frame. Via the use of a Heckit model (two-step approach), this study claims to achieve consistent and asymptotically normal estimators of covariates in the final model, identifying that weekly sales contributes significantly to retailer revenue of household i in a week t .

Before delving into our specific research objectives and hypotheses, we note major differences that distinguish our approach from the two works just described. First, our response of interest is the number of visits in a period. Visit rate is a primary target for retailers; with a fixed budget, increased visits indicate that a retailer is winning a larger share of customers’ overall expenditures. In fact, we note that little to no existing research exists assessing the marginal impact of treatment effects on limited dependent variables such as in-store visits in an unbalanced panel format. The clear benefit of focusing on a single variable, discrete integer-valued visit counts, as a response, is simplicity. Other benefits for modeling the conditional mean of a discrete response include the availability of easy-to-apply parametric and semi-parametric models to measure impacts of the exposure effects, direct interpretations, and ease of reproducibility. Secondly, although the previous works do focus on exposure and redemption effects after treatment distribution of customized coupons, neither of the studies take into account the simultaneous receipt of campaigns and what the marginal impact of this simultaneity may actually be on their response variable(s) of interest during the promotional phase of the loyalty program. Perhaps the biggest oversight in these prior works involves the issue of endogeneity. While both works attempt to mitigate endogeneity of personalized rewards in both a Bayesian and strictly parametric frameworks, they fail to apply a formal, robust statistical test in order to empirically validate the claim that endogeneity is indeed present. Both authors rely on the pre-promotional spending behavior¹ as an indication of a higher likelihood of personalized rewards. However, without a formal test, such a modeling approach could result in an overly complex model for mitigating a verisimilitudinous claim of endogeneity for the personalized campaigns.

1.2.2 Application I: Promotional Loyalty Marketing Program Selection-Based Endogeneity For Unbalanced Panel - Quasi-Experiment

Within the quasi-experiment being analyzed, 30 separate campaigns are planned for distribution to a distinct subset of households. The campaigns are then mailed at varying intervals throughout a 16 month promotional test cycle and represent our principal covariates of interest. Each campaign has a respective start day and end day for which the product-specific coupons from distributed campaigns are redeemable. The span of

¹The pre-promotional spending behavior refers to weekly spending patterns prior to the distribution of the first promotional campaign. Equivalently, this refers to all tractable customer behavior in each respective household’s baseline periods.

days between the start and end days represents the *duration* of the campaigns effects. In many cases, after completion or ending of a campaign, an individual household, for example, may revert to a period of non-promotion within the test cycle while another household may immediately receive another promotion. We rely on quantitative tracking measures of individual-level household expenditures, items purchases, campaign durations and descriptions, and other measures to assist in our modeling approaches. The promotional loyalty program spans a two-year time period in which days 1-223 serve as the **control period** in which no campaigns are received by any household. The **test period** lasts from day 224-720. However, not all campaigns containing the promoted coupons begin at day 224. The **baseline period** will be a household-specific period in which no promotional campaigns are received prior to the first campaign received. Hence, every household’s baseline contains its respective control period.

Regarding the treatments of interest, we begin by defining the *exposure effect*. The exposure effect is the main treatment effect of interest in which households are exposed to one of three types of customized coupon campaigns via mail. The primary categories/levels of exposure are labeled Type A, B, and C. The Type A campaign contains a set of personalized coupons specifically tailored to fit the customers purchasing behavior based on an established, transactional buying profile. The Type B campaign denotes a set of coupons that were sent to households that offer discounts on products during specified time frames where items are typically bought in conjunction with one another. Lastly, Type C denotes those coupons sent to households that offered discounts on standard products. We note that, in line with prior work, the Type A campaign coincides with the *personalized* rewards (P) whereas Type B and Type C campaigns are *targeted* (T). We also note that there are no restrictions on the simultaneous receipt of campaigns in the promotional cycle of the study, so the treatment effect can also be defined jointly as the combination of the single campaign types (e.g. Type AB, Type ABC, Type AC, Type BC). Table 1.1 gives more detailed descriptions of the categories [10].

The dataset used for modeling is composed of totals or aggregates of all time-varying, numeric, household-level features for both our response of interest (in-store visits) and relevant explanatory covariates (e.g. expenditures)² Additionally, we capture time-invariant covariates per each row-level aggregate. These household level totals are aggregated for each campaign exposure’s period as well as any periods representing intermittent cycles of non-promotion within the test cycle such that each household has a multi-period vector of in-store visit counts and associated total expenditures.

For completeness, it is necessary to explicitly define the type of study that frames the data (and loyalty promotional program) used in the majority of this thesis. The nature of this particular customer promotional loyalty program in terms of our research must be treated as a quasi-experimental design. A quasi-experimental design tests: “descriptive causal hypotheses about manipulable causes as well as many structural details, such as the frequent presence of control groups and pretest measures, which support counterfactual inferences about what could transpire in the absence of a treatment which lack

²Further details are provided in Chapter 2.

Table 1.1: Campaign Type Descriptions

Campaign Type	Campaign Definition	Campaign Customization Level
Type A	A personalized set of ~14–16 coupons chosen for a household with a tag: ‘Thanks for being such a great customer!’ ‘Please enjoy the savings reserved only for premier customers like you.’ Contains a combination of ‘\$Y off Brand X’ and/or ‘\$Y off Category X when spending \$Z’ coupons	High
Type B	A themed pamphlet stating: ‘Wellness’, ‘Back to School’ (etc.) plus ~10 coupons for products related to the theme (e.g., pencils and notebooks, health products and juice, etc.).	Medium
Type C	A single manufacturer sponsored booklet with ~1–6 coupons and a free sample.	Low

random assignment” ([11]). In essence, the quasi-experimental nature of a study design presents the scenario in which treatments (or covariates of interest) are consciously allocated without randomization to subjects for an intended outcome. In quasi-experimental designs, the extent to which one can reliably estimate causal effects of interest is the necessity to remain conscious of internal validity³. Such threats to internal validity include biases resulting from *maturation*, *experimental mortality* (attrition), *instrumentation* (change in measurement method on the outcome), *history* (current event forces change in outcomes, *design contamination* (cross-sectional unit intervention which can degrade the outcome), and *selection* (cross-sectional units’ differences that are unmeasured between treatment and control groups that are related to the outcome) ([12]). The goal of employing quasi-experimental designs is to approximate randomized controlled designs as closely as possible wherein empirical researchers presume that the primary distinction between subjects for those who are exposed to a treatment versus those who are not is simply the exposure ([13]). The issue of non-randomization of treatment exposures in this quasi-experiment is not insurmountable as studies have been successfully analyzed by controlling for the variation of the treatment variable of interest that has been contaminated by endogeneity ([14]). In certain cases, more advanced econometric techniques primed for assessing causal effects in quasi-experiments can approximate randomization by mitigating the sources of endogeneity for consistent parameter estimation ([15]). From an econometrics modeling perspective, the most popular method for dealing with identifying causal impacts for a cross-sectional unit over time is the traditional fixed effects model for panel data ([16]), which is what we will use as the basis for our modeling approach. Lastly, the two studies previously explained prior provide evidence that personalized campaigns (Type A) are distributed to specific households based on specific control period behavior. However, in addition to this reasonable claim, we also suspect that the endogeneity associated with the distribution of Type A campaigns are associated with time-varying characteristics realized in periods spanning the promotional cycle. Therefore, the personalized rewards are likely allocated to the best customers in hopes of inducing additional, in-store shopping frequency within the promotional cycle temporally at multiple periods. The type of endogeneity is attributed to self-selection as these Type A exposures are specifically designated to a certain subsample of the cross-section over specific sub-cycles within the test period. Failure to adequately account for such a phenomenon in a standard longitudinal regression context violates the

³Internal validity refers to the ability or likelihood in which a study, either a pure experimental or quasi-experimental design, is able to establish causality ([12]).

zero-covariance assumption between the contemporaneous errors and covariates resulting in highly biased and inconsistent parameter estimates ([17]).

Lastly, for proper estimation of model parameters of interest in longitudinal settings, one should carefully consider the cause of why a panel may be unbalanced. For instance, there can exist an underlying phenomenon related to missing observations that could induce otherwise balanced panels to result in unbalanced panels. We explore a complete-case (CC) approach by employing a selection indicator which results in a scalar multiple indicator applied to the full covariate vector and response at each time t similar to [18]. More details will be expounded upon in the body of Chapter 2.

The described application as well as missingness inspires our research direction in the chapter 2. The principal research questions are as follows:

1. Given an unbalanced panel as a result of missingness, is it possible to derive a fully robust test for endogeneity using a complete-case approach?
2. If endogeneity is present in an unbalanced panel, what estimation technique can be employed in order to achieve appropriate inference?

In Chapter 2, we utilize a combined correlated random effects model with a control function approach to develop a test for idiosyncratic endogeneity given a nonlinear, unbalanced panel data model with a suspected endogenous binary covariate in the presence of time-constant endogeneity. If endogeneity is identified, an instrumental variable, generalized method-of-moments (IV-GMM) approach is shown to be a solution for estimation in the presence of endogeneity.

1.3 Estimating Generalized Additive Mixed Models

Generalized additive mixed models (GAMMs) can be viewed as an intuitive extension of generalized linear mixed models (GLMMs) that offers additional flexibility of included model covariates as they can be estimated via unknown smooth functions ($f(\cdot)$). The traditional linear predictor of a generalized linear model is replaced, or augmented, by a sum of smooth functions. The genesis of the model forms follow from [19] and [20] in which the smooth functions are estimated via a scatterplot smoother using a local scoring algorithm. Additionally, if an empirical researcher were interested in a smoothed covariate's interaction with a covariate in the parametric portion of a GAMM, then one can include factor smooths interactions. We offer an exploration into GAMM factor smooth interactions as they have been rarely explored in previous marketing studies. Moreover, in the case of estimating generalized additive mixed models, there has been little investigation regarding the finite-sample performance of variance components with complex correlation structures. While the estimation of variance components for traditional linear and generalized linear mixed models have been explored, we seek to explore the joint estimation of variance components, unknown factor smooths that vary by levels of the estimated smooth function, and parametric fixed effects in GAMMs

under stationary and non-stationary AR(1) structures for longitudinal, Poisson counts. This finite-sample performance analysis is the focal point for the principle contribution in this chapter. We construct simulations in which we manipulate the test functions corresponding to the smooth functions, the coefficient associated with the variance of the random effects, the coefficient associated with the autocorrelation parameter, cluster size, and the number of repeated measurements per each subject in order to assess the sophistication of GAMM models in proper estimation of parameters of interest. Additionally, we consider cases of misspecified data generating processes where we assess the GAMM models' ability to estimate variance components on the boundary of the respective parameter's space. Estimation will explicitly be employed utilizing a penalized quasi-likelihood (PQL) approach after re-casting a GAMM into an overparameterized GLMM. To summarize, in Chapter 3, we explore:

1. Estimation of GAMMs re-parameterized as GLMMs with factor smooth interactions.
2. Simultaneous estimation and performance of GAMMs with factor smooths utilizing simulations for estimation of the fixed, parametric component of the model as well as the nonlinear component as well as all variance components in a variety of scenarios.

1.3.1 Application II: Carryover effect estimation into the First Intermittent Cycle of a Campaign Exposures

Utilizing the data from Chapter 2, we estimate potential nonlinear carryover effects in the epoch immediately after the end of households' first campaign promotion in the promotional cycle. This epoch can also be considered as the first intermittent period of non-promotion beyond the baseline periods for all households.

1.4 Extending the Granger-Ramanathan Optimal Solution via a Penalized Approach

Combination forecasting models have been studied heavily in recent history with recurring consideration given to linear forecasting combinations. We offer a brief review of the basic forms considered by the general form of the "linear-in-weights" forecasting combination proposed by [21]. Following the form of [22], let the variable of interest in forecasting, h -step ahead at time t , be denoted as y_{t+h} . Let us also assume we have a vector of forecasts, computed ex-ante and based on a corresponding information set $\mathcal{F}_t$, of size $N \times 1$, that is available and denoted as $\hat{\mathbf{y}}_{t+h,t}$.

Consider the regression

$$y_{t+h} = \omega_{0h} + \boldsymbol{\omega}'_h \hat{\mathbf{y}}_{t+h,t} + \varepsilon_{t+h} \quad (1.1)$$

in which the goal is to choose weights $(\omega_{0h}, \boldsymbol{\omega}_h)$ to produce a combined forecast of the form $\hat{y}_{t+h,t}^c = \omega_{0h} + \boldsymbol{\omega}'_h \hat{\mathbf{y}}_{t+h,t}$.

If we assume that y_{t+h} and $\hat{\mathbf{y}}_{t+h,t}$ have a joint distribution, $D(y_{t+h}, \hat{\mathbf{y}}_{t+h,t})^4$, then under the mean squared loss function (quadratic loss function), the first two, finite moments are of significance and can be represented as:

$$E \begin{pmatrix} y_{t+h} \\ \hat{\mathbf{y}}_{t+h,t} \end{pmatrix} = \begin{pmatrix} \mu_{y_{t+h,t}} \\ \boldsymbol{\mu}_{\hat{\mathbf{y}}_{t+h,t}} \end{pmatrix} \quad \text{Var} \begin{pmatrix} y_{t+h} \\ \hat{\mathbf{y}}_{t+h,t} \end{pmatrix} = \begin{pmatrix} \sigma_{y_{t+h,t}}^2 & \boldsymbol{\sigma}'_{y\hat{\mathbf{y}}_{t+h,t}} \\ \boldsymbol{\sigma}_{y\hat{\mathbf{y}}_{t+h,t}} & \Sigma_{\hat{\mathbf{y}}_{t+h,t}} \end{pmatrix}$$

where the optimal weights (population valued parameters) of both the constant and combination weights, respectively, are $\omega_{0h}^* = \mu_{y_{t+h,t}} - \boldsymbol{\omega}_h^{*'} \boldsymbol{\mu}_{\hat{\mathbf{y}}_{t+h,t}}$ and $\boldsymbol{\omega}_h^* = \frac{\boldsymbol{\sigma}_{y\hat{\mathbf{y}}_{t+h,t}}}{\Sigma_{\hat{\mathbf{y}}_{t+h,t}}}$. If the forecast itself is linear in combination weights and under the presumption of quadratic loss (MSE), the weights will depend only on the first two moments of the joint distribution of the vector of forecasts and the response variable of interest. This equation is estimable by OLS regression where the response is the dependent variable and the forecasts' vector along with a constant are explanatory covariates. This is the most general form with the inclusion of the intercept, which does not require the forecasts in the linear combination composition to be unbiased as any bias can be adjusted through ω_{0h} ([23]). The sample version of these weights is the usual least squares estimates for the response regressed on vector of forecasts and a constant ([21],[23]).

Two other noteworthy considerations are

- (i) $y_{t+h} = \boldsymbol{\omega}_h' \hat{\mathbf{y}}_{t+h,t} + \varepsilon_{t+h}$
- (ii) $y_{t+h} = \boldsymbol{\omega}_h' \hat{\mathbf{y}}_{t+h,t} + \varepsilon_{t+h}$ s.t. $\boldsymbol{\omega}_h' \boldsymbol{\iota} = 1$

Consideration (i) can also be estimated by ordinary least squares with no constant, which mandates that the forecasts from each model be unbiased. Also, (ii) is easily estimated by constrained ordinary least squares. The restriction itself that the weights sum to 1 is employed to assure that the combined forecast of unbiased forecasts is unbiased and that the forecast combination solution uses all available information ([24]).

In Chapter 4, we look to extend the Granger-Ramanathan (GR) solution to combination forecasting vis-a-vis an L1, LASSO penalty that discards useless underlying forecasting models while keeping those forecasts that contribute positively to out-of-sample prediction at certain levels of a forecasting target. Following the work of [25], under the quadratic loss function, the optimal solution to the Granger-Ramanathan form is the conditional mean, which can be approximated by the aggregation of conditional quantiles. Given this observation, a LASSO penalty can be applied on a quantile regression, which will be utilized as a tool that aids in classifying underlying forecasting models into three categories: “strong”, “partially weak”, and “weak”. The resulting combination of quantile forecasts yields a prediction equation that has a conditional mean interpretation as in [21], but now the associated, estimated coefficients reflect the presence of “strong”, “partially weak” and “weak” forecasting models.

⁴We suppress the joint distribution and the finite moments on past information $\mathcal{F}_t$.

1.4.1 Application III: Forecasting Combination Problem of Monthly Annualized Inflation Gap with Penalization

The principle forecasting target considered in our application is the monthly annualized inflation gap, which is computed as the difference between two temporal functions of the consumer price index (CPI). In the first phase of our forecasting application, we estimate the target utilizing 10 univariate forecast models, utilizing a recursive forecasting scheme, and record the forecasts at various forecasting horizons $h = \{1, 2, 3, 4, 8, 12\}$. In the next phase of our forecasting analysis, these point forecasts over varying horizons are combined via a penalized quantile regression method, where our proposed L1-penalized quantile forecasting combination (L1PQFC) model is compared against other popular combination forecasting models. We compare these models on a variety of metrics over a diversified range of forecasting horizons in order to assess our proposed model utility in terms of forecasting accuracy, our proposed models' robustness, and the tradeoff between squared bias and forecast variance relative to an equal weight benchmark.

References

- [1] W. H. Greene, *Econometric analysis*. Pearson Education India, 2003. [2](#)
- [2] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. MIT press, 2010. [2](#)
- [3] D. F. Hendry and J. N. Muellbauer, “The future of macroeconomics: macro theory and models at the bank of england,” *Oxford Review of Economic Policy*, vol. 34, no. 1-2, pp. 287–328, 2018. [2](#)
- [4] E. W. Frees *et al.*, *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press, 2004. [2](#)
- [5] A. W. Allaway, R. M. Gooner, D. Berkowitz, and L. Davis, “Deriving and exploring behavior segments within a retail loyalty card program,” *European Journal of Marketing*, vol. 40, no. 11/12, pp. 1317–1339, 2006. [3](#)
- [6] C. Mauri, “Card loyalty. a new emerging issue in grocery retailing,” *Journal of Retailing and Consumer Services*, vol. 10, no. 1, pp. 13–25, 2003. [3](#)
- [7] FD, *Grocery Rewards Programs Consumer Insights*, 2008. (accessed November 30, 2016). [3](#)
- [8] R. Venkatesan and P. W. Farris, “Measuring and managing returns from retailer-customized coupon campaigns,” *Journal of marketing*, vol. 76, no. 1, pp. 76–94, 2012. [3](#)
- [9] A. Baik, “Personalized coupon campaigns: Understanding and directing customers,” 2015. [4](#)
- [10] A. Chaudhry, *ESSAYS ON MULTICATEGORY DEPENDENCE MODELS OF CONSUMER CHOICE*. PhD thesis, Rice University, 2014. [5](#)
- [11] T. D. Cook, D. T. Campbell, and W. Shadish, *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, 2002. [6](#)
- [12] M. A. Handley, C. R. Lyles, C. McCulloch, and A. Cattamanchi, “Selecting and improving quasi-experimental designs in effectiveness and implementation research,” *Annual review of public health*, vol. 39, pp. 5–25, 2018. [6](#)
- [13] E. C. Strumpf, S. Harper, and J. S. Kaufman, “Fixed effects and difference in differences,” 2017. [6](#)
- [14] I. Petrakis, “The application of quasi-experimental methods in the context of marketing: a literature review,” 2016. [6](#)
- [15] J. D. Angrist and J.-S. Pischke, *Mastering’metrics: The path from cause to effect*. Princeton University Press, 2014. [6](#)

- [16] O. Deschenes and K. C. Meng, “Quasi-experimental methods in environmental economics: Opportunities and challenges,” tech. rep., National Bureau of Economic Research, 2018. [6](#)
- [17] J. Antonakis, S. Bendahan, P. Jacquart, and R. Lalive, “Causality and endogeneity: Problems and solutions,” [7](#)
- [18] J. M. Wooldridge, “Correlated random effects models with unbalanced panels,” *Journal of Econometrics*, 2018. [7](#)
- [19] T. Hastie and R. Tibshirani, “Generalized additive models: some applications,” *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 371–386, 1987. [7](#)
- [20] T. Hastie and R. Tibshirani, *Generalized additive models*. Wiley Online Library. [7](#)
- [21] C. W. Granger and R. Ramanathan, “Improved methods of combining forecasts,” *Journal of Forecasting*, vol. 3, no. 2, pp. 197–204, 1984. [8](#), [9](#)
- [22] A. Timmermann, “Forecast combinations,” 2005. [8](#)
- [23] A. Timmermann, “Forecast combinations,” *Handbook of economic forecasting*, vol. 1, pp. 135–196, 2006. [9](#)
- [24] C. G. Gibbs, “Forecast combination, non-linear dynamics, and the macroeconomy,” *Economic Theory*, vol. 63, no. 3, pp. 653–686, 2017. [9](#)
- [25] L. R. Lima and F. Meng, “Out-of-sample return predictability: A quantile combination approach,” *Journal of Applied Econometrics*, vol. 32, no. 4, pp. 877–895, 2017. [9](#)

CHAPTER 2

Quantifying the Exposure Effects of Customized and Targeting Promotions on Shopping Frequency in the Presence of Dual Sources of Endogeneity

Abstract

Endogeneity is a concern that must always be accounted for as its existence in any parametric model completely invalidates statistical inference due to unrectifiable bias and parameter estimate inconsistency. For panel data, this phenomenon is further complicated as it can exist in plurality as time-constant heterogeneity as well as through time-varying unobservables non-zero covariance with an included endogenous regressor. As the general objective for many studies is to rightfully overcome endogeneity in any form, there is a lack of focus in overcoming dual endogeneity in the context of nonlinear, unobserved effects panel data models with unbalancedness. Moreover, for endogeneity that manifests via non-zero covariance between time-varying unobservables and an included endogenous regressor, little consideration has been fully given regarding a discrete endogenous regressor's form in an unbalanced panel context, which is the focal point of this chapter. We seek to develop a fully robust Wald test for idiosyncratic endogeneity of a discrete regressor using a combined correlated random effects/control function approach given panel unbalancedness. We even give consideration for our proposed test using a complete-case (CC) approach as a protection against panel unbalancedness resulting from a missingness mechanism. Utilizing simulations, we identify the rejection rates, nominal sizes, and statistical power in our robust tests for identifying endogeneity of a discrete regressor in the presence of an unobserved effect in multiple panel structures. If endogeneity is present, we propose an instrumental variables, generalized method-of-moments (IV-GMM) approach producing feasibly efficient GMM estimators under varying assumptions about the covariance matrix. We link the aforementioned theory and estimators to a promotional marketing application in which there is suspected endogeneity associated with the distribution of a particular campaign via selection bias in a promotional loyalty program for a sample of household shoppers. We seek to properly estimate the impact of single and joint campaign exposure effects on visits for households who participate in the loyalty program.

2.1 Introduction

Nonlinear panel data models, in which the response covariate is of limited range such as (e.g. binary, counts), are ever-present and of utmost interest to amongst econometric researchers. When the response represents counts of a cross-sectional unit i across a designated time t is of interest, one of the most attractive model estimator is the Poisson fixed effects (FE) estimator, fully explored by [1]. The most notable assumption of this model is that it allows for any arbitrary correlation between the included, time-varying covariates and the multiplicative (to the conditional mean function), time-constant, unobserved heterogeneity. Explicitly, this signifies that any latent, time-invariant cross-sectional characteristics are permitted to be entirely correlated with the time-varying explanatory variables. We briefly review the Poisson FE estimator here.

Let $y_{it} = 0, 1, 2, \dots$ represent a count variable, $\mathbf{x}_{it}$ is a complete covariate vector that can contain time-invariant and time-varying predictors which also includes a constant for the intercept, $\boldsymbol{\beta}$ corresponds to the parameter vector associated with the covariates, and α_i be unobserved heterogeneity. Assuming conditional independence of $y_{it}, y_{iu}, t \neq u$ given $(\mathbf{x}_{it}, c_i)$ and Poisson distributed counts, then the Poisson probability specification is

$$P(y_{it}|\mathbf{x}_{it}, c_i) = (\mu_{it}^{y_{it}} \exp(-\mu_{it})) (y_{it}!)^{-1}$$

where the conditional mean $\mu_{it} = \alpha_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i)$. To emphasize, the two necessary assumptions for success in estimation are that the responses for a given i across t , y_{it} , conditional upon both the covariates and the time-constant heterogeneity, are Poisson distributed and the temporal independence of the counts for each cross-sectional unit i at time t . Additionally, no restriction is placed upon the relationship between unobserved heterogeneity and the covariates in $\mathbf{x}_{it}$ (or $E[\mathbf{x}_{it}'c_i] \neq 0$). Using a conditioning argument on component sums, the method of attaining a FE Poisson estimator via a conditional maximum likelihood (CML) approach is proposed.

As the sum of Poisson counts across time, $s_i = \sum_{t=1}^T y_{it}$, is Poisson distributed with the specified mean $s_i = \sum_{t=1}^T y_{it} \sim \text{Pois}\left(\sum_{t=1}^T \mu_{it}\right)$, then the conditional joint distribution of the Poisson counts can be written as

$$P(y_{it}|\mathbf{x}_{it}, c_i, s_i) = \prod_{t=1}^T q_{it}^{y_{it}} \left(\sum_{t=1}^T y_{it}! \right) \left(\prod_{t=1}^T y_{it}! \right)^{-1}$$

where $q_{it} = \left(\mu_{it} \right) \left(\sum_{t=1}^T \mu_{it} \right)^{-1} = \left(\exp(\mathbf{x}_{it}'\boldsymbol{\beta}) \right) \left(\sum_{t=1}^T \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) \right)^{-1}$ and $\sum_{t=1}^T q_{it} = 1$. The resulting log-likelihood from the joint distribution is $\ell(\boldsymbol{\beta}) = \sum_{i=1}^N \sum_{t=1}^T y_{it} \ln(q_{it})$.

This argument completely circumvents the issue of parameter inconsistency¹ from traditional MLE as the conditional maximum likelihood estimation (CMLE) approach is utilized to extricate the fixed effects from the likelihood function. As a beneficial extension from [2], the authors derive an estimator, given the first two aforementioned assumptions² in the Poisson FE case holds, the parameter estimate consistency remains achievable given correct specification of the conditional mean function with hardly any restrictions on the (conditional) variance function. The resulting estimator is estimated via quasi-maximum likelihood estimation³.

Another popular panel data estimator is the Poisson random effects (RE) estimator, which, like the FE case, maintains the same two assumptions regarding the conditional distribution of the response and the temporal independence between counts (strict exogeneity). For a random draw of a cross-section i from a population of interest and a conditional mean specification $E[y_{it}|\mathbf{x}_{it}, \alpha_i] = \alpha_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \equiv \exp(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)$, it is assumed c_i is independent of $\mathbf{x}_{it}$ and $\exp(c_i) \equiv a_i \sim \text{Gamma}(\delta, \delta)$ ⁴. Parameters of the joint probability distribution⁵ can be estimated by maximum likelihood estimation. The QMLE analog for this random effects estimator for consistency holds true as well as long as the conditional mean is properly specified and the strict exogeneity assumption holds. Similarly, either of these models can also be applied to unbalanced panels without distorting theoretical asymptotics when any missingness occurs at random ([3], [4]). The primary difference in the RE vs. FE model specifications is that, for the former case, the unobserved heterogeneity is presumed to be completely uncorrelated with the complete covariate vector $\mathbf{x}_{it}$ as it thought to be a random variable specified with its own distribution. Explicitly, to emphasize in clear notation, c_i is a random effect with no correlation between it and the time-varying explanatory covariates (e.g. $E[c_i|\mathbf{x}_{i1} \dots \mathbf{x}_{iT}] = 0$) ([3]). Clearly, this differs from the FE estimator, which treats c_i as a parameter to be estimated.

The most attractive benefit of a fixed effects estimator, particularly when applied to data from a non-experiment, arises due to its ability to completely extricate a large source of bias resulting from controlling for subject-specific effects, at the expense, however, of increased sampling variability due to the variance attributed to the exogenous covariates, which vary both within and between subjects ([5]). The FE estimator completely disregards any between-subject variability (and, thus, disregards information), which ultimately can increase standard errors associated with coefficients relative to estimators or methods that account

¹The likely inconsistency is due to the “incidental parameters problems” in the case of the “Small T and Large N asymptotics”. Assuming the time-constant heterogeneity, c_i , for example, is a parameter that one wants to estimate via maximum-likelihood estimation (MLE), if one has only T observations to estimate the time-constant effect for each i , then as $N \rightarrow \infty$, c_i remains stochastic or noisy, resulting in likely general inconsistency. The traditional MLE estimator cannot remove c_i and it does not dissipate as $n \rightarrow \infty$.

²These two assumptions mandate a “strict exogeneity” presumption as we condition upon all observables including the covariates at all time periods and time-constant heterogeneity, e.g. $E[y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, c_i] = E[y_{it}|\mathbf{x}_{it}, c_i]$. Therefore, once controlling for the time-constant heterogeneity and the time-varying explanatory variables, $\mathbf{x}_{is}$ exhibits no partial influence on y_{it} , $s \neq t$.

³The estimator is known as the quasi-maximum likelihood estimator (QMLE)

⁴The multiplicative heterogeneity $\exp(c_i)$ follows a gamma random variable with parameters (δ, δ) such that $E[\exp(c_i)] = 1$ and $\text{Var}(\exp(c_i)) = 1/\delta$. Additionally, the multiplicative heterogeneity, need not follow a Gamma distribution, but it must follow a parametric distribution $c_i \sim N(0, \sigma^2)$ (which is the log-normal).

⁵As $\mathbf{x}_{it}$ has an intercept, the normalization of the mean and variance of multiplicative heterogeneity results in no loss of generality. Integration by parts can be utilized to obtain the joint probability distribution that is free of the multiplicative heterogeneity [1].

for both sources of variance. Also, a fixed effects framework assumes the cross-sectional units and their corresponding within-unit variances are all identical. However, the FE specification will be completely ineffective if there is little to no within-subject variability in subjects which will result in standard errors that may be too large for any inference utility. Nevertheless, in most cases for which conclusions are essential for the sample at hand, the fixed effects framework is preferred as statistical inferences are made conditional upon the effects of interest in the sample ([6]). The random effects estimator, conversely, is most beneficial when both time-invariant and time-variant covariates are of principal interest⁶ and statistical inference is of interest to an entire population (instead of a sample). That is, in a random effects framework, the units i are assumed to be a random sample drawn from a population, thus the variance between them is of primary interest that allows one to claim inferences about a population. As an empirical test for practitioners, a statistical test, known as the Hausman test, is often suggested as a viable statistical test to for RE and FE specification ([7]). Let $\hat{\beta}^{\text{RE}}$ and $\hat{\beta}^{\text{FE}}$ represent the RE and FE estimators, respectively, with the estimated variance denoted as $\hat{\mathbf{V}}$. Under the null hypothesis, the time-varying covariates are assumed to not be correlated with the time-constant effect, and the test statistic can be computed as

$$\mathbf{H} = (\hat{\beta}^{\text{RE}} - \hat{\beta}^{\text{FE}})' [\hat{\mathbf{V}}[(\hat{\beta}^{\text{RE}} - \hat{\beta}^{\text{FE}})]]^{-1} (\hat{\beta}^{\text{RE}} - \hat{\beta}^{\text{FE}}) \stackrel{a}{\sim} \chi_K^2 \text{ } ^7$$

where, if the null hypothesis fails to be rejected, the test indicates that a random effects specification is preferred as it is both consistent and efficient relative to the consistent only fixed-effects framework⁸ ([8]).

Additionally, from a modeling perspective, much debate has been had regarding which estimator should be employed in order to properly estimate an unobserved effects model, specifically when there is any likelihood of covariate omissions. An unobserved effects model for counts can be explicitly represented as

$$E[y_{it} | \mathbf{x}_{it}, c_i] = \exp(\mathbf{x}_{it}' \boldsymbol{\beta} + c_i)$$

where $y_{it} = 0, 1, 2, \dots$. This model specification for counts is extremely beneficial as it can be completely specified by the conditional mean, and, conveniently, it shares the conditional expectation structure for the aforementioned Poisson FE and Poisson RE models. Therefore, to summarize from a practical perspective in applied work, if one believes that there exists omitted variables that are time-invariant, then a fixed effects estimation is the appropriate modeling choice given these two estimators.

Both variants of time-constant covariates, measured and unmeasured, are completely parsed out of the likelihood function. If it is truly believable that there are virtually no covariate omissions whatsoever, or whatever omissions that exist are likely to not be correlated with any included covariates in $\mathbf{x}_{it}$, then a

⁶Relative to the FE framework, the RE model does not ignore between-subject variability.

⁷Note, K denotes the rank of $[\hat{\mathbf{V}}[(\hat{\beta}^{\text{RE}} - \hat{\beta}^{\text{FE}})]]$, which signifies this test statistic is asymptotically chi-squared distributed with K degrees of freedom.

⁸As a point of clarification from [7], the FE estimator is consistent whether the multiplicative heterogeneity term is fixed or random, but the RE estimator achieves consistency and will be efficient, or have smaller standard errors associated with the coefficients, relative to the FE if and only if the heterogeneity term is truly uncorrelated with the covariate vector. The RE estimator will be inconsistent if there is any correlation between the covariate vector and the multiplicative heterogeneity term.

random effects estimator is appropriate as it will produce unbiased estimates of the coefficients, reduced standard errors (relative to the fixed effects specification), and parameter estimates on any time-constant covariates ([5]). However, if there are likely time-varying covariate omissions, which estimator is preferable? Moreover, how is one able to detect the presence of time-varying covariate omissions? Additionally, to what extent is the choice of either of these estimators (or any estimator) preferable if endogeneity is found as a result in a non-zero covariance of a discrete time-varying covariate and time-varying omissions? To what extent will the estimator choice be affected if the panel structure is unbalanced?

A primary contribution from [9] utilizes Mundlak’s projection as an attempt to unify the FE and RE estimators for linear and nonlinear panel data. This model is commonly referred to as the correlated random effects (CRE) model. Correlated random effects models can be thought of as a linked framework between both fixed effects (FE) and random effects (RE) estimators. The intuition behind this model is to partition the time-varying covariates into a portion representing the within-subject variability and a portion that represents between-subject variability ([10]). CRE models are defined to relate the unobserved heterogeneity to the observed explanatory covariates (or any distinct set of exogenous variables) by specifying some restrictive, conditional distribution for the unobserved effect given the covariates ([9]). The approach explicitly models the conditional distribution of unobserved heterogeneity given a set of observable covariates as a linear function. The unobserved, time-constant heterogeneity itself is modeled as a linear projection on the averages of all (time-varying) covariates which results in an error term that is completely uncorrelated and independent of the explanatory covariates. Therefore, the FE assumption for correlation can be accounted for through the group means of the time-varying covariates. Advantageously, this model will result in coefficients associated with time-varying covariates that exactly mimics to the FE estimator’s coefficients, and additionally, will simultaneously permit coefficients on any time-invariant characteristics via their respective time averages. In terms of previous literature, this correlated random effects specification has proven to be indispensable in its application to nonlinear, panel data models. These models have been studied in previous literature applying to both linear and nonlinear (probit and tobit) panel models ([3]). Another recent work has employed the ability of the CRE model to assess the appropriateness of the fixed-effects specification using within-family variability by nesting the conventional fixed-effects specification via additional parameter heterogeneity ([11]). This modeling approach has very recently been extended to estimating dynamic⁹, nonlinear panel data models given an unbalanced panel structure ([15]) using a minimum distance approach that is asymptotically equivalent to maximum likelihood estimation. In addition to these empirical applications, both general methodology and estimation strategies under various settings such as unbalanced panels ([16],[17]) have been carefully considered. A summary of the primary assumptions, advantages, and disadvantages for the FE, RE, and CRE models can be found in Table 2.1.

As a brief example illustrating the CRE model in a relevant framework, recall the conditional mean formulation for a unobserved effects model for counts as $E[y_{it}|\mathbf{x}_{it}, c_i] = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i)$. We recall the

⁹Dynamic panel data signifies that a necessarily included covariate is at least a one-period lag of y_{it} as y_{it-1} . See [12], [13], [14], and [3].

Table 2.1: General Advantages and Disadvantages for FE, RE, and CRE models

Panel Model	Assumption	Advantages	Disadvantages
Fixed Effects Model	c_i is correlated with $\mathbf{x}_{it}$	• Removes potential large biases associated with time-invariant unmeasured covariates • Relevant when inferences about the observed sample is of interest • Generally consistent under correct assumptions	• No estimation on any time-invariant effects • Useless if there is no variation within-subject
Random Effects Model	c_i is uncorrelated with $\mathbf{x}_{it}$	• Allows for estimation of time-invariant covariates • Relevant when inferences about a population is of interest • Can result in smaller standard errors relative to the fixed effects approach	• Primary assumption is hard to verify • Results in inconsistency if primary assumption is not met
Correlated Random Effects Model	c_i is assumed to be linearly related with $\mathbf{x}_{it}$	• Unifies the FE and RE approaches • Allows for estimates of both time-varying, time-invariant covariates • The coefficients on the time-varying covariates are identical to the FE estimates	• The linear dependence using a linear projection on time averages (Mundlak's device) can be restrictive

assumptions of the unobserved heterogeneity in relation to the covariate vector are $E[\mathbf{x}_{it}'c_i] = 0$ and $E[\mathbf{x}_{it}'c_i] \neq 0$ for the RE and FE model specifications, respectively. Given the assumption of strict exogeneity, Mundlak proposes a linear projection for the time-constant effect on the time dependent averages as $c_i = \bar{\mathbf{x}}_i'\gamma + a_i$ where the error term a_i , is presumed Gaussian and independent for each i and is independent of the covariate vector $\mathbf{x}_{it}$ at all time points. The result signifies that the time constant heterogeneity can linearly depend on the time-varying covariates via the time averages, and such an inclusion of $\bar{\mathbf{x}}_i$ into the conditional mean function as $E[y_{it}|\mathbf{x}_{it}, c_i] = \exp(\mathbf{x}_{it}'\beta + c_i) \equiv \exp(\mathbf{x}_{it}'\beta + \bar{\mathbf{x}}_i'\gamma + a_i)$, mimics a FE estimator with a notable correlation between the time averages and time constant unobserved heterogeneity. As for the shared benefit with random effects estimator, this model is flexible enough also to include and estimate time-constant effects. However, as the strict exogeneity applies for the CRE formulation, then the time-varying covariate vector $\mathbf{x}_{it}$ cannot be correlated with any time-varying omitted variables across any time period t .

The propensity of an unobserved effects, nonlinear panel to be simultaneously prone to both sources of endogeneity provides the basis of focus for this chapter. It would be very beneficial if there existed a method that allows controls for time-constant, unobserved heterogeneity (or endogeneity that is time constant) while isolating a test solely for idiosyncratic endogeneity (time-varying unobservables correlated with included regressors) particularly in the context of unbalanced nonlinear panels. A starting point for being able to handle these two sources of endogeneity would be a fixed effects instrumental variables FE-IV approach. A notable work, [18], explores the allowance for the time-varying covariates and the time-constant

heterogeneity as well as idiosyncratic endogeneity in the context of a correlated random coefficient (CRC) model (linear panel case) employing this IV approach¹⁰. Moreover, the conditions for which the FE-IV estimator is consistent are derived as well for the CRC model with consideration for an unbalanced panel structure due to selection. Using the CRE model in the context of a fractional response, a study of the log amount spent on test pass rates for Michigan schools likely correlated with time-varying unobservables is one of the first studies that allows for contemporaneous endogeneity of student performance in the pursuit of estimating average partial effects ([19]). Two contemporary works from [17] and [20] further investigate the previously unexplored CRE models in the case of both nonlinear and linear (respectively) unbalanced panels where, specifically, the presumption of unbalancedness is a result of selection. Moreover, the latter work investigates the result of a pure test of idiosyncratic endogeneity attributed specifically to the non-zero correlation between the explanatory covariates and idiosyncratic shocks while permitting all explanatory variables and instruments to be correlated with the time-constant endogeneity. We seek to explore the analog of [20] with an exponential mean function most commonly employed for count responses.

As recognized in the working paper by [21], the contemporary literature states that a pure test of idiosyncratic endogeneity has not yet been developed for a nonlinear, unobserved effects panel data model that is unbalanced and in the presence of time-constant endogeneity. While [20] has recently explored this case fully for linear panels with unbalancedness, we seek to explore the analog with an exponential mean function most commonly employed for count responses. To our knowledge, this is the first work that will utilize the CRE model with instrumental variables in tandem with a control function approach to provide a fully robust test of pure idiosyncratic endogeneity with an unbalanced, nonlinear panel structure in the presence of time-constant endogeneity using a complete-case (CC) approach. The panel’s unbalancedness resulting from the selection process is unmodeled as the CC approach can be conditioned upon for valid estimation, yet can be correlated with the time-constant endogeneity with no consequence. We will highlight the utility of our proposed Wald test for valid and trustworthy inference on the coefficient that captures the covariance between the unobservables and the suspected endogenous culprit with an adjustment for standard errors. For estimating our constructed final estimating equation, the robustness of the final stage Poisson QMLE is indispensable as it only requires the conditional mean be correctly specified. Moreover, we do not even restrict the time-varying endogenous covariate to be continuous.

The remainder of this chapter is structured as follows: Section 2.2 will expound upon the definition, causes, and remedies of endogeneity. Section 2.3 will highlight the methodology underlying our test of idiosyncratic endogeneity as well as some theoretical background regarding our development for the pure test of idiosyncratic endogeneity in nonlinear unbalanced panels. Additionally, we will provide the Monte Carlo simulation results for the test of idiosyncratic endogeneity under a variety of panel structures and varying parameters regarding the magnitude of endogeneity and strength of instruments. Section 2.4 will

¹⁰The strict exogeneity condition for the usual FE estimator not only prohibits idiosyncratic endogeneity, but also lagged values of y_{it} or any other regressors that where any change in the idiosyncratic errors (unobservables) potentially affects changes in future values of $x_{i(t+h)}$, $h \geq 1$.

describe our GMM approach for estimation and the development of three feasibly efficient GMM estimators. Section 2.5 provides an overview of our quasi-experiment of assessing the marginal effect of campaign exposure on visit rates in the test cycle and highlights the results from both our pure test for endogeneity and our IV-GMM results. Section 2.6 will conclude our findings.

2.2 Endogeneity

Endogeneity, as a phenomenon, is most often represented as the instance when explanatory covariates are correlated with an error term in a regression model. From a linear regression perspective, with an equation represented as $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_{it}$, endogeneity occurs if $E[\mathbf{x}_{it}'u_{it}] \neq 0$ (or $\text{Cov}(\mathbf{x}_{it}, u_{it}) \neq 0$). From a modeling perspective, endogeneity manifests itself in a variety of different ways for both linear and nonlinear cross-sections and panel data structures. Two well-explored instances of endogeneity are measurement errors (erroneously measured explanatory variables that can distort parameter estimates in a model) and simultaneity (explanatory covariates and response variable are conjointly determined) ([22]). However, in this chapter, we focus on a form of endogeneity that is extremely pervasive (and arguably more misunderstood) in econometric regression modeling: the unobserved heterogeneity/omitted variable bias. Omitted variable bias occurs when there is a failure to incorporate a relevant predictor into a regression equation, and these omitted predictors are then correlated with an included covariate in the model resulting in endogeneity. Additionally, an omitted variable bias could also arise when a model has excluded functions of explanatory covariates or if the conditional mean is completely misspecified, which is known as functional form misspecification ([23],[3]). A succinct explanation of the bias resulting from omitted variable is the difference between the true value of the parameter and the expected value of its estimator attributed to the lack of inclusion of relevant explanatory covariates ([24]). In the case in which omitted variables are ignored and ordinary least squares is employed, for example, the OLS estimator from linear regression is purely uninterpretable as it is both biased and inconsistent¹¹ resulting in invalid statistical inference. In more contemporary literature, it is argued that omitted selection is an underlying subsegment of the bias resulting from the omitted variables form. This case is quite common when researchers have observational data from a quasi-experiment which most often is derelict of the proper use of randomization of assigned treatments of interest. Without explicitly accounting for the underlying mechanism of treatment allocation or distribution to subjects in quasi-experimental analyses, the endogeneity resulting from omitted selection is likely imminent as differences between those experiencing an effect in a standard cross-section of observational data is likely to be correlated with the response of interest. This is further complicated if treatments are allocated at varying times temporally across the cross-sectional unit. In other words, we speak to a particularly noteworthy endogeneity issue that plagues quasi-experiments frequently in attempting to quantify effects of treatments,

¹¹As $n \rightarrow \infty$, the parameter estimate fails to converge in probability to the true parameter.

distributed to or for a target base, wherein these discrete treatments are naturally endogenous. These treatments manifest from decisions made by experimenters with intended, behavior consequences in terms of measured outcomes of interest. This is known as selection-based endogeneity ([25]).

In marketing and management, selection-based endogeneity can further be parsed in two ways: sample-selection and self-selection. In the sample-selection case, the process in which subjects are in studied or selected into the sample is dependent upon the response in a way that creates endogeneity due to correlations between the included covariates and the errors. From a modeling perspective, the outcome variable is solely observed for a subsample, resulting in a censoring mechanism observed in the study as all outcomes are observed solely for a fraction of the true sample. For the sample-selection case, we refer the seminal, award-winning work by [26] and expounded upon in [27]¹² for which the proposed Heckman model is formulated with two equations, the structural equation of interest as well as a selection equation that represents the likelihood of subject selection, as well as a variety of other reviews, extensions, endogeneity considerations, and critiques ([28], [29], [30], [3], [31], [32]). However, as relevant to our application, is the selection-based endogeneity resulting from self-selection. According to [25], since there is no underlying censoring that occurs within the sample, self-selection bias arises when the dependent responses are observed for differing sub-segments of the population; however, there is a non-random element related to these dependent outcomes. From a managerial perspective, this is done purposely by overseers for favorable outcomes. In essence, in an attempt to enhance outcomes or responses, the treatments observed and allocated to subjects are a function of organizational choices based on a variety of both observable and unobservable factors ([33]).

A clear contrast from which this self-selection bias differs from sample selection is that the response is observed in all cases, yet there exists a non-randomness involved in the construction of the response due to unobserved covariates. For an application study which we will analyze in a section 2.5, the potential bias resulting from self-selection is most relevant to the quasi-experiment as the treatments allocated by the designers of the loyalty program are specifically chosen, at designated temporal points across the cross-sectional households, with the goal of stimulating the response of interest: visits. Failure to account for this form of omitted variable bias represents both a hazard to internal validity of the design of interest which is likely to affect both the likelihood to receive a treatment of interest (such as personalized campaigns sent only to the best customers) over time as well as the performance outcomes (in-store visits) resulting in endogeneity. Therefore, coefficients from resulting statistical models are uninterpretable and untrustworthy due to statistical inconsistency.

¹²The famed sample-selection bias case is the impact of education on the wage equation, wherein a sizeable subset of unemployed individuals are included in the sample. As there is no information of wages (the response), these subjects cannot be included in the estimating equation, and a researcher is left with a non-random, truncated sample.

2.2.1 Conventional Remedies for Endogeneity

There have been a myriad of methods proposed from a variety of different texts and prior research that mitigates the omitted variable bias resulting from endogeneity issue in both linear and non-linear cross-section and panel framework. Popular approaches for dealing with a single endogenous regressor are control functions (CF), instrumental variables (IV), and generalized method-of-moments (GMM). We note that, dependent on the authors or literature, there can be a bit of confusion regarding the class of estimators that correspond to IV estimation compared to GMM estimation. The standard IV estimator is a special case of a GMM estimator. That is, dependent upon a weight matrix utilized in the GMM estimation, minimization of the corresponding objective function yields an IV estimator ([34]). Instrumental variables, in the case of mitigating endogeneity, must exhibit properties of being (1) strongly correlated with an endogenous regressor, yet (2) completely uncorrelated with the principal response except indirectly through its association with the endogenous culprit ([3]), and, intuitively, should only be associated to the response variable primarily through the relationship with the suspected endogenous regressor. As an illustration, we highlight the canonical causal graph of an instrumental variable model in a regression setting in Figure 2.1¹³. These two aforementioned properties are known as the relevance and exclusion restrictions, respectively, that enables the resulting IV estimator to approximate randomization in terms of proper estimation of covariates or treatment effects. The randomization of the treatment assignment absent in a quasi-experiment is approximated via the exogenous variation of properly chosen instruments associated with the endogenous variables ([35], [36]). In quasi-experimental designs, the relevance and exclusion edicts associated with instrumental IVs are indispensable if endogeneity is suspected as they can achieve consistent (albeit, biased) causal estimates.

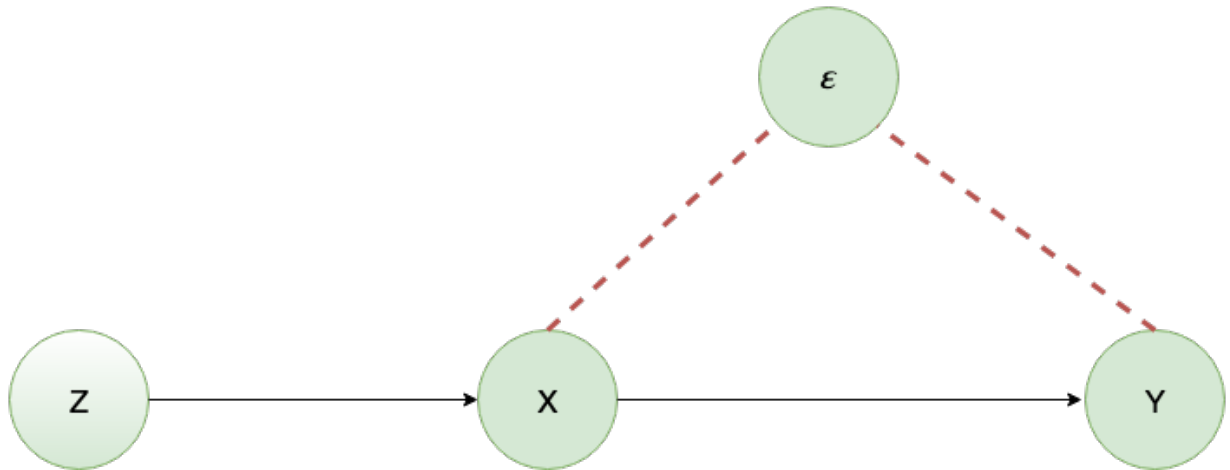


Figure 2.1: Instrumental Variables Causal Graph

¹³As seen in this Figure 2.1, the relationship between the covariate X and the response Y is confounded by an unobservable ϵ (as indicated by the dashed, bidirectional lines linking ϵ to both X and Y), prohibiting the direct marginal effect estimation of X on Y using a traditional regression model. However, given an independently determined instrument Z , which has a direct effect on X as indicated by “ $\rightarrow$ ” but not Y , then the figure shows that this additional variable induces variation in X that is independent of the unobservable ϵ . As this is the case, an auxiliary experiment can be performed utilizing Z that allows one to estimate the effect of X on Y , traditionally, via a two-stage, instrumental variables regression.

We proceed with a few pieces of literature that deal with endogeneity demonstrating these three methods. Utilizing a carefully constructed control function approach, [31] provides bias-corrected estimators for nonlinear panel unobserved effects model plagued with both endogeneity of invariant and temporally dependent forms. Seminal control function approaches involving the investigation of 2SPS (two-stage predictor substitution) and 2SRI (two-stage residual inclusion) from [37] observes endogeneous effects in two distinct simulation settings (duration models with endogeneous variables and an ordered logit) as well as an application to a cigarette smoking application for a cross-section of mothers. Another work employs a pure control function approach, which is an econometric method used to correct for biases that arise as a consequence of selection observed for correlated random coefficients models ([38]). Regarding the notion of traditional instrumental variable (IV) estimation, [39] surveys recent works employing IV methods and walks through a very detailed process for understanding the assumptions under which chosen instruments are appropriate as well as exploring these procedures through the STATA language. For prior studies involving generalized method-of-moments, a cross-sectional application studying household data to estimate the association between health insurance with other controls and the number of patient visits (outpatient) to the distinct categories of providers is explored by [40]. These authors apply both a GMM approach and a linear IV approach to correct for the potential endogeneity given different health insurance types (private vs. public). These set of authors, [41], take interest in exploring the similarities and differences in utilizing a control function approach versus a two-stage least squares approach (2SLS)¹⁴ as well as a combined formulation of these two modeling approaches in an attempt at estimating nonlinear causal effects models. A two-step, fixed effects instrumental variables (FE-IV) approach permitting arbitrary serial correlation has also been well-established ([42]). In the case of discrete endogenous variables, [43] seeks to mitigate the effects of endogeneous treatments utilizing data from a retrospective study has been considered using a full information maximum likelihood (FIML) approach¹⁵, requiring the joint distribution of the response and a dummy endogenous regressor conditional upon a vector of instrumental variables. Lastly, a Hausman-type test of a dummy endogenous regressor under a nonlinear least squares framework has been proposed by [47].

For panel data considerations, a study interested in modeling nonlinear panel data subjected to asymmetric dynamics, simultaneity, cross-sectional subjects' heterogeneity, and endogeneity existing in both standard covariates and a threshold covariate is explored using first-differenced GMM as well as a first-differenced, two-stage, least squares approach ([48]). In a recent analysis studying a cross-section of firms in the United Kingdom for expenditures related to research and design and time-invariant firm characteristics

¹⁴As a noteworthy point, the methodology surrounding the notion of a control function approach versus a 2SLS can be perplexing. While both of these methods involve two-stage estimation, the 2SLS seeks to parse variation in a treatment of interest that is independent of unmeasured unobservables and use the variation to estimate the treatment effect, whereas the control function approach partitions the unobservables into (i) a portion that is correlated with a treatment and (ii) an uncorrelated portion of the treatment (a "residual"). This uncorrelated portion, or the residual, is included in the second-stage estimation and is taken to be the control function. The residual of the first-stage regression seeks to account for the unobservables, and any effect of the unmeasurables is controlled for via its inclusion in the structural estimation ([41]).

¹⁵The full information maximum likelihood (FIML) procedure results in an asymptotically efficient estimator for system of simultaneous regression equations conditional upon normally distributed errors terms. Additionally, for models that are nonlinear in their parameters, it is the sole known efficient estimator ([44], [45], [46]).

on company performance, the dynamic GMM method is compared to traditional OLS and fixed effects estimation for understanding endogeneity bias resulting from simultaneity in the response and several explanatory covariates ([49]). Another advanced study springboards off of traditional GMM methodology to their novel method of penalized-focused, method-of-moments (FGMM) in order to achieve dimension reduction in high dimensions ([50]). In an effort to mitigate measurement error, omitted variable bias, simultaneity, and panel attrition resulting in sample selection bias in a common setting, [51] builds off of the techniques and methodology first established by [52] and [42] in an effort to appropriately study the interconnectedness between health and wages.

Thus, in both cross-sections and longitudinal data contexts and under specific assumptions, these methods can be utilized in order to mitigate endogeneity bias as well as achieve consistent estimation. However, all estimation methods do have a distinct set of advantages and disadvantages depending on the data structures, model assumptions, forms, and responses. Although we do not provide a complete survey of these for control function, IV, and GMM methods (although we reserve a more focused discussion for GMM in Section 2.4), we will highlight a few obstacles that plague both our test for endogeneity as well as parameter estimate consistency in an unbalanced panel.

We now proceed with the methodological discussion in the next section testing for idiosyncratic endogeneity in the presence of a time-constant unobserved effect for nonlinear, unbalanced panels using a combined correlated random effects/control function approach.

2.3 Correlated Random Effects with a Control Function Approach

- Test for Endogeneity in Nonlinear, Unbalanced Panels

As alluded to previously in Section 2.1, arguably the greatest utility offered by the usage of CRE models is that they can be unified with a control function approach for mitigating endogeneity in nonlinear models that are subject to both time constant endogeneity and idiosyncratic endogeneity. We will argue it is possible to both simultaneously conduct a fully robust test for idiosyncratic endogeneity in the presence of unobserved heterogeneity and produce coefficients that are considered asymptotically consistent under the main assumption of the integrity¹⁶ of the endogenous variable after a parametric correction of the standard errors ([3]). Alternatively, this proposed procedure can solely test for idiosyncratic endogeneity regardless of the limited nature regarding the range of values of the variable such as binary or count endogenous explanatory variables. In what follows, we focus solely on the latter case as the nature of our endogenous variable is binary and the parametric correction for producing consistent estimates will not apply. If the null hypothesis of idiosyncratic endogeneity is rejected, consistent estimation of our parameters of interest can be achieved via an even more robust method of estimation: the generalized method-of-moments (GMM) procedure. We proceed with the development of the test.

¹⁶The method is appropriate for endogenous variables that are *not* of restricted range (e.g. continuous).

2.3.1 Correlated Random Effects/Control Function Approach: Idiosyncratic Endogeneity Test

We now outline the procedure for testing for idiosyncratic endogeneity (at least one regressor is correlated with time-varying unobservables) while allowing for unobserved heterogeneity (at least one or more structural regressor is correlated with time-constant heterogeneity) ([21]). We start with the structural equation represented as a conditional exponential mean model, which is a log-linear¹⁷ model, with multiplicative, unobserved heterogeneity (α_{i1}) on the mean function and a single, binary endogenous regressor (y_{it2}). This form of a structural equation represents a nonlinear, unobserved effects panel data model.

The multiplicative heterogeneity term can be re-written algebraically in the additive sense. That is, we write

$$\begin{aligned} E[y_{it1} | \mathbf{z}_i, y_{it2}, r_{it1}, \alpha_{i1}, t_{it}] &= E[y_{it1} | \mathbf{z}_{it1}, y_{it2}, r_{it1}, \alpha_{i1}, t_{it}] = t_{it} \exp(\mathbf{z}_{it1}' \boldsymbol{\beta} + \varphi_1 y_{it2} + r_{it1}) \alpha_{i1} \\ &= t_{it} \exp(\mathbf{z}_{it1}' \boldsymbol{\beta} + \varphi_1 y_{it2} + \ln(\alpha_{i1}) + r_{it1}) \\ &= t_{it} \exp(\mathbf{z}_{it1}' \boldsymbol{\beta} + \varphi_1 y_{it2} + c_{i1} + r_{it1}) \end{aligned}$$

for $i = 1, \dots, n$, $t = 1, \dots, T$, with $c_{i1} = \ln(\alpha_{i1})$ and $E[\mathbf{z}_{it1}' c_{i1}] \neq 0$. This heterogeneity term under this transformation of course can still be interpreted as a multiplicative effect, or more clearly, an additive shift constant for the intercept and presumed to be correlated with the covariates in $\mathbf{z}_{it1}$ (thus, a fixed effect). Also, here, we presume that $\mathbf{z}_{it1}$ contains an constant for the intercept. Note, the time-varying unobservables, r_{it1} , are written inside the exponential function. Its specification within the exponential function leads to the interpretation as unobservable, unseen covariates that affect the expectation of y_{it} and treated symmetrically to the observed covariates ([53]). The structural equation of interest, representing the conditional mean of our response variable of interest for each cross-sectional unit i across time period t , is

$$E[y_{it1} | \mathbf{z}_i, y_{it2}, c_{i1}, r_{it1}, t_{it}, \mathbf{h}_i] = E[y_{it1} | \mathbf{z}_{it1}, y_{it2}, c_{i1}, r_{it1}, t_{it}] = t_{it} \exp(\mathbf{z}_{it1}' \boldsymbol{\beta} + \varphi_1 y_{it2} + c_{i1} + r_{it1}) \quad (2.1)$$

Before proceeding with other notational definitions, we define the selection indicator as

$$h_{it} = \begin{cases} 1, & \text{if } (y_{it1}, \mathbf{z}_{it1}, y_{it2}, t_{it}) \text{ is fully observed} \\ 0, & \text{otherwise} \end{cases}$$

That is, $\mathbf{h}_i \equiv \{h_{i1}, \dots, h_{iT}\}$, for each cross-sectional unit i , represents the series selection indicator. This implies that if $h_{it} = 1$, the entire row vector is observed, and time period t for cross-section i can be used

¹⁷The transformation using the log on the mean is linear in parameters.

in estimation¹⁸, representing the complete-case (CC) analysis. Formally, for the construction of our test, we assume the following:

Assumption 1 (Strict Exogeneity of Selection): $E(r_{it1}|\mathbf{z}_{it}, c_{i1}, t_{it}, \mathbf{h}_i) = E(r_{it1}|\mathbf{z}_{it}, c_{i1}, t_{it}) = 0, \quad t = 1, \dots, T.$

Therefore, the observed row vector in any time period t cannot be related to our time-varying unobservables r_{it1} once controlling for our exogenous variables and unobserved heterogeneity. As properly formalized in relatively recent publications ([42], [17]), this assumption implies a few things. Firstly, this assumption always holds true under the following two conditions: (I.) $\mathbf{z}_{it}|c_i$ is strictly exogenous¹⁹, and (II.) the vector of selection indicators, $\mathbf{h}_i \equiv \{h_{i1}, \dots, h_{iT}\}$, is completely random such that $\mathbf{h}_i \perp (\mathbf{z}_{it}, t_{it}, y_{it2}, r_{it1}, c_{i1}), \forall t = 1, \dots, T$. It can also hold true if $\mathbf{h}_i$ is a function of, as in our case, $(t_{it}, c_{i1}, y_{it2}, \mathbf{z}_{it})$. Lastly, this assumption does still allow selection $\mathbf{h}_i$ at any time t to be correlated with exogenous variables in the instrument vector (including the $\mathbf{z}_{it1}$ in the structural equation) as well as unobserved heterogeneity c_{i1} . Another common name of this assumption in econometrics is the conditional mean independence assumption.

For the remaining definitions, we have: y_{it1} representing the dependent variable of interest, $\mathbf{z}_{it1}$ represents a set of strictly exogenous covariates to the structural equation that is of size $1 \times K$, β represents the associated population parameters to be estimated, y_{it2} represents the time-varying, potentially endogenous regressor²⁰ with φ_1 being an associated coefficient. The variable t_{it} represents the offset variable, or the exposure effect, multiplicative to the conditional mean function. The latent, time-invariant variable c_{i1} represents the unobserved effect or the time-constant heterogeneity endogeneity for the cross-section of subjects, and r_{it1} represents the time-varying omitted unobservables/variables that is suspected to be correlated with y_{it2} . Lastly, $\mathbf{z}_{it}$ represents the set of all exogenous variables with an additional set of instruments (e.g. $\mathbf{z}_{it1} \subseteq \mathbf{z}_{it}$) relative to the structural equation. In other words, the exogenous variables are $\mathbf{z}_i = \mathbf{z}_{it} \equiv (\mathbf{z}_{it1}, \mathbf{z}_{it2})$ where the vector contains strictly exogenous variables in $\mathbf{z}_{it2}$ to be excluded from the structural equation. The dimension of $\mathbf{z}_{it}$ is $1 \times L$, where $L \geq K$. Conditional on the unobserved heterogeneity, the regressors in each time period are uncorrelated with the time-varying unobservables in each time period. We also assume, as for a traditional fixed effects specification, that all exogenous variables including those excluded in the structural equation are allowed to explicitly be correlated with the fixed, time-constant heterogeneity. This leads to a second assumption of strict exogeneity of our instruments on the structural equation as well as a third assumption traditional fixed effects assumption.

Assumption 2 (Strict Exogeneity of Instruments): $E[r_{it1}|\mathbf{z}_i, c_{i1}, t_{it}] = E[r_{it1}|\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT}, c_{i1}, t_{it}] = 0, \quad t = 1, \dots, T$. Although implied in **Assumption 1**, we formalize this assumption here. All instruments,

¹⁸The total number of periods for which each cross-sectional unit is observed can be defined as $\sum_{r=1}^N h_{ir} = T_i$

¹⁹Strictly exogenous here is the case in which conditional on the unobserved heterogeneity, c_{i1} , the time-varying unobservables are uncorrelated with all instruments $\forall T$. This prohibits lagged responses in the instrument vector ([3]).

²⁰These methods easily extend to a vector of time-varying endogenous variables, $\mathbf{y}_{it2}$, but for simplicity, we consider a single, endogenous case, y_{it2} .

$\mathbf{z}_i = \mathbf{z}_{it}$, are uncorrelated with the time-varying unobservables r_{it1} in each time period $\forall t = 1, \dots, T$ given the time-invariant heterogeneity c_{i1} . Thus, the idiosyncratic, time-varying unobservables has mean zero conditional on the time constant heterogeneity as well as all exogenous instruments for all time periods t .

Assumption 3 (Non-Zero correlation between all instruments and time-constant unobserved heterogeneity as well as selection): $E[\mathbf{z}'_{it}c_{i1}] \neq 0$, $E[\mathbf{h}'_i c_{i1}] \neq 0$. Again, this is implied by the conditional mean independence, **Assumption 1**, but we explicitly state it here. We assume that all exogenous variables can be correlated with the fixed effect and the selection indicator.

Assumption 4 (Non-singularity): $\sum_{t=1}^T \left[E[\mathbf{h}_i \mathbf{z}'_{it} \mathbf{z}_{it}] \right]$ is non-singular. None of elements of the instrument vector are assumed perfect linear combinations of each other (no multicollinearity).

Assumption 5 (Instruments and Endogenous Covariate Sufficiently Related): $\text{rank} \left[\sum_{t=1}^T E[\mathbf{h}_i \mathbf{z}'_{it} y_{it2}] \right] = L$. The instruments are correlated, at least partially, with the endogenous covariate y_{it2} .

For the remainder of outlying the methodology in this subsection, we will represent the conditional mean function multiplying through by the selection indicator h_{it} as

$$E[h_{it} y_{it1} | \mathbf{z}_i, y_{it2}, c_{i1}, r_{it1}, t_{it}] = h_{it} t_{it} \exp(h_{it} \mathbf{z}'_{it1} \boldsymbol{\beta} + h_{it} \varphi_1 y_{it2} + h_{it} c_{i1} + h_{it} r_{it1}) \quad (2.2)$$

where observations are only used for estimation when $h_{it} = 1$.

The endogenous variable y_{it2} , presumably, has a linear reduced form that is a function of the exogenous variables including the additional instruments, $\mathbf{z}_i$, and the associated unobserved heterogeneity, as $h_{it} y_{it2} = h_{it} \mathbf{z}'_{it} \boldsymbol{\Lambda}_2 + h_{it} c_{i2} + h_{it} u_{it2}$. Given we intuit a fixed effect associated with the linear reduced form for y_{it2} is present also, we assume a fixed effect specification as well as no idiosyncratic endogeneity issue (e.g. $E[\mathbf{z}'_{it} c_{i2}] \neq 0$ and $E[\mathbf{z}'_{it} u_{it2}] = 0$ where complete cases are observed $h_{it} = 1$) for the linear reduced form. Following [21] and [17], the correlated random effects (CRE) approach specifies a model for the time constant heterogeneity²¹. We can specify a model for $D(c_{i2} | (h_{it}, h_{it} \mathbf{z}_{it}) : t = 1, \dots, T)$. If we allow $\mathbf{v}$ to be a vector of known functions of $(h_{it}, h_{it} \mathbf{z}_{it})$ that behave as sufficient statistics, including time dummies and time-varying covariates' averages, in each unit spanning the cross-sectional unit, (e.g. $\mathbf{v}_i \equiv (1, T_i, \mathbf{z}_i, \bar{\mathbf{z}}_i)$ where $\bar{\mathbf{z}}_i$ necessarily includes the time averages of the time dummies for the complete cases) then this suffices as an appropriate specification. Thus, $D(c_{i2} | (h_{it}, h_{it} \mathbf{z}_{it}) : t = 1, \dots, T) = D(c_{i2} | \mathbf{v}_i)$.

Following [9], the form $h_{it} c_{i2} = h_{it} \delta_2 + h_{it} \bar{\mathbf{z}}'_i \boldsymbol{\phi}_2 + h_{it} a_{i2}$, $a_{i2} | \mathbf{z}_i \sim N(0, \sigma_{a_{i2}}^2)$ applies with the Mundlak adjustment, which projects the unobserved heterogeneity, c_{i2} , associated with the linear reduced form of the potentially endogenous variable y_{it2} , onto a vector of time averages with an associated intercept and error term with the time-constant errors assumed independent of $\mathbf{z}_i$ for the observed cases. The equation for y_{it2} can be written as

²¹A huge benefit from a modeling perspective is that the CRE approach allows the inclusion of the time-constant variables in $\mathbf{z}_i$ and at the same time delivers the FE estimates on the time-varying exogenous regressors.

$$h_{it}y_{it2} = h_{it}\delta_2 + h_{it}\mathbf{z}_{it}'\mathbf{\Lambda}_2 + h_{it}\bar{\mathbf{z}}_i'\phi_2 + h_{it}a_{i2} + h_{it}u_{it2} \quad (2.3)$$

for complete cases. Note, we displace the intercept constant from $\mathbf{z}_{it}$ for clarity. Thus, we have imposed the Mundlak adjustment device to allow heterogeneity affecting the endogenous regressor, y_{it2} , to be correlated with all exogeneous variables $\mathbf{z}_i$ via corresponding time averages the complete instrument vector $\bar{\mathbf{z}}_i$. Time averages are computed as $h_{it}\bar{\mathbf{z}}_i = \frac{\sum_{t=1}^T h_{it}\mathbf{z}_{it}}{h_{it}T}$. Also, for convenience, we define a composite error term in (2.3) as $h_{it}\nu_{it2} = h_{it}(a_{i2} + u_{it2})$.²²

We turn to explicitly finding a reduced form for c_{i1} in the structural equation via the Mundlak device as before for a pure test of idiosyncratic endogeneity (or time-varying explanatory variables be correlated with idiosyncratic errors/time varying omitted variables). We now project the unobserved heterogeneity in the structural equation c_{i1} onto both $(\bar{\mathbf{z}}_i, \bar{\nu}_{i2})$ via Mundlak's device as the goal is to isolate the source of endogeneity purely attributed to the time-varying unobservables r_{it1} . Following the same convention for the CRE approach for the time-constant heterogeneity c_{i2} associated with the linear reduced form for the potential endogenous regressor, we specify the time constant heterogeneity as $D(c_{i1} | (h_{it}, h_{it}\mathbf{z}_{it1}) : t = 1, \dots, T) = D(c_{i1} | \mathbf{g}_i)$ with $\mathbf{g}_i \equiv (1, T_i, \bar{\mathbf{z}}_{i1}, \bar{\mathbf{z}}_i, \bar{\nu}_{i2}) \equiv (1, T_i, \bar{\mathbf{z}}_{i1}, \bar{\mathbf{z}}_i, \bar{y}_{i2})$. We construct a linear reduced form of c_{i1} as

$$h_{it}c_{i1} = h_{it}\delta_1 + h_{it}\bar{\mathbf{z}}_i'\phi_1 + h_{it}\Phi_1\bar{\nu}_{i2} + h_{it}a_{i1}$$

where $a_{i1} | \mathbf{z}_i \sim N(0, \sigma_{a_{i1}}^2)$ for the complete cases $h_{it} = 1$. This projection is crucial. It is clear that the term a_{i1} (remaining error) is now completely uncorrelated with y_{it2} and $\mathbf{z}_{it}$, $\forall t = (1, \dots, T)$. That is, given the Mundlak device on the form of c_{i1} , we now have $\text{Cov}(\mathbf{z}_i, a_{i1}) = 0$ as well as $\text{Cov}(y_{i2}, a_{i1}) = 0$.

Recalling that ν_{it2} is the composite error associated with y_{it2} , it easily follows that

$$h_{it}\bar{\nu}_{i2} = h_{it}\bar{y}_{i2} - h_{it}\delta_2 - h_{it}\bar{\mathbf{z}}_i^*\mathbf{\Lambda}_2$$
²³

We can plug the projection into the structural equation as

$$\begin{aligned} E[h_{it}y_{it1} | \mathbf{z}_{it1}, y_{it2}, c_{i1}, r_{it1}, t_{it}] &= h_{it}t_{it}\exp(h_{it}\mathbf{z}_{it1}'\beta + h_{it}\varphi_1y_{it2} + h_{it}c_{i1} + h_{it}r_{it1}) \\ &= h_{it}t_{it}\exp(h_{it}\mathbf{z}_{it1}'\beta + h_{it}\varphi_1y_{it2} + h_{it}\delta_1 + h_{it}\bar{\mathbf{z}}_i'\phi_1 + h_{it}\Phi_1\bar{\nu}_{i2} + h_{it}a_{i1} + h_{it}r_{it1}) \\ &= h_{it}t_{it}\exp(h_{it}\mathbf{z}_{it1}'\beta + h_{it}\varphi_1y_{it2} + h_{it}\delta_1 + h_{it}\bar{\mathbf{z}}_i'\phi_1 + h_{it}\Phi_1[\bar{y}_{i2} - \delta_2 - \bar{\mathbf{z}}_i^*\mathbf{\Lambda}_2] + h_{it}a_{i1} + h_{it}r_{it1}) \\ &\equiv h_{it}t_{it}\exp(h_{it}\mathbf{z}_{it1}'\beta + h_{it}\varphi_1y_{it2} + h_{it}\kappa_1 + h_{it}\bar{\mathbf{z}}_i'\eta_1 + h_{it}\Phi_1\bar{y}_{i2} + h_{it}a_{i1} + h_{it}r_{it1}) \end{aligned}$$

²²For the linear reduced form in (2.3), the obvious assumptions are I.) no instruments in $\mathbf{z}_{it}$ are correlated with u_{it2} , II.) no perfect linear dependencies of the covariates in $\mathbf{z}_{it}$ (no multicollinearity), and III.) relevant and strong instruments are needed to explain variation in y_{it2} .

²³Note, $\bar{\mathbf{z}}_i^*$: The asterisk (*) is used to denote the combined vector of time averages explicitly from $\mathbf{z}_{it}$ and the Mundlak mean already defined from (2.3).

Now, a_{i1} is completely uncorrelated with $\bar{\mathbf{z}}_i$ as well as $\bar{y}_{i2}$. It is clear that a_{i1} is uncorrelated with all of the variables in this estimating equation: $(y_{it2}, \bar{y}_{i2}, \bar{\mathbf{z}}_i, \mathbf{z}_{it1})$. Therefore, by the strict exogeneity assumption on $\mathbf{z}_{it}$, r_{it1} is now uncorrelated with the exogenous regressors, $\mathbf{z}_{it1}$, $\bar{y}_{i2}$, and the corresponding time averages, $\bar{\mathbf{z}}_i$ for $h_{it} = 1$. We write $h_{it}\nu_{it1} = h_{it}(a_{i1} + r_{it1})$ as a composite error term from the structural form above (with ν_{it2} defined previously) and now consider the joint distribution of these composite errors.

Assuming that the joint composite errors are independent of $\mathbf{z}_i$ (e.g. $(\nu_{it1}, \nu_{it2}) \perp\!\!\!\perp \mathbf{z}_i$) and have a bivariate normal distribution that is $D \begin{pmatrix} \nu_{it1} \\ \nu_{it2} \end{pmatrix} \Big| \mathbf{z}_i \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix} \right]$ ²⁴, then $E[h_{it}\exp(\nu_{it1})|h_{it}\nu_{it2}] = \exp(h_{it}\pi_1 + h_{it}\theta_1\nu_{it2})$ ²⁵. The systematic endogeneity of y_{it2} is allowed through this specification (the unobserved heterogeneity, ν_{it1} , and the time-varying composite error ν_{it2} .) We now have $E[h_{it}y_{it1}|\mathbf{z}_i, y_{it2}, \nu_{it2}, t_{it}] = \exp(h_{it}\pi_1 + h_{it}\theta_1\nu_{it2})h_{it}t_{it}\exp(h_{it}\mathbf{z}_{it1}'\beta + h_{it}\varphi_1y_{it2} + h_{it}\kappa_1 + h_{it}\bar{\mathbf{z}}_i'\boldsymbol{\eta}_1 + h_{it}\Phi_1\bar{y}_{i2})$. Thus, we can obtain the final estimating equation

$$E[h_{it}y_{it1}|\mathbf{z}_{it}, y_{it2}, \nu_{it2}, t_{it}] = h_{it}t_{it}\exp(h_{it}\alpha_1 + h_{it}\mathbf{z}_{it1}'\beta + h_{it}\bar{\mathbf{z}}_i'\boldsymbol{\eta}_1 + h_{it}\varphi_1y_{it2} + h_{it}\Phi_1\bar{y}_{i2} + h_{it}\theta_1\nu_{it2}) \quad (2.4)$$

for all complete cases. Note, the inclusion of $\bar{y}_{i2}$ and $\bar{\mathbf{z}}_i$ into this equation allows for unobserved heterogeneity, (formerly c_{i1}), to be correlated with all explanatory variables as well as the excluded instruments. Given the Mundlak application and the instruments are strictly exogenous, the idiosyncratic endogeneity test of $H_0 : \theta_1 = 0$.

The process for estimation is a two-step estimation process as

- I. Employ pooled linear probability model (LPM) on y_{it2} on $\mathbf{1}, \mathbf{z}_{it}, \bar{\mathbf{z}}_i$, across t and i to obtain the residuals $\hat{\nu}_{it2}$ for all complete cases ($h_{it} = 1$)²⁶.
- II. Employ pooled Poisson Quasi-Maximum Likelihood Estimation (QMLE) on the mean (exponential) function, with $\bar{\mathbf{z}}_i, y_{i2}, \bar{y}_{i2}, \hat{\nu}_{it2}, \ln(t_{it})$, and $\mathbf{z}_{it1}$ as covariates ($h_{it} = 1$).

We now discuss the specifics of the two-stage estimation process in detail.

Firstly, we discuss the control function methodology. The control function method first partitions the variation of endogenously suspected, binary covariate y_{it2} into two distinct components: (i) exogenous variation that is a result from regression using the strictly exogenous controls and strictly exogenous instrumental variables all contained within $\mathbf{z}_{it} = (\mathbf{z}_{it1}, \mathbf{z}_{it2})$ on the response, y_{it2} , and (ii) the problematic variation that is explicitly correlated with the time-varying unobservables, r_{it1} , from the initial exponential mean function in (2.1). The intuition is that the problematic variation in (ii) is left in the errors or residuals from the first-stage, pooled linear probability regression. In (2.1), the suspected endogeneity in

²⁴Note, the variances here are normalized to 1, and θ represents the covariance between the error terms. That is, if $\text{Cov}(\nu_{it1}, \nu_{it2}) = \theta$.

²⁵Clearly, $\theta_1 = \frac{\text{Cov}(\nu_{it1}, \nu_{it2})}{\text{Var}(\nu_{it2})}$, or the covariate coefficient on the generated regressor ν_{it2} .

²⁶An LPM is a conditional probability model in which the response probability is specified as a linear function in parameters exclusively.

y_{it2} is correlated with the time-varying omissions r_{it1} (e.g. $E[y_{it2}r_{it1}] \neq 0$); yet, through this method, the endogeneity is thought to be significantly captured by the composite error term from the first stage regression $\hat{\nu}_{it2}$ provided we have appropriate instrumental variables that satisfy Assumption 4 and particularly Assumption 5. Therefore, the inclusion of $\hat{\nu}_{it2}$ in the second stage regression for the fully derived structural equation in (2.4) renders y_{it2} exogenous as the suspected endogeneity of y_{it2} (formerly correlated with r_{it1} which no longer exists in the structural equation), is now controlled for by the inclusion of $\hat{\nu}_{it2}$ (the control function). We explicitly test for the statistical significance of $\hat{\nu}_{it2}$, which now contains the endogenous portion of y_{it2} , from the first stage through assessing the magnitude of the coefficient $\theta_1 = \frac{\text{Cov}(\nu_{it1}, \nu_{it2})}{\text{Var}(\nu_{it2})}$ from the second stage pooled Poisson QMLE. This conditional exponential mean in the structural equation including the first stage residual allows for a pure test of idiosyncratic endogeneity in the presence of heterogeneity endogeneity using Mundlak's projection.

For the pooling procedure of both stages, we define the pooled regression models. Pooled regression models applied to any longitudinal data structure essentially disregards the panel element of the data and assumes homogeneous models in all time periods t for each cross-sectional unit i ([3]). The only sufficient assumptions required to utilize this estimation technique in this particular context are: I.) no explanatory variables or instruments are correlated with time-varying unobservables, II.) no perfect linear dependencies between any of the instruments, and III.) there is significant correlation between the instrumental variables and the endogenous culprit, which by our previous assumptions are all attained. Additionally, there is no restriction for serial dependence as the entire estimation procedure for the pooled method, in both stages, can be made impervious to any autocorrelation and heteroscedasticity of unknown form by employing cluster-robust standard errors ([3], [20]). Again, it is worth reiterating the QMLE estimation in (2.4) such that the conditional variance need not be properly specified. Also, as well, the selection indicator $\mathbf{h}_i$ allowing for unbalanced panels cannot be correlated with ν_{it2} in (2.3) ([3], [21], [20]).

Regarding the first stage (I.), we note that time dummies related to our temporal component of the panel, is necessarily included in $\mathbf{z}_{it}$ as well as the corresponding time averages $\bar{\mathbf{z}}_i$. Following [17] and [3], $\mathbf{z}_{it}$ includes aggregate time dummies and $\bar{\mathbf{z}}_i$ include the time averages. Notably, the inclusion of these covariates ensures the coefficient vector β of this pooled LPM estimator is identical/equivalent to the within (or FE) estimator on the unbalanced panel ([42]). As we can attain an identical coefficient vector equivalent to the FE estimator on this unbalanced panel, then we have attained a coefficient vector that is robust to any time-constant endogeneity/heterogeneity and $\mathbf{z}_{it}$ via this pooled method. The intuition of this result is that the time average of the aggregate time variables changes across the units in the given cross-sectional unit i as we average different time periods t for a given i . Also, in this step, it is possible to add time-constant covariates and still yield the coefficient vector of a pooled regression estimate equivalent to the FE parameter vector for Λ_2 as well as, most importantly, additional variation in the residuals $\hat{\nu}_{it2}$. Although the correlated random effects model has been thoroughly explored and the following claim is alluded to in previous literature ([9], [21],[3]), we take time to emphasize the power of this combined approach. The ability to add time-constant

effects to the first stage regression allows access a broader array of potentially effective instruments to further capture variation in a suspected, time-varying endogenous variable that otherwise would be swept out opting to use, for instance, a fixed-effect, instrumental variables (FE-IV) approach.

Regarding the second stage estimation (II.), we employ a pooled Poisson QMLE regression on the mean exponential function including $\bar{\mathbf{z}}_i$, $\ln(t_{it})$, $\bar{y}_{i2}$, $\hat{\nu}_{it2}$, and $\mathbf{z}_{it1}$. This pooled Poisson regression explicitly bases parameter estimation not on the joint distribution of counts $\{y_{i1}, \dots, y_{iT}\}$ for a cross-sectional i , but on the marginal distributions of the individual counts y_{it} . Thus, the resulting Poisson QML estimator, $\hat{\beta}_{PQML}$, is consistent under an assumption that conditional mean is correctly specified without the explicit need for the response to truly be drawn from a Poisson distribution ([54]). These estimates can be suitable when overdispersion is present provided the resulting regression's standard errors are appropriately adjusted for overdispersion. Additionally, we speak to the use of Mundlak's projection in second-stage (II.) of our estimation process. Recalling the fact that the inclusion of including time dummies and time averages results in an identical coefficient vector equivalent to the fixed effects, then applying Mundlak's projection of c_{i1} in the conditional mean function onto $(\bar{\mathbf{z}}_i, \bar{\nu}_{i2})$, also ultimately yields a coefficient vector $\hat{\beta}_{PQML}$ that mimics the Poisson FE estimator (estimated by conditional maximum likelihood estimation); thus, the parameter vector is robust to any correlation between the time-constant endogeneity and $\mathbf{z}_{it1}$ ([8], [55]²⁷, [3]). Thus, the time-constant, heterogeneity endogeneity, formerly c_{i1} , is completely controlled for. Therefore, this enables a pure test of idiosyncratic endogeneity.

Bootstrapping has also been proposed for computing the correct standard errors²⁸ due to the estimated residual vector included in this estimation of the structural equation. However, as we will discuss in subsection 2.3.3, this will not serve as an appropriate correction in our case as we know a crucial assumption regarding this two-step procedure is explicitly violated. Lastly, the test of idiosyncratic endogeneity can be employed on the presumed endogenous variable, y_{it2} , as an asymptotic test on $\theta_1 = 0$ ²⁹. It can be made robust to arbitrary serial correlation and heteroscedasticity in most standard software packages. As we have discussed our testing procedure, we now turn to asymptotics associated with the final estimating equation representing the conditional mean function of our two-step approach.

²⁷It is robust in the specific sense that a FE approach removes the time constant heterogeneity, and this approach mimics an equivalent coefficient vector, which yields estimate that must be robust as well.

²⁸Failing to properly bootstrap the standard errors in the second stage with the included generated regressor included from the first stage leads to many type I errors as the generated regressor itself has sampling variance of its own ([56]).

²⁹Explicitly, we choose the convenient Wald hypothesis test as it is adapted simply in the case of a quasi-maximum likelihood setting, where, under the null hypothesis and a possibly misspecified variance function, the Wald test statistic has a chi-squared distribution with the associated degrees of freedom provided the regularity conditions hold ([57]).

2.3.2 Idiosyncratic Test for Endogeneity - Asymptotics for the Conditional Mean Function

We will now discuss the asymptotic variance specifically regarding our final estimating equation. Recall the final estimating equation is:

$$E[h_{it}y_{it1}|\mathbf{z}_{it}, y_{it2}, \nu_{it2}, t_{it}] = h_{it}t_{it}\exp(h_{it}\alpha_1 + h_{it}\mathbf{z}_{it1}'\boldsymbol{\beta} + h_{it}\bar{\mathbf{z}}_i'\boldsymbol{\eta}_1 + h_{it}\varphi_1y_{it2} + h_{it}\Phi_1\bar{y}_{i2} + h_{it}\theta_1\nu_{it2})$$

For simplicity, let us define the conditional mean as $E[h_{it}y_{it1}|\mathbf{z}_{it}, y_{it2}, \nu_{it2}, t_{it}] = m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*)$ with $\mathbf{q}_{it}^* = (\mathbf{z}_{it1}, y_{it2}, \bar{\mathbf{z}}_i, \bar{y}_{i2}, t_{it}, \hat{\nu}_{it2})$ as the vector of all covariates including our generated regressor from the first stage residuals for the control function procedure and for the selected sample $h_{it} = 1$, and $\boldsymbol{\gamma}^* = (\alpha_1, \boldsymbol{\beta}, \boldsymbol{\eta}_1, \Phi_1, \varphi_1, \theta_1)$ as the parameter vector for the selected sample $h_{it} = 1$. We denote the vector of covariates and corresponding parameter vector with an asterisk to highlight the complete cases, specifically.

Following [57], we derive a simple quasi-maximum likelihood procedure for estimating $\boldsymbol{\gamma}^*$ as well as derive the asymptotic variance for the endogeneity test for pooled estimation of the structural equation under the “fixed T large N ” case. We disregard the time series for each time period t and use a Poisson assumption for the conditional mean. We write down the log-likelihood of the Poisson as if $\mathbf{q}_{it}^*, t = 1, 2, \dots, T$ are independent across t and as if $E[h_{it}y_{it}] = m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*)$. For each i in our cross-section for the complete cases, the quasi log-likelihood for the pooled Poisson estimation is:

$$\begin{aligned} \ell_i(\boldsymbol{\gamma}^*) &= \sum_{t=1}^T h_{it}y_{it}\ln[m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*)] - \sum_{t=1}^T m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*) - \sum_{t=1}^T \ln(h_{it}y_{it}!) \\ &\equiv \sum_{t=1}^T h_{it}y_{it}\ln[m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*)] - \sum_{t=1}^T m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*) \\ &\equiv \sum_{t=1}^T \ell_{it}(\boldsymbol{\gamma}^*) \end{aligned}$$

The pooled Poisson QMLE maximizes the sum of $\ell_i(\boldsymbol{\gamma}^*)$ across $i = 1, \dots, N$ which does not mandate independence across t nor the Poisson assumption. Thus, the pooled Poisson QMLE is robust in the sense that it estimates $\boldsymbol{\gamma}^*$ under the assumption $E[h_{it}y_{it}] = m(\mathbf{q}_{it}^*, \boldsymbol{\gamma}^*)$. Note, this does not necessarily require the strict exogeneity assumption placed on the covariate vector $\mathbf{q}_{it}^*$ which enables any covariate to be included such as endogenous variables as well as lagged dependent variables. Also, this conditional mean form does not accommodate an unobserved fixed effect which poses no issue in our case due to the Mundlak’s adjustment applied to derive the final estimating equation for the conditional mean function previously.

We now proceed with estimating the asymptotic variance of $\hat{\gamma}^*$. Firstly, the limiting distribution of the Poisson QMLE under weak conditions can be written as:

$$\begin{aligned}\sqrt{N}(\hat{\gamma}^* - \gamma^*) &= \left(-N^{-1} \sum_{i=1}^N \mathbf{H}_i(\gamma^*) \right)^{-1} \left(N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\gamma^*) \right) + o_p(1) \\ &= \mathbf{A}^{-1} \left(N^{-1/2} \sum_{i=1}^N \mathbf{s}_i(\gamma^*) \right) + o_p(1)\end{aligned}$$

Let $\mathbf{s}_i(\gamma^*)$ be the score (vector of partial derivatives) of size $K \times 1$ for $\ell_i(\gamma^*)$, where $\mathbf{s}_i(\gamma^*) = \sum_{t=1}^T \mathbf{s}_{it}(\gamma^*)$, and where $\mathbf{s}_{it}(\gamma^*)$ is the score of $\ell_{it}(\gamma^*)$. Note, $\mathbf{s}_{it}(\gamma^*) = \nabla_{\gamma^*} m(\mathbf{q}_{it}^*, \gamma^*)' (h_{it} y_{it} - m(\mathbf{q}_{it}^*, \gamma^*)) / m(\mathbf{q}_{it}^*, \gamma^*)$ ³⁰. The matrix $\mathbf{A}$ is the expected value of the negative Hessian (e.g. $E[-\mathbf{H}_i(\gamma^*)]$) of $\ell_i(\gamma^*)$ for the cases used in estimation. The Hessian, $\mathbf{H}_i(\gamma^*)$, is a symmetric $K \times K$ matrix of second-order partial derivatives which results in the definition of $\mathbf{A}$ as

$$\mathbf{A} = \sum_{t=1}^T E[\nabla_{\gamma^*} m_{it}(\gamma^*)' \nabla_{\gamma^*} m_{it}(\gamma^*) / m_{it}(\gamma^*)]$$

This matrix can consistently be estimated by $\tilde{\mathbf{A}} = N^{-1} \sum_{i=1}^N \sum_{t=1}^T \nabla_{\gamma^*} \hat{m}_{it}' \nabla_{\gamma^*} \hat{m}_{it} / \hat{m}_{it}$. A consistent estimator of $\mathbf{B}$, the variance of the score, is $\tilde{\mathbf{B}} = E[\mathbf{s}_i(\gamma^*) \mathbf{s}_i(\gamma^*)'] = N^{-1} \sum_{i=1}^N \mathbf{s}_i(\hat{\gamma}^*) \mathbf{s}_i(\hat{\gamma}^*)'$ for complete cases.

Therefore, the limiting distribution of the Poisson QMLE for the values used in estimation can be re-written in more convenient notation as

$$\sqrt{N}(\hat{\gamma}^* - \gamma^*) \sim \mathbf{N}(\mathbf{0}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1})$$

This results in an estimate of $\text{avar}(\hat{\gamma}^*)$ as $\text{avar}(\hat{\gamma}^*) = \tilde{\mathbf{A}}^{-1} \tilde{\mathbf{B}} \tilde{\mathbf{A}}^{-1} / N$ which is a symmetric matrix whose terms shrink to zero at a rate to $1/N$. Note, the same arguments apply to deriving the asymptotic variance of the structural equation without the selection h_{it} . Beneficially, this estimator achieves consistency as well as asymptotically normality as shown above.

We derived our test under the assumption the suspected endogenous variable is indeed discrete. We now turn to exploring the added complication of having non-continuous endogenous covariate utilizing our two-step approach.

2.3.3 Idiosyncratic Endogeneity Test: Form of Endogenous Regressor

We emphasize a note on the pooled regression employed in the first step regression given the form of the endogenous covariate y_{it2} . If y_{it2} is continuous, the corresponding residuals from the pooled, first-stage ordinary least squares regression, $\hat{\nu}_{it2}$, can likely be assumed reasonably homoscedastic-normal (provided

³⁰The notation ∇_{γ^*} is taken to mean the partial derivative with respect to γ^*

strong and relevant instruments). More importantly, however, is that the independence assumption of the residuals from $\mathbf{z}_{it}$ is quite likely. However, if y_{it2} is binary, as in our application and as our main focal point for our testing procedure, the resulting pooled regression of y_{it2} on $\mathbf{1}, \mathbf{z}_{it}, \bar{\mathbf{z}}_i$, across t and i for the residuals $\hat{\nu}_{it2}$ in (I.) are estimated via a pooled linear probability model (LPM), as previously stated. It is noteworthy to briefly review the LPM.

If y_{it2} is binary, then it is natural to interpret its expected value as a probability. Given a random sample, the unconditional probability that $y_{it2} = 1$ is equal to its unconditional expected value as (given $\mathbf{h}_i$)

$$P(h_{it}y_{it2} = 1) \equiv E[h_{it}y_{it2}]$$

Following from (2.3), this implies that the conditional probabilities are

$$P(h_{it}y_{it2} = 1|\mathbf{z}_{it}, \bar{\mathbf{z}}_i) \equiv E[h_{it}y_{it2} = 1|\mathbf{z}_{it}, \bar{\mathbf{z}}_i]$$

Assuming $\nu_{it2} = a_{i2} + u_{it2}$ is mean zero, then under a correct model specification, we have

$$\begin{aligned} P(h_{it}y_{it2} = 1|\mathbf{z}_{it}, \bar{\mathbf{z}}_i) &\equiv h_{it}\delta_2 + h_{it}\mathbf{z}_{it}'\mathbf{\Lambda}_2 + h_{it}\bar{\mathbf{z}}_i'\phi_2 \\ P(h_{it}y_{it2} = 0|\mathbf{z}_{it}, \bar{\mathbf{z}}_i) &\equiv 1 - (h_{it}\delta_2 + h_{it}\mathbf{z}_{it}'\mathbf{\Lambda}_2 + h_{it}\bar{\mathbf{z}}_i'\phi_2) \end{aligned}$$

Therefore, the probability of success is a linear function of instrumental variables and corresponding time averages. This poses a further issue. Given y_{it2} is binary, the independence assumption made of the residuals from the instruments is quite unlikely to be true ([53], [57]) as well as the homoscedastic-normal assumption of the residuals resulting from the linear reduced form of y_{it2} .

If ν_{it2} is defined as the associated time-varying error term of (2.3) and recalling the variance of a Bernoulli random variable (e.g. For $Y \sim \text{Bern}(p)$, then $\text{Var}(Y) = P(Y = 1) * (1 - P(Y = 1)) \equiv P(Y = 1) * P(Y = 0)$), then the conditional variance of the error term is

$$\begin{aligned} \text{Var}(h_{it}\nu_{it2}|\mathbf{z}_{it2}, \bar{\mathbf{z}}_i) &= \text{Var}(h_{it}y_{it2} - (h_{it}\delta_2 + h_{it}\mathbf{z}_{it}'\mathbf{\Lambda}_2 + h_{it}\bar{\mathbf{z}}_i'\phi_2)|\mathbf{z}_{it2}, \bar{\mathbf{z}}_i) \\ &= \text{Var}(h_{it}y_{it2}|\mathbf{z}_{it}, \bar{\mathbf{z}}_i) \\ &= P(h_{it}y_{it2} = 1|\mathbf{z}_{it}, \bar{\mathbf{z}}_i) * P(h_{it}y_{it2} = 0|\mathbf{z}_{it}, \bar{\mathbf{z}}_i) \\ &= [h_{it}\delta_2 + h_{it}\mathbf{z}_{it}'\mathbf{\Lambda}_2 + h_{it}\bar{\mathbf{z}}_i'\phi_2] * [1 - (h_{it}\delta_2 + h_{it}\mathbf{z}_{it}'\mathbf{\Lambda}_2 + h_{it}\bar{\mathbf{z}}_i'\phi_2)] \\ &\neq \sigma_\nu^2. \end{aligned}$$

As seen above, the construction of the variance of the error terms resulting from linear probability models reveals that the variance is dependent upon a covariate vector. which indicates that the variance of

the error changes as the covariate vector changes. This results in inherent heteroscedasticity (as opposed to homoscedasticity). Therefore, the violation of the assumption of homogeneous finite variance is imminent. Additionally, when y_{it2} takes on two values, $y_{it2} = \{0, 1\}$, then the resulting residuals produced from the vector of any explanatory covariates or instruments can only take on two values, resulting in a non-Gaussian distribution of residuals. These two violations will result in likely inconsistency for our parameter vector β_{PQML} ([58], [41]). To refer back to the ending of Section 2.3.1, the suggested bootstrapping procedure proposed in [3] is futile as y_{it2} is of a non-continuous form (and thus cannot yield a residual vector that is both homoscedastic and non-deterministic of $\mathbf{z}_{it}$ as shown above), which further emphasizes that this is simply a proposed test for idiosyncratic endogeneity given a binary regressor and not suitable correction for appropriate and consistent parameter estimates. It would be extremely unlikely to assume that binary y_{it2} can be written as a strict linear equation with an additive independent error; yet, a more appropriate setup as in [43] could be more realistic. Therefore, what is the utility of this test assuming a discrete y_{it2} ? As stated by [21], this approach is still perfectly valid and a robust test of idiosyncratic endogeneity. Under the null hypothesis, y_{it2} is presumed exogenous. This allows us the flexibility of computing the residual vector in any way regardless of the structure of y_{it2} as this test simply seeks to find a linear combination of y_{it2} and $\mathbf{z}_i$ appearing in the structural form of the final estimating equation. The test is fully robust as it does not matter how the residuals are computed under the null for the purpose of finding a limiting chi-square distribution; none of the assumptions regarding the bivariate nature of the residuals need to hold for seeking the linear combination between $(y_{it2}, \mathbf{z}_{it})$, and any inference made on any covariates (including the control function) of interest can be made robust to the effects of serial correlation and heteroscedasticity of unknown form provided the reasonable assumption of independence between the cross-sectional units i .

The entire construction of the procedure described previously is simply a idiosyncratic test for a binary endogenous regressor, y_{it2} . If $H_0 : \theta_1 = 0$ is rejected, in order to potentially achieve consistency for the parameter vector of interest, we must rely on another estimator beyond the combined correlated random effects and control function approach due to the binary nature of y_{it2} . Conveniently, we have an estimator, the instrumental variables, generalized method-of-moments (IV-GMM) estimator, which, under a specific set of moment conditions, can yield asymptotically consistent estimates given valid instruments. The true benefit of this estimator will mitigate the endogeneity bias that could exist in y_{it2} regardless of whether this regressor is continuous, binary, or discrete ([21], [59], [57]). This method also does not require the restrictive assumption of the linear reduced form for y_{it2} nor the requirement that the joint distribution of the composite residual terms be independent from our exogenous instruments.

In our next subsection, as an additional novelty or contribution to the literature, we perform a Monte Carlo simulation analysis to assess whether detecting idiosyncratic endogeneity in the presence of time-constant heterogeneity of a binary regressor via this two-staged approach is effective given a nonlinear, unobserved effects panel data model under three differing design structures: an unbalanced panel structure with selection, an unbalanced panel structure without selection, and a balanced panel. For clarity, this

balanced case exactly refers to the proposed case in the working paper by [21], and we extend this approach to the two cases of unbalanced panels. We are particularly interested in assessing our results under “weak” instrumental variables versus “strong” instrumental variables in all cases as we intuit that instrument strength is crucial in this testing procedure. Rejection rates of the test of endogeneity given instrument strength is our primary focus. Additionally, we will evaluate the nominal size and statistical power (primarily contingent upon effect size) of this test as well in each of our simulations.

2.3.4 Monte Carlo Simulation - Endogeneity Test: Data Generating Processes

We present the pure test of idiosyncratic endogeneity. Firstly, recall the structural equation as well as the linear reduced form for our binary endogenous variable:

$$E[y_{it1}|\mathbf{z}_i, y_{it2}, t_{it}, c_{i1}, r_{it1}, \mathbf{h}_i] = E[y_{it1}|\mathbf{z}_{it1}, y_{it2}, t_{it}, c_{i1}, r_{it1}] = t_{it} \exp(\mathbf{z}_{it1}'\boldsymbol{\beta} + \varphi_1 y_{it2} + c_{i1} + r_{it1})$$

$$y_{it2} = \mathbf{z}_{it}'\boldsymbol{\Lambda}_2 + c_{i2} + u_{it2}$$

where $E[\mathbf{z}_{it1}'c_{i1}] \neq 0$, $E[\mathbf{z}_{it}'c_{i2}] \neq 0$, and $E[y_{it2}u_{it2}] = 0$ are assumed (the fixed-effects assumption as well as no idiosyncratic endogeneity assumed in this reduced form). For the time-constant fixed effects in both equations, we explicitly draw iid observations for both time-invariant unobserved heterogeneity terms: $c_{i1} \sim N(0, 1)$ and $c_{i2} \sim N(0, 1)$.

For the linear reduced form for the binary endogenous variable³¹, we have:

$$y_{it2}^* = \delta_2 + \mathbf{z}_{it}'\boldsymbol{\Lambda}_2 + c_{i2} + \nu_{it2}^*$$

$$\text{where } \mathbf{z}_{it}' = (x_{it1}, x_{it2}, w_{it1}, w_{it2}) \text{ and } x_{it1} \sim N(0, 1) * .5 + .8 * c_{i1}, \quad x_{it2} = \begin{cases} 1, & U(0, 1) * c_{i1} \geq 0.5 \\ 0, & U(0, 1) * c_{i1} < 0.5 \end{cases},$$

$w_{it1} \sim \chi_1^2 * c_{i2}$, $w_{it2} \sim N(0, 1) * c_{i2}$. Therefore, $\mathbf{z}_{it}$ contains a continuous exogenous regressor, a binary exogenous regressor, which are both allowed to be correlated explicitly with the time-invariant heterogeneity in the structural equation, along with two instrumental variables (w_{it1}, w_{it2}) that must be excluded from the structural equation (and allowed to be correlated with the resulting time-invariant heterogeneity from the linear reduced equation) as well as arbitrary time dummies³². We simulate instrumental variables under two scenarios: “weak” instrumental variables and “strong” instrumental variables³³. Following similarly from

³¹We explicitly reform the equation with the intercept coefficient shown and $\mathbf{1}$ displaced from $\mathbf{z}_{it}$.

³²The exact number of time dummies included will be explained in the panel structure choices. Additionally, these time dummies need not have simulated coefficients associated with them.

³³As a technical aside, the idea of weak vs. strong IVs of this two-stage process follows primarily from [60]. For valid IVs, they must meet the criteria of both being completely uncorrelated with the error term in the structural equation, yet strongly associated with the endogenous regressor itself. Thus, in setting our coefficients for all scenarios, we choose coefficients for weak IVs in such a way that guarantees the mean of the F-statistics in the first-stage, linear probability model (LPM) is necessarily smaller than the F-statistics of the strong IVs. Specifically, in the case of strong IVs, the mean F-statistics are well over 10

[25], we directly specify the parameter vector corresponding to the controls and instruments in the linear reduced form and conditional mean function to induce instrument strength. The parameter vector for our weak IV scenario is $\mathbf{\Lambda} = (0.2, 0.2, 0.1, 0.3)'$ with $\delta_2 = -1$. The parameter vector for our strong IV scenario multiplies the coefficients of our instruments by 10 resulting in $\mathbf{\Lambda} = (0.2, 0.2, 1, 3)'$ with $\delta_2 = -1$, as before. We simulate the binary dependent variable y_{it2} as follows

$$y_{it2} = \begin{cases} 1, & \text{if } y_{it2}^* > 0 \\ 0, & \text{if } y_{it2}^* \leq 0 \end{cases}$$

As the Poisson assumption for the response is quite restrictive, we consider a more general DGP with the shared structure on the conditional expectation. For simulating the dependent variable in the structural equation, we have

$$y_{it1}^* = t_{it}^* \exp(\alpha_1 + \mathbf{z}_{it1}'\boldsymbol{\beta} + \varphi_1 y_{it2} + c_{i1} + \nu_{it1}^*)$$

where $\mathbf{z}_{it1} = (x_{it1}, x_{it2})^{34}$, $t_{it}^* \sim U[0, 1] * 99 + c_{i1}$, $t_{it} = \lceil t_{it}^* \rceil^{35}$, and $y_{it1} = \lceil y_{it1}^* \rceil$. As before, we set $\boldsymbol{\beta} = (0.2, 0.2)'$ and the coefficient on our binary endogenous variable is $\varphi = 0.2$. The constant is set to $\alpha_1 = 1$. The error terms are generated jointly as $D \begin{pmatrix} \nu_{it1}^* \\ \nu_{it2}^* \end{pmatrix} \middle| \mathbf{z}_i \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta_0 \\ \theta_0 & 1 \end{pmatrix} \right]^{36}$. We now will discuss the three panel structures for consideration in our combined correlated random effects/control function approach for idiosyncratic endogeneity³⁷.

1. Design 1(a) - Unbalanced Panel with selection: For the primary design, each cross section

i is observed to time $T = 5$. For the selection indicator, to introduce the unbalancedness, $\mathbf{h}_i =$

$\{h_{i1}, \dots, h_{iT}\}$, we draw $\mathbf{h}_i^* \sim U[0, 1] * c_{i1}$ where $\mathbf{h}_i = \begin{cases} 1, & \text{if } \mathbf{h}_i^* \geq 0 \\ 0, & \text{if } \mathbf{h}_i^* < 0 \end{cases}$ which indicates if the entire

row vector $(y_{it}, \mathbf{z}_{it}, t_{it}, y_{it2})$ is fully observed and can be used for estimation and 0 otherwise³⁸. This

will result in unbalancedness and the resulting value, T_i , corresponds to the number of time dummies

in all three panel structures, which is the suggested, indicative value for instrument relevancy. We will present the average F-statistics for all scenarios in our results section.

³⁴Note, again, $\mathbf{z}_{it1}$ includes time dummies which corresponds to the same time dummies used in the linear reduced form for y_{it2}^* .

³⁵Note, although not necessary, t_{it} of course serves the purpose of an offset.

³⁶For the DGPs which constitutes the creation of y_{it1} and y_{it2} , these errors (and responses) are denoted with an asterisk as they are not yet composed of the time-invariant error substitution resulting from the linear projections from Mundlak's method onto the time averages of all exogenous variables. That is, we construct the response variable without the explicit substitution of the linear projections onto time averages; however, yet instead, we allow their construction with the "equivalent" time-invariant fixed effects c_{i1} and c_{i2} . This construction has been used as in [61].

³⁷As a final aside regarding the instrumental variables, clearly, (w_{it1}, w_{it2}) are deemed valid instruments as they are directly correlated with y_{it2} by the construction of the linear reduced form of the suspected endogenous regressor, yet completely uncorrelated with the source of endogeneity in the structural equation represented by the conditional mean y_{it1} .

³⁸Note, this selection indicator extends to both the linear reduced form for the binary endogenous regressor as well as the conditional exponential mean function in the structural equation. That is, the same observations for each cross section i for time T_i is used in each estimation step of the full two-step process.

less 1³⁹ for each i in covariate/instrument vectors in both the linear reduced form and the structural equation.

2. **Design 1(b) - Unbalanced Panel without selection:** For this unbalanced design, each cross section i can be observed **at most up to time** $T_i = 5$. The value of T_i corresponds to the number of time dummies less 1 for each i in covariate/instrument vectors in both the linear reduced form and the structural equation.⁴⁰
3. **Design 1(c) - Balanced Panel:** This is a purely balanced design where, for each cross section i , we observe these observations all for $T = 5$. This corresponds to the number of time dummies less 1 for each i in covariate/instrument vectors in both the linear reduced form and the structural equation. This balanced design will explicitly serve as our benchmark as it directly mimics the proposed procedure by [21].

For our idiosyncratic endogeneity test using our combined Correlated Random Effects/Control Function approach, we estimate our first stage regression of our binary regressor including both the time dummies as well as the time averages for the production of the generated regressor corresponding to the first stage error and include this into our second stage Poisson QMLE regression of the conditional mean. Recall, the conditional exponential mean function will also include the time dummies, time averages for all exogenous variables/instruments (including the averages associated with the time dummies), and the inclusion of cross-sectional average of the potentially endogenous binary regressor, $\bar{y}_{i2}$, as outlined in the prior section. For each of the three designs, we simulate $i = 1500$ observations with the number of Monte Carlo replications as $R = 5000$. For assessing endogeneity, we manipulate the source of covariance between our jointly distributed error terms with a range of values as $\theta_0 = \{0, 0.5, 0.7, 0.9\}$ for the explicit second stage estimation of the conditional mean function

$$E[y_{it1} | \mathbf{z}_{it}, y_{it2}, \nu_{it2}] = t_{it} \exp(\alpha_1 + \mathbf{z}_{it1}' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\eta}_1 + \varphi_1 y_{it2} + \Phi_1 \bar{y}_{i2} + \theta_0 \nu_{it2})$$

When $H_0 : \theta_0 = 0$, this indicates no endogeneity as there is no covariance (or, in this case, synonymously, correlation as the variances are normalized to 1) between the error terms in the errors of the linear reduced form and the structural equation, allowing us to assess the size of the different tests for idiosyncratic endogeneity with a nominal size of $\alpha = 5\%$ in each of the three designs. Additionally, we will assess the power of these endogeneity tests where $\theta_0 = -0.25$ with errors drawn with positive covariance. For inference, we will use cluster-robust standard errors to account for the two-stage estimation process.

³⁹The number of time dummies is subtracted by 1 in order to avoid the “dummy trap” to avoid the additional dummy being entangled with the intercept resulting in collinearity.

⁴⁰This is accomplished by expanding each cross-section by at most: $5 + \text{round}(-5 * U[0, 1])$.

2.3.5 Monte Carlo Simulation Results

The results for our three designs can be displayed in tables below. We first direct our attention to the results highlighting the combined CRE/CF approach for testing endogeneity in our unbalanced design with selection displayed in Table 2.2. We see that when idiosyncratic endogeneity is induced via a non-zero, positive covariance/correlation from the random draws of bivariate distributed errors, our simulation given this design was able to detect, to varying degrees, the induced idiosyncratic endogeneity quite well. This is achieved by estimating the conditional mean function with the proposed (second-staged) pooled Poisson QMLE with both weak IVs and strong IVs. Intuitively, for the CRE/CF tests in all three scenarios, the higher the positive covariance/correlation of the errors, the larger the rejection rate of exogeneity. As a higher degree of correlation/covariance is induced between errors in both equations, indicating a larger degree of endogeneity, our proposed approach rejects at a higher rate. Note the disparity between the rejection rates or proportion of times exogeneity is rejected given the case of weak versus strong IVs. When the errors are positively correlated at .5, the null hypothesis of exogeneity ($\theta_0 = 0$) was rejected at a rejection rate of 14% out of 5000 replications (the proportion of times out 5000 iterations the null was rejected), at an $\alpha = 5\%$ level, estimating the conditional mean function with Poisson QMLE in the weak IVs case, yet, at 78%, with our strong IVs. when the bivariate errors were drawn with $\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$ and strong IVs are employed, the rejection rate is 93.2% out of 5000 replications compared to 30% with the weaker IVs highlighting the disparity in results given IV strength. These results are promising and indicative of correctly identifying idiosyncratic endogeneity once controlling for unobserved, time-constant heterogeneity and ignorable selection with instrumental variables that are strongly associated with our potentially endogenous regressor, yet heavily dependent on the proper exclusion restrictions of the chosen IVs. This notion highlights the importance of instrumental variables that are indeed relevant.

Regarding the nominal size of our test, or the probability of rejecting a true null hypothesis, we note that second-stage conditional mean estimations in both scenarios are in line with the null hypothesis, albeit with a bit of slight under-rejection in the Strong IV case given a nominal $\alpha = 0.05$. Lastly, we assess the statistical power of our test, or the likelihood of correctly rejecting a false null in Design 1(a) given our sample size and chosen α . With our positively correlated errors and setting $\theta_0 = -0.25$, we see that our test is convincingly and increasingly able to detect the falsity of idiosyncratic endogeneity induced through negative correlation (as $\theta_0 = -0.25$ suggests), when, in reality, the errors are drawn with positive covariance. As the trajectory of positive covariance/correlation of error terms increases, the more convincingly the test is able to detect such falsity, particularly when employing strong IVs. In the case of $\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$, the tests were able to correctly reject a false null with a probability of 0.98 in the case of strong IVs and only marginally well with weak IVs with a probability of 0.376. As we can see in the subsequent tables for the remaining panel designs, unbalanced panel without selection as represented in Table 2.3 and the balanced design with $T = 5$ in Table 2.4, we see similar results regarding our tests ability to detect idiosyncratic endogeneity in both scenarios.

The rejection rates generally increase in our simulations as higher levels of correlation between error terms were induced. The idiosyncratic endogeneity test in these scenarios using weak IVs are still clearly very ineffective in aiding in detection compared to strong IV by very large margins. Also, the nominal test sizes for our α in Tables 2.3 and 2.4 are notably appropriate and expected with regards to our test. The statistical power based on effect size of both of these designs are quite impressive as well. Overall, this approach seems to suggest that the CRE/CF approach has utility in correctly identifying idiosyncratic endogeneity when present in all three panel structures presented here, and our explicitly contribution of employing this test given two unbalanced scenarios closely matches the benchmark proposed by [21]. Lastly, we assess the average coefficients of the two control regressors, the endogenous regressor, and the first stage LPM F-statistics as an additional check for our simulation results in Table 2.5. Regarding the F-stage statistics, we see that in all 3 cases for our panel structures, the mean F-stats for all Strong IV cases exceeds the Weak IV case by at least twice as much. For our coefficients, recall, we set the coefficient for the continuous regressor x_{it1} as $\beta = 0.20$, the coefficient for x_{it2} , our binary regressor, as $\beta = 0.20$, and the coefficient the endogenous regressors y_{it2} as $\varphi = 0.20$. For all three designs, we note that the mean coefficients for our controls in 5000 replications seem to be relatively close to the specified value from our data generating processes. There is slight variability, however, of some overestimation and underestimation to the specified values of our controls. Most important, though, in all three designs, is the mean coefficient value of the binary endogenous variable. The mean value in all scenarios seems to be severely undersized (or far away from 0.20, generally) in both weak and strong IV cases, indicating likely demonstrably unrectifiable bias and perhaps likely inconsistency given our designs.

On average, there seems to be a more adequate approximation to the true coefficient value in the strong IV cases relative to the alternative. This seems to suggest that stronger IVs tend to help mitigate the coefficient biases of the endogenous binary regressor relative to weaker IVs, but results in inappropriately sized coefficient estimation regardless. Also, although the associated standard errors were not bootstrapped for correct inference and statistical significance for covariates in our case (recommended in the previous section due to residual inclusion from our control function approach), the adjustment would clearly not affect the mean of the coefficient values for the replications. This gives further evidence of the inappropriateness of utilizing the first-stage, linear probability model (LPM) on the binary endogenous regressor for the combined correlated random effects/control function approach in the conditional mean function as the assumptions are not met for joint bivariate assumption of the error terms and its independence from our exogenous variables as a full correction. However, as stated previously, the CRE/CF approach here is still quite appropriate for the test of idiosyncratic endogeneity itself in the presence of heterogeneity endogeneity (time-constant endogeneity), particularly with unbalanced panels distinctly prone to selection and no selection cases as evidenced by the results and estimates in Tables 2.2, 2.3, 2.4, and 2.5. With these results, we will now explore a nonlinear, IV-GMM procedure for a panel discussed next as an avenue for achieving consistent estimates in the next section.

Table 2.2: Design 1(a) - Unbalanced Panel with Selection - Monte Carlo Simulation Results

CRE/CF Test for Endogeneity	Poisson QMLE - Weak IV (RR)	Poisson QMLE - Strong IV (RR)
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5$	0.14	0.78
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7$	0.228	0.864
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$	0.3	0.932
Nominal Size of Test (5%)	Poisson QMLE - Weak IV (Prob)	Poisson QMLE - Strong IV (Prob)
$\theta_0 = 0$	0.052	0.038
Statistical Power	Poisson QMLE - Weak IV (Prob)	Poisson QMLE - Strong IV (Prob)
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5, \theta_0 = -0.25$	0.196	0.752
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7, \theta_0 = -0.25$	0.28	0.944
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9, \theta_0 = -0.25$	0.376	0.98

Table 2.3: Design 1(b) - Unbalanced Panel without Selection - Monte Carlo Simulation Results

CRE/CF Test for Endogeneity	Poisson QMLE - Weak IV (RR)	Poisson QMLE - Strong IV (RR)
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5$	0.132	0.712
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7$	0.28	0.92
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$	0.392	0.972
Nominal Size of Test (5%)	Poisson QMLE - Weak IV (Prob)	Poisson QMLE - Strong IV (Prob)
$\theta_0 = 0$	0.046	0.047
Statistical Power	Poisson QMLE - Weak IV (Prob)	Poisson QMLE - Strong IV (Prob)
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5, \theta_0 = -0.25$	0.184	0.899
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7, \theta_0 = -0.25$	0.388	0.96
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9, \theta_0 = -0.25$	0.512	0.988

Table 2.4: Design 1(c) - Balanced Panel - Monte Carlo Simulation Results

CRE/CF Test for Endogeneity	Poisson QMLE - Weak IV (RR)	Poisson QMLE - Strong IV (RR)
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5$	0.292	0.864
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7$	0.416	0.972
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$	0.604	0.998
Nominal Size of Test (5%)	Poisson QMLE - Weak IV (Prob)	Poisson QMLE - Strong IV (Prob)
$\theta_0 = 0$	0.04	0.048
Statistical Power	Poisson QMLE - Weak IV (Prob)	Poisson QMLE - Strong IV (Prob)
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5, \theta_0 = -0.25$	0.348	0.972
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7, \theta_0 = -0.25$	0.536	0.992
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9, \theta_0 = -0.25$	0.696	0.998

Table 2.5: Monte Carlo Results: Mean Coefficient and F-Statistics Estimates

Design 1(a)								
	Poisson QMLE - Weak IV				Poisson QMLE - Strong IV			
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\varphi}_1$	F	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\varphi}_1$	F
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5$	0.190799	0.2078236	0.048512	8.17783	0.2071202	0.2211091	0.1545175	21.77778
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7$	0.1773231	0.1992557	0.2874633	9.17783	0.197887	0.2038951	0.1293719	22.10428
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$	0.1863547	0.1851858	0.1074751	9.394507	0.1898469	0.1961161	0.1425634	22.799
Design 1(b)								
	Poisson QMLE - Weak IV				Poisson QMLE - Strong IV			
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\varphi}_1$	F	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\varphi}_1$	F
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5$	0.1819423	0.2019428	0.3532889	15.4544	0.2033725	0.2185424	0.1350968	39.33982
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7$	0.1832131	0.2104009	0.1779756	16.00481	0.1952411	0.2256341	0.1525228	39.33833
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$	0.1812587	0.2148332	0.097009	16.05151	0.1898075	0.2197119	0.1264474	39.71649
Design 1(c)								
	Poisson QMLE - Weak IV				Poisson QMLE - Strong IV			
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\varphi}_1$	F	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\varphi}_1$	F
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.5$	0.1949151	0.2123987	0.0374947	33.11585	0.1983426	0.2153993	0.1750732	78.71061
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.7$	0.1842371	0.2146333	0.1237842	32.51317	0.1951561	0.1907651	0.1666198	76.20364
$\text{Cov}(\nu_{it1}^*, \nu_{it2}^*) = 0.9$	0.1782523	0.1933755	0.1382255	32.39034	0.1956517	0.2015236	0.1454183	77.86736

2.4 Generalized Method of Moments: Construction of Three Feasibly Efficient Estimators

The most seminal work regarding the generalized method-of-moments (GMM) estimator, [62], provides a cornerstone definition for the estimator (or class of estimators) that, conveniently, absorbs or subsumes a variety of econometric estimators as well as desired asymptotic results in large samples. GMM is an optimization that sets a family of orthogonality conditions imposed for testing certain restrictions. For estimating a vector population parameters, a vector of estimators are chosen such that the sample analogs of corresponding population moment conditions are as close to zero in their respective population values. Under arbitrary regularity conditions and a given weighting matrix, a vector of parameter estimates minimizes a quadratic objective function of the difference in sample and population moments. In the case of nonlinear GMM estimation, the procedure mimics the standard GMM estimator except that the moment conditions specified are nonlinear functions of the parameter vector. Advantageous aspects of GMM estimation are attributed to flexible identification, great large sample properties as the estimators are strongly consistent, and the benefit offered by the lack of having to specify parametric distributions for an associated error vector of the model. The GMM method offers an invaluable utility in both linear and non-linear cross-sections and panel data (both of dynamic and static forms). A pivotal work, [63], proposes the generalized method-of-moments estimation technique for count panel data models with multiplicative heterogeneity under a nonlinear conditional mean specification while relaxing the strict exogeneity assumption of explanatory variables. Evaluating the inconsistency of a probit model with an endogenous regressor is also evaluated ([64]). Another principal study that framed the foundation for the instrumental variables, two-stage nonlinear GMM approach is presented in [65], where a system of simultaneous equations are modeled, which is nonlinear in both parameters and covariates conjointly, that are contaminated by endogeneity with additive errors. Another work explores the consideration of a class of GMM estimators that is devoid of imposing a distribution of the count response that requires only the correct conditional mean specification via quasi-conditional maximum likelihood estimation ([66]). Additional popular works of this procedure involving creating and evolving the “difference-GMM” methodology in the presence of endogeneity are investigated by both [12] and [67] with multivariate time series considerations (VARs). Other previous literature for utilizing GMM methods to mitigate endogeneity are vast. For example, in estimating the relationship between governance and performance, [68] uses GMM methods to control for both time constant and time related endogeneity for a panel of 6,000 firms. Prior work ([69]) springboards off the aforementioned literature ([68]) by assessing the same scope of the association between company and governance performance of the ASX 200 index at any time from 2000 to 2007.

As our framework does not center around a dynamic effect, we briefly more literature directly related to our application utilizing GMM for nonlinear functions. Another central piece of literature for panel count data using GMM estimation in the presence of time-constant heterogeneity, predetermined regressors, endogenous regressors, and dynamic models is well-explored with the open-ended GAUSS software *ExpEnd* for both additive and multiplicative model specifications in addition to asymptotics ([70],[71]). These authors' work explores mitigating endogeneity bias utilizing a exponential fractional regression model (EFRM) for structural parameters identification with additive errors utilizing a quasi-differencing approach for fixed effects removal with an application to firm decisions [4]. Beneficially, and unlike the control function approach, this model can accompany endogenous regressors without the obligation of assuming a linear reduced form of the suspected endogenous covariate nor any parametric assumptions about the error terms. Lastly, [72] assesses the relationship between labor mobility and patent propensity in a firm-level panel context utilizing GMM methods for endogeneity correction of R&D expenditures as well as time-varying mobility of firm employees.

As directly applicable for our application, we rely on utilizing this estimation technique in order to mitigate the potential endogeneity bias of a binary endogenous variable in order to achieve consistent parameter estimation. The GMM estimators is dependent upon correctly specified moment conditions, eschewing the mandatory maximum likelihood estimation's reliance on assumptions of the distributional form of the response. A beneficial consequence, in this case, is a clearer level of robustness to functional form and less sensitivity to arbitrary requirements imposed by parametric assumptions at the risk of possible efficiency loss. Again, conveniently differing from the CRE/CF approach, we are not restricted to assuming a linear reduced form for the binary endogenous regressor and simply rely on a standard rank condition for identification of the (exponential) conditional mean ([59]).

As before, we begin with the multiplicative error model specification with an unobserved heterogeneity term entered multiplicatively as⁴¹ as well as time-varying unobservables.

$$\begin{aligned}
y_{it} &= \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i + r_{it}) & i = 1, \dots, N \quad t = 1, \dots, T_i \\
&= \exp(\mathbf{x}_{it}'\boldsymbol{\beta})\exp(c_i)\exp(r_{it}) \\
&= \mu_{it}\alpha_i v_{it}
\end{aligned} \tag{2.5}$$

where $\alpha_i = \exp(c_i)$ is the time-constant, unobserved heterogeneity, and $v_{it} = \exp(r_{it})$ denotes the time-varying unobservables⁴² This form assumes that $E[\exp(r_{it})] = E[v_{it}] = 1$ such that $E[y_{it}|\mathbf{x}_{it}, \alpha_i] = \mu_{it}\alpha_i$ ⁴³.

⁴¹As before, an offset can be included in $\mathbf{x}_{it}$ as its a covariate whose estimated parameter is constrained at 1. This functional form is interceptless as it cannot be disentangled from α_i .

⁴²Both errors are multiplicatively separable under this model specification.

⁴³As an aside and particularly important for GMM estimation, the additive or multiplicative specification panel models are observationally equivalent when the conditional mean only is specified ([73], [59]) provided that $P(\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i + r_{it}) > 0) = 1$. Differences arise, however, for instrument choice under suspected endogeneity.

In addition to a fixed effects assumption, other standard assumptions are that there exists no serial correlation in the unobservables and that they are completely independent of the time-constant unobserved heterogeneity α_i .

If at least one of the regressors in $\mathbf{x}_{it}$ is endogenous, then $E[\mathbf{x}_{it}'v_{i(t-s)}] \neq 0, s \geq 0$. This implies that at a regressor is correlated with current and (perhaps past) time-varying unobservables. We outline the two-step GMM estimation for a multiplicative error model for consistency of parameter estimation. Under the assumption that at least one regressor in $\mathbf{x}_{it}$ is endogenous for an exponential model, we assume we have a set of exogenous instruments $\mathbf{z}_{it}$ ⁴⁴.

In this unobserved effects panel case, we still have the fixed effect in the population moment condition as with our proposed test for endogeneity. We can circumvent the issue of a potential endogeneity bias resulting from its existence (or if the unobserved heterogeneity being directly correlated with contemporaneous and possibly past time-varying unobservables) in order to develop appropriate moment conditions for parameter estimation. A quasi-differencing approach is suggested by Wooldridge to produce orthogonality conditions that extricates the fixed effect for consistent estimation of count data models with endogeneity. The resulting transformation via differencing results in an error term process in differences ([74], [75], [72], [70]). The time-constant, fixed effect must be eliminated. The explicit steps are

1. Divide equation (2.5) by μ_{it} .
2. Derive a single lag of equation (2.5).
3. Subtract equation (2.5) from its lagged version.

resulting in

$$\frac{y_{it}}{\mu_{it}} - \frac{y_{i(t-1)}}{\mu_{i(t-1)}} = \alpha_i(v_{it} - v_{i(t-1)}) \equiv \pi_{it}(\beta) \quad i = 1, \dots, N \quad t = 1, \dots, T_i$$

where the parameter vector is defined as $\beta = (\beta_1, \dots, \beta_k)'$. We can write Wooldridge's transformation to accommodate endogenous regressors more formally as:

$$\pi_{it}(\beta) = \left[\frac{y_{it}}{\mu_{it}} - \frac{y_{i(t-1)}}{\mu_{i(t-1)}} \right] = \alpha_i(v_{it} - v_{i(t-1)}) \quad i = 1, \dots, N \quad t = 1, \dots, T_i \quad (2.6)$$

where we can define the GMM orthogonality conditions as conditional, nonlinear population moment restrictions:

$$E[\pi_{it}(\beta)|\mathbf{z}_{it}] = E\left[\left(\frac{y_{it}}{\mu_{it}} - \frac{y_{i(t-1)}}{\mu_{i(t-1)}}\right)\middle|\mathbf{z}_{it}\right] = \mathbf{0} \quad i = 1, \dots, N \quad t = 1, \dots, T_i \quad (2.7)$$

where $\mathbf{z}_{it}$ can be drawn from $x_i^{t-2} \equiv (x_{i(t-2)}, x_{i(t-3)}, \dots, x_{i1})$ as well as other external instruments ([4]). This requires a bit of discussion. The GMM procedure for estimating is advantageous in the regard that the endogeneity of an explanatory variable in $\mathbf{x}_{it}$ can be rectified by using appropriate lags of this endogenous

⁴⁴Instruments exogenous such that the ℓ population unconditional moment holds $E[\mathbf{z}_{it}'\pi_{it}(\beta)] = \mathbf{0}$.

variable as valid instruments ([12], [67], [76]). The result from the quasi-differencing transformation in equation (2.7) have zero covariance between all $\pi_{it}(\beta)$ and $\mathbf{x}_{it}$ dated $t - 2$ and earlier. The exogeneity of such instruments is a direct feature of the assumed absence of serial correlation in the idiosyncratic errors. Nevertheless, the entire panel structure from period $t - 2$ and beyond can be used to obtain internal instrumental variables for extricating the correlation between an endogenous regressor in $\mathbf{x}_{it}$ and $\pi_{it}(\beta)$ ([77]).

Proof: In line with the presented transformation above and the defined vector of exogenous instruments $\mathbf{z}_{it}$, let $\mathbf{x}_{it} \equiv \mathbf{z}_{it}$ be endogenous with no serially correlated errors. Therefore, if $E[\mathbf{z}_{it}'\pi_{is}(\beta)] \neq 0$ for $s \leq t$, but $E[\mathbf{z}_{it}'\pi_{is}(\beta)] = 0$ for $s > t$. Therefore, two lags or more of $\mathbf{x}_{it} \equiv \mathbf{z}_{it}$ is permissible as $E[\mathbf{z}_{i(t-s)}'\pi_{it}(\beta)] = 0, t \geq 3, s \geq 2$. Therefore, the vector of instrumental variables for parameter identification is $\mathbf{z}_{it} = [x_{i1}, \dots, x_{i(t-2)}]$.

45

Under the quasi-differencing transformation using instrument lags for changes in period t , we note: For each cross-section i , let $\boldsymbol{\pi}_i(\beta) = (\pi_{i3}, \dots, \pi_{iT})'$ be of size $(T - 2) \times 1$ and $(T - 2) \times L$ be a block-diagonal matrix of instruments, $\mathbf{Z}_i$, with L representing the number of instruments as $L \equiv L_1 + L_2 + \dots + L_{T-2}$.

Thus:

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}_{i3} & 0 & 0 & 0 \\ 0 & \mathbf{z}_{i4} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{z}_{iT} \end{bmatrix}, \quad \boldsymbol{\pi}_i(\beta) = \begin{bmatrix} \pi_{i3}(\beta) \\ \pi_{i4}(\beta) \\ \vdots \\ \pi_{iT}(\beta) \end{bmatrix} \quad (2.8)$$

where, for instance, $\mathbf{z}_{i3} = [x_{i1}, \dots, x_{i(t-2)}]$ for the endogenous variable in $\mathbf{x}_{it}$, and with the unconditional population moment condition:

$$E[\mathbf{Z}_i'\boldsymbol{\pi}_i(\beta)] = \mathbf{0}$$

Then, an arbitrary GMM estimator, $\hat{\beta}_{GMM}$, using the quasi-differencing form, minimizes the given objective function

$$Q_N(\beta) = \arg \min_{\beta} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i'\boldsymbol{\pi}_i(\beta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i'\boldsymbol{\pi}_i(\beta) \right] \quad (2.9)$$

where $\mathbf{W}_N$ is a positive, symmetric, semi-definite⁴⁵ weighting matrix of size $\ell \times \ell$, $\mathbf{W}_N \perp \beta$, and $\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i'\boldsymbol{\pi}_i(\beta)$ is the sample complement of the population unconditional moment $E[\mathbf{Z}_i'\boldsymbol{\pi}_i(\beta)] = \mathbf{0}$. The process of the minimization of this objective function finds the solution that makes the value of this as close to 0 as possible.

⁴⁵Of course, other external instruments outside of endogenous lags are possible in $\mathbf{z}_{it}$ ([72], [78], [12], [4]).

⁴⁶The positive-definite condition for the weight matrix here is essential in order to ensure that all moment conditions are assigned a positive, non-zero weight.

Under certain regularity conditions ([62]) for a just-identified case⁴⁷ and irrespective choice of the weighting matrix $\mathbf{W}_N$, a GMM estimator has the following large sample properties

1. $\hat{\beta}_{GMM} \xrightarrow{P} \beta^*$
2. $\sqrt{N}(\hat{\beta}_{GMM} - \beta^*) \xrightarrow{d} N(\mathbf{0}, (\Upsilon' \mathbf{W}_N \Upsilon)^{-1} \Upsilon' \mathbf{W}_N \Psi \mathbf{W}_N \Upsilon (\Upsilon' \mathbf{W}_N \Upsilon)^{-1})$

where:

$$\Upsilon = E \left[\frac{\partial}{\partial \beta'} \gamma(\mathbf{Z}_i, \beta^*) \right], \quad \Psi = E[\gamma(\mathbf{Z}_i, \beta^*) \gamma(\mathbf{Z}_i, \beta^*)']^{48}$$

and Υ is the expected value of the of first derivatives of the moments (in which the instruments exceeds the moments). If the instruments surpasses the number of moments, the choice of the weight matrix $\mathbf{W}_N$ is crucial. The optimal choice of a weighting matrix $\mathbf{W}_N$ for minimizing the asymptotic variance in large samples is the covariance matrix's inverse for the moments Ψ^{-1} . With an optimal weight $\mathbf{W}_N^{\text{opt}}$, the property $\text{plim} \mathbf{W}_N^{\text{opt}} = \Psi^{-1}$ holds [62], and in the case of an overidentified system, the asymptotic distribution becomes

$$\sqrt{N}(\hat{\beta}_{GMM} - \beta^*) \xrightarrow{d} N(\mathbf{0}, (\Upsilon' \Psi^{-1} \Upsilon)^{-1})$$

This would result in an efficient GMM estimator that has the smallest possible asymptotic variance ([34]). However, this estimator is infeasible as Ψ^{-1} is unknown. In order to get a feasibly efficient estimator in the two-step process, we follow the procedure for a feasible solution proposed by [62] and [34] by finding an estimate for Ψ^{-1} (denoted as $\hat{\Psi}_N^{-1}$). The resulting estimator $\hat{\beta}$ from this two-step estimation process has the same asymptotic distribution as the minimizer of the objective function with the true, optimal weight matrix. In essence, the overall goal is to choose such weight matrices that highlights moments with small variance and weight them the most in estimation.

We highlight the estimation process for feasibly efficient, two-step GMM estimators with a weight matrix assuming errors are homoscedastic, a two-step GMM estimator with a heteroscedastic-robust weight matrix, and a two-step optimal GMM estimator that is impervious to both the effects of heteroscedasticity and autocorrelation. These errors correspond specifically to the variance-covariance matrix of the error process of $\pi_i(\beta)$. In all cases, $\mathbf{W}_N = \Psi_N^{-1}$ ⁴⁹.

Step 1: For our application, in the first step we choose an initial weight matrix Ψ (for the first step out of the two-step estimation) necessarily requires a submatrix such that

$$\Psi_{jk} = N^{-1} \sum_{i=1}^N \mathbf{z}_{ij}' \mathbf{z}_{ik}$$

⁴⁷In a just-identified case, the number of instruments equals the number of moments $L = \ell$.

⁴⁸This is the covariance matrix of the population moment conditions, where γ can be linear or non-linear in β .

⁴⁹Note, in all cases, the matrix Ψ_N is an estimate of the true covariance matrix of the error process under certain assumptions $\Psi = E[\mathbf{Z}_i' \pi_i(\beta) \pi_i'(\beta) \mathbf{Z}_i]$.

with $j = 1, \dots, \ell$ and $k = 1, \dots, \ell$ firstly used to compute a consistent, yet inefficient, estimate $\pi_i(\tilde{\beta})$ ⁵⁰.

Step 2(a): Proceeding with the second step, we choose:

$$\Psi_{jk} = \omega_{jk} N^{-1} \sum_{i=1}^N \mathbf{z}_{ij}' \mathbf{z}_{ik}$$

with

$$\omega_{jk} = N^{-1} \sum_{i=1}^N \pi_{ij}(\tilde{\beta}) \pi_{ik}(\tilde{\beta})$$

where j, k index the given moment conditions and Ψ_{jk} represents the (estimated) covariance matrix between the moment conditions. Thus:

$$\Psi_{jk} = \frac{\sum_{i=1}^N \pi_{ij}(\tilde{\beta}) \pi_{ik}(\tilde{\beta}) \sum_{i=1}^N \mathbf{z}_{ij}' \mathbf{z}_{ik}}{N} \equiv \Psi_N$$

Given $\mathbf{W}_N = \Psi_N^{-1}$, and under homoscedastic errors, then a feasibly efficient, two-step GMM estimator, $\hat{\beta}_{GMM}$, minimizes the objective function:

$$\begin{aligned} Q_N(\beta) &= \arg \min_{\beta} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right] \\ &= \arg \min_{\beta} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right]' \frac{N}{\sum_{i=1}^N (\pi_i(\tilde{\beta}))^2 \sum_{i=1}^N \mathbf{z}_i' \mathbf{z}_i} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right] \end{aligned}$$

Step 2(b): Under a heteroscedastic-robust specification of the weighting matrix and using the inefficient estimate, $\pi_i(\tilde{\beta})$, from step 1, then $\Psi_N = N^{-1} \sum_{i=1}^N \mathbf{z}_i \pi_i(\tilde{\beta}) \pi_i'(\tilde{\beta}) \mathbf{z}_i'$. Then, the heteroscedastic-robust feasibly efficient GMM estimator, $\hat{\beta}_{ROGMM}$, minimizes:

$$\begin{aligned} Q_N(\beta) &= \arg \min_{\beta} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right] \\ &= \arg \min_{\beta} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right]' \frac{N}{\sum_{i=1}^N \mathbf{z}_i \pi_i(\tilde{\beta}) \pi_i'(\tilde{\beta}) \mathbf{z}_i'} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{z}_i' \pi_i(\beta) \right] \end{aligned}$$

⁵⁰We also assume $\Psi_{jk} = \mathbf{0}$ for $j \neq k$. Thus, the initial weight matrix assumes unconditional moment equations are independent. Also, $\mathbf{z}_{ij}, \mathbf{z}_{ik} \subseteq \mathbf{z}_i$. Additionally, initial consistency of $\tilde{\beta}$ induces GMM estimator produced from it is to be asymptotically efficient.

Step 2(c): Finally, for a heteroscedastic and autocorrelation consistent (HAC) weight matrix and using the inefficient estimate, $\pi_i(\tilde{\beta})$, from step 1, then, $\Psi_N = N^{-1} \sum_{i=1}^N \mathbf{Z}_i \pi_i(\tilde{\beta}) \pi_i'(\tilde{\beta}) \mathbf{Z}_i' +$

$$N^{-1} \sum_{l=1}^{l=n-1} \sum_{i=1}^N K(l, m) \{ \mathbf{Z}_i \pi(\tilde{\beta}) \pi_{i-1}'(\tilde{\beta}) \mathbf{Z}_{i-1}' + \mathbf{Z}_{i-1} \pi_{i-1}(\tilde{\beta}) \pi_i'(\tilde{\beta}) \mathbf{Z}_i' \}^{51}.$$

Thus, the feasibly efficient GMM estimator, under heteroscedastic and autocorrelation assumption, is $\hat{\beta}_{HACGMM}$, minimizes:

$$Q_N(\beta) = \arg \min_{\beta} \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i' \pi_i(\beta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \mathbf{Z}_i' \pi_i(\beta) \right]$$

with $\mathbf{W}_N = \Psi_N^{-1}$ given as above.

Note, the resulting estimators in 2(a), 2(b), and 2(c) result are all feasibly efficient GMM estimators with the latter two being robust to heteroscedasticity, and heteroscedasticity/autocorrelation, respectively.⁵² A feasibly efficient GMM estimator is one that, under assumptions surrounding the true covariance matrix of the error process, yields the smallest asymptotic variance compared to all possible GMM estimators based on the derived population moment conditions ([81]).

Lastly, we test the overidentifying restrictions to verify if chosen, relevant instruments are unrelated to the error process, or, equivalently, the model is specified appropriately given the moment conditions. We test these overidentifying restrictions by employing the Hansen's J statistic, which is simply the returned value of the GMM objective function evaluated at an efficient GMM estimator ([34], [70], [62]).

In the case of the optimal, nonrobust GMM estimator, $\hat{\beta}_{OGMM}$, we have:

$$J(\hat{\beta}_{OGMM}) = N \times Q_N(\hat{\beta}_{OGMM})$$

With the optimal, robust GMM estimator, we have

$$J(\hat{\beta}_{ROGMM}) = N \times Q_N(\hat{\beta}_{ROGMM})$$

With the optimal, HAC GMM estimator, we have

$$J(\hat{\beta}_{HACGMM}) = N \times Q_N(\hat{\beta}_{HACGMM})$$

For a valid test of overidentifying restrictions, under $H_0 : E[\mathbf{Z}_i' \pi_i(\beta)] = \mathbf{0}$ (all moments have expectation equal to zero at the true value of the parameter) will follow a $\chi_{\ell-k}^2$, and a strong rejection of this null via a large test statistic casts doubt on the joint hypotheses of both the orthogonality of our moment conditions

⁵¹Note, let m correspond to the number of lags with $o = \frac{l}{m+1}$. Employing a Newey-West kernel, we have $K(l, m) = 1 - o$, $\iff 0 \leq o \leq 1$, 0 otherwise. As there is little theoretical guidance on optimal lag choice, we opt for current practice lag choice. Lags are chosen such that $m \approx \lfloor N^{1/4} \rfloor$ ([79]).

⁵²Note, the forms of the weight matrices matches forms from previous literature ([34],[80], [79]).

in addition to a correctly specified model ([34], [82]). Following that the null hypothesis is not rejected, then the moment conditions are appropriate and the GMM estimator is consistent. We now describe our application.

2.5 Application: Quantifying the Exposure Effects of Promotional Campaigns on In-Store Visits

The primary goal of this section is to quantify the marginal effect of campaigns on visiting behavior relative to reversions to no campaigns in the promotional periods for all households in loyalty promotion. Specifically, we take primary interest in both the single and joint treatment effects over periods in the promotional period while controlling for other variables. The covariates of interest are highlighted and describe in Table 2.6. We directly hypothesize a positive effect on our response variable over time which can be succinctly written as follows:

H1: Households' exposure to single or overlapping campaign(s) has (have) a higher differential impact on visits per period relative to households with no campaign in the test cycle.

2.5.1 Data Description

Data from this quasi-experiment contains all the transactions over a two-year time period from March 2005 to March 2007. This scanner transaction data captures the principal response of interest: in-store visits. The unbalanced nature of the cross-section of households in the promotional period is attributed to a variety of differing issues. Firstly, the duration of 30 customized and targeted campaigns, all exposure effects of interest, are varied. These chosen durations of the 30 campaigns mailed to consumers of interest are unknown to us. Secondly, the temporal decision to allocate a campaign via mail that lasts for the designated durations by the designers of the loyalty campaign is unknown, but presumably or likely associated with promotional cycle behavior for the active sample of households. Also, there is overlap in the main exposure/campaign effects which results in joint effects as a household receives two or three campaigns with coupons that are redeemable throughout the same duration. In any instance of this occurrence, we treat this as a joint effect, incrementing a distinct period, signifying the duration in which the simultaneous receipts of coupons are redeemable within a given time interval.

Table 2.6: Promotional Modeling: Response and Exogenous Variables: Descriptions

Variable	Operalization
Response Variable	
Visits per period	Counts of visits within a specific period
Independent Covariates	
Campaign Type	Coupon set received in a period (Type A, Type B, Type C, Type AB, Type AC, Type BC, Type ABC, and Type D)
Sales per period	The total per dollar amount attributed to the retailer
Period	Integer-valued counter representing the change in exposure effect of the campaigns
Diffdays	The duration, in days, representing the duration of the exposure effect for a redeemable time of a coupon set

We define period as an integer-valued covariate that represents a change in the campaign received across the test cycle. The type of campaigns that households receive in terms of both their regularity of distribution and the type of campaigns distributed are presumably attributed to the trip revenue certain households contribute in the control period. Specifically, the Type A campaign was likely sent more often to consumers with a higher spending amount in the control period which induces the promotional cycle’s endogeneity as affirmed previously ([83]). Households that spent more before the loyalty promotional program began were more likely to receive these personalized rewards in the promotional periods. Thus, households in this particular study experience these personalized campaigns as endogenous treatments. From an econometrics modeling standpoint, and following [25], we assume that the endogenous treatment (the personalized campaigns) will have an intercept-only effect on the outcome which shifts the mean visit rate upward or downward throughout the promotional cycle.

We observe aggregated overall average visits for all households throughout the duration of the study for both the baseline (period 1) and promotional (periods 2-34) cycles displayed in Figure 2.2. It should be noted that there is no transactional activity in period 32 (no households visited in this period). The overall average rate of visitation during the baseline period of the study is 0.1106747 visits per day. We observe that the overall average visitations per day for all periods oscillate from low to high from the odd to even numbered periods with an overall clear trend upward until period 18. The rate of visitation reaches its apex in the test cycle at period 26 with an average visit of 0.6423841 per day. Beyond period 26, the average visitation per day trends downward with an average visit of 0.1428571 during the final period the promotional cycle.

Of 30 campaigns spanning the duration of this shopper-card, loyalty promotional program, the specific coupons sets are associated with differing campaigns. For example, the first household in our data set received a themed-event coupon set (Type B) which was valid from the 281st day until the 334th day of this household’s two-year, time profile during the 29th campaign. Yet, after receiving the personalized mailing campaign (Type A) for a specified window of time thereafter, the household received, again, Type B beginning on the 477th day during 12th campaign. We presume no difference between Type B received in different windows of time or different campaigns in terms of the overall treatment effect of interest. Any personalized or targeted campaign received during any at any point in the promotional period represents the same exposure effect for any household. As stated previously, there are no restrictions imposed on the simultaneous receipt of campaigns to customers within the same window of time. For example, a customer could both be exposed to a Type A and Type B coupon sets during overlapping campaigns concurrently over the same time period. Cases in which this occurred signifies a combined set of campaigns which are modeled as joint, yet distinct treatment effects (e.g. Type AB, Type ABC, Type AC, and Type BC). Lastly, in the promotional period, a household can regress to a washout state in which no treatments are received (Type D).

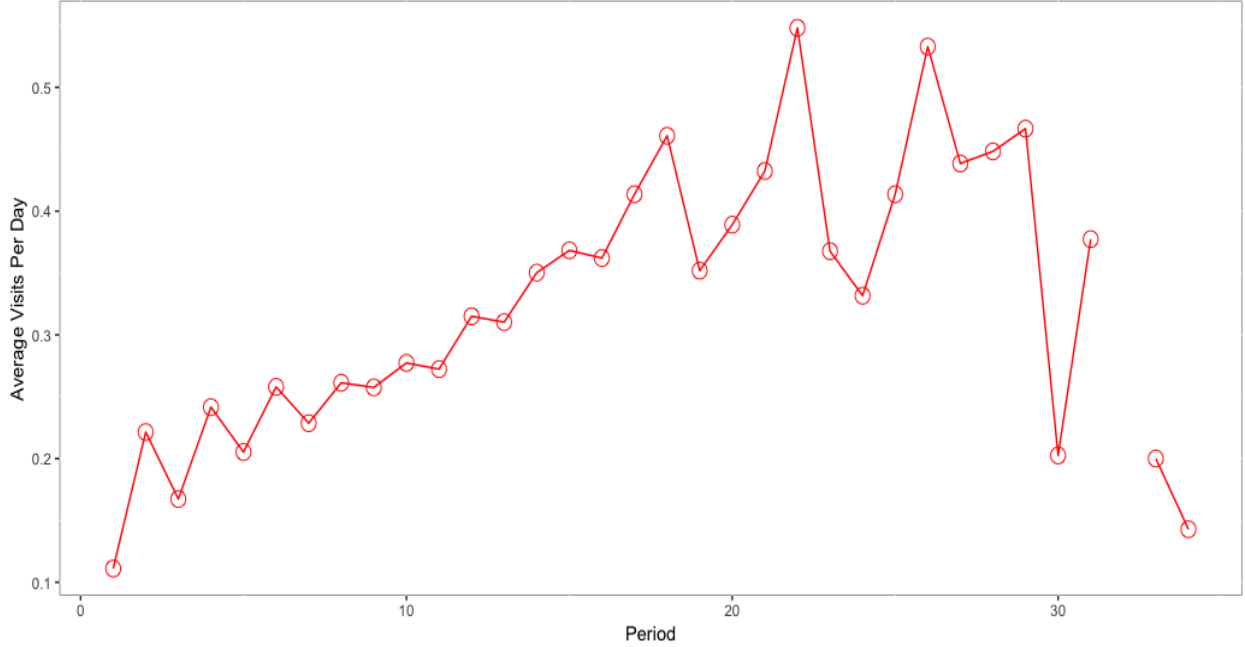


Figure 2.2: Average visits per day per period

We display the average visits given main exposure effects of interest (Type A, Type B, and Type C campaigns) throughout the entire duration of the study, focusing on the individual treatments received in the test cycle relative to no campaign (Type D). Figure 2.3 shows the overall average visit is higher relative to Type D (indicative of no campaign received in periods 2-34) throughout the entire duration of the study. Interestingly, it appears the visit rate for Type B (themed coupon set) has the highest impact on overall visitation rate increase compared to Type A and Type C campaigns throughout the duration of the test cycle across the time series. For completeness, we demonstrate the average visits per day for each period for the joint campaigns as well in Figure 2.4. We overwhelmingly observe higher average visits given a joint campaign relative to no campaign as expected.

We formally hypothesize that the singular exposure effect of the Type A campaign is indeed endogenous. We will conduct the clean test of idiosyncratic endogeneity of the Type A exposure effect.

H2: The Type A exposure effect is endogenous in the promotional cycle.

We make an important note of distinction here. The Type A exposure effect is presumably “singly” endogenous, meaning that the joint exposures containing Type A (e.g. Type AB, Type ABC, Type and AC), are not endogenously determined and is treated as a separate, exogenous effect. Recall, the joint exposures are measured via simple overlap in campaign durations in which the starting date of a new campaign received by a household does not begin before the ending date of the previous campaign, thus are not presumed endogenous.

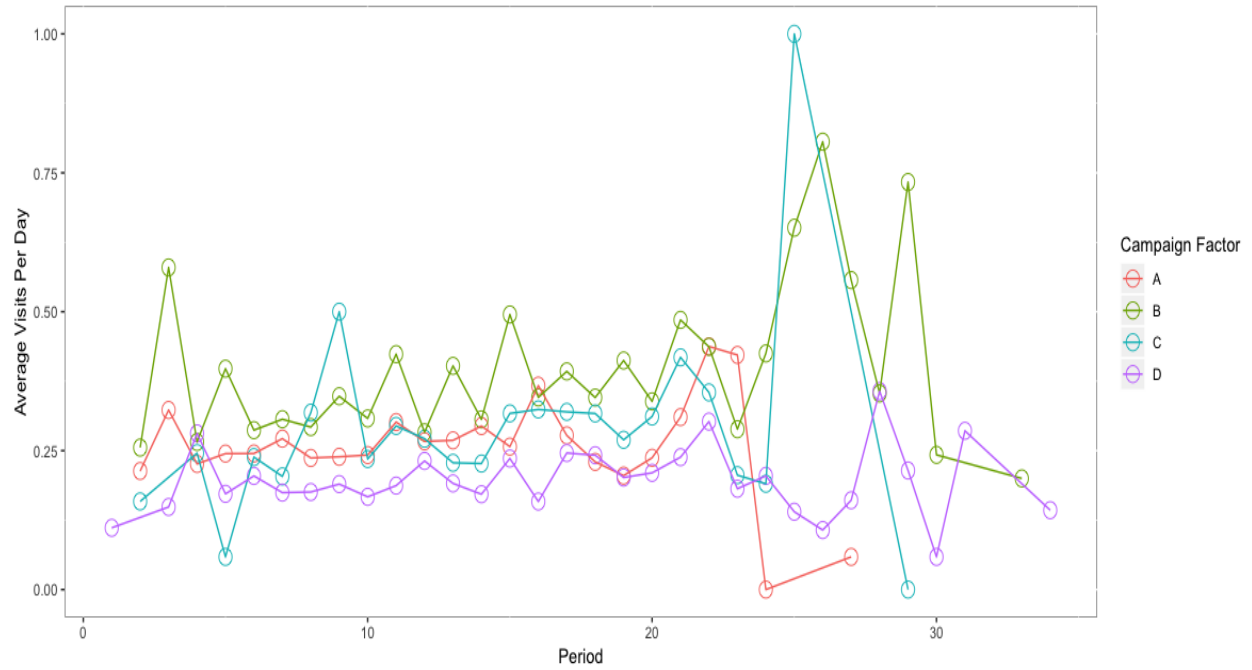


Figure 2.3: Average visits per day (Single Campaign Effects)

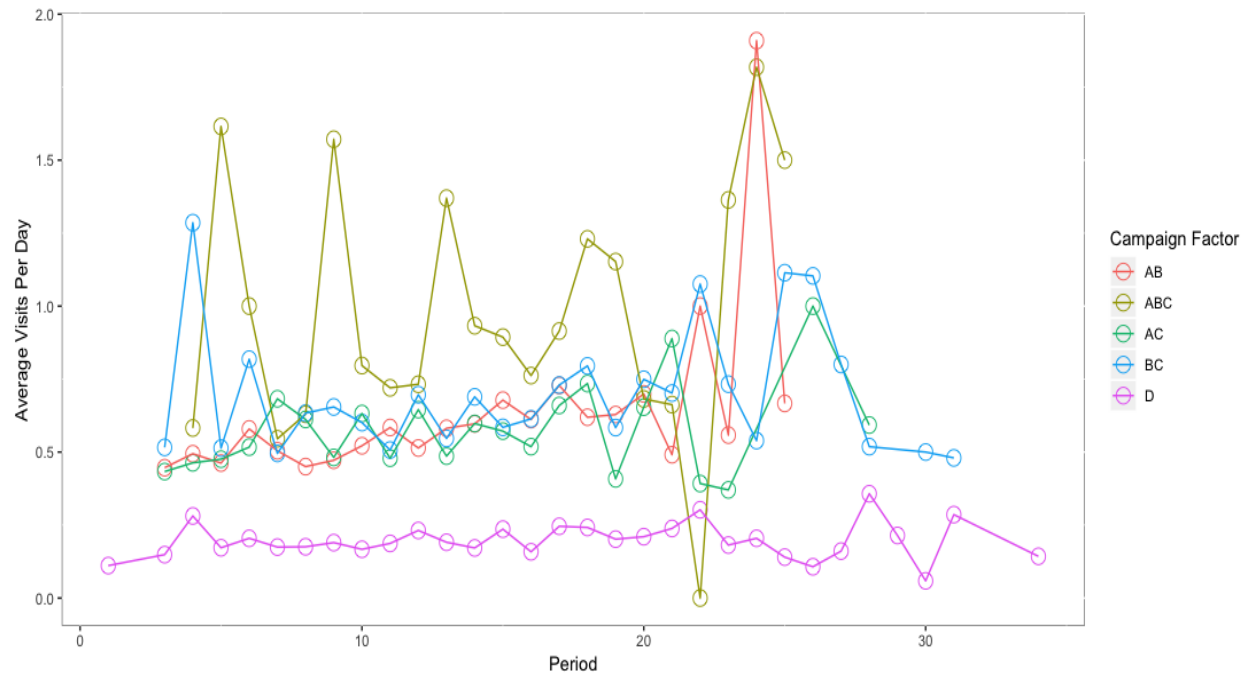


Figure 2.4: Average visits per day (Joint Campaign Effects)

We now define the instrumental variables that will aid in this test of idiosyncratic endogeneity in the case of time-invariant heterogeneity. In order to utilize the combined CRE/CF technique, we need valid instruments that we assume are directly correlated with the potential endogenous variable that does not directly impact our responses, yet indirectly impacts the responses only through its correlation with the endogenous covariate. Table 2.7 summarizes our additional instruments. We briefly discuss the practicality of a valid instrumental variable beyond *Assumption 2* in the previous section. Beyond its mandatory non-correlation with the error in the structural equation, a valid instrumental variable must explicitly explain variation or have an association with the potential endogenous variable ($E(\mathbf{z}'_{it}y_{it2}) \neq 0$).

Following [83] and [84], we have evidence that control period variables are directly associated with the distribution of Type A campaigns throughout the entire test period’s epoch. We utilize the covariates “Pre Days Between” and “Pre Sales”, which serve as instruments, computed in the control phase before any campaigns are distributed in the test phase (Days 1-223 or, correspondingly, weeks 1-32). Regarding the practicality of instrument validity in our application, we’d expect a negative association between the total number of days between visits and the receipt of Type A campaigns (the longer the duration of no store visits, the less likely a household receives a personalized campaign containing personalized coupons tailored to their purchasing behavior), but a positive association with “Pre Sales”. In addition, we utilize two other time-invariant covariates that are household specific. “Baseline Visit Rate” represents the visit rate per day of the first period and “Baseline Spend”, which represents the total revenue contribution of each household in the first period. Note, the slight difference between the control phase and a household-specific first period. The control phase is uniform for all periods (days 1-223), whereas a household’s first period is non-uniform in terms of the duration due to the fact that the 30 campaigns begin at differing start days. For example, the first household is not exposed to its first promotional campaign (Type B) until day 281, whereas the second household is not exposed to its first promotional campaign (Type A) until day 587. Positive associations are expected between these time-invariant variables and the receipt of Type A Campaigns in the test period. Lastly, we utilize two additional, time-varying instruments: “Time in Previous NPP” (the number of days a household spent in the prior washout period) and “Lag Items per period” (total in basket items purchased from the retailer in the prior period). We believe that likely determinants of endogeneity of the personalized reward, in addition to control period behavior as well as household-specific baseline behavior, are time-varying explanatory variables that are associated with the receipt Type A campaigns. A likely association is expected between “Time in Previous NPP” and Type A as well as a positive association between prior period item purchases and receipt of Type A campaigns.

Following [60], we can directly check for instrumental relevance by regressing our endogenous variable, Type A, on the seven instruments given above and compute a naive regression in order to empirically check instrument relevance. A joint F-statistic greater than 10 indicates relevant instruments. We report the individual F-statistic for each individual instrument as well as the joint F-statistic for them all.⁵³

⁵³We also checked for collinearity between these instruments, and little evidence is found other for severe collinearity with the regressors with the highest VIF being 5.3 for Pre Days Between.

Table 2.7: Promotional Modeling: Endogenous Variable and Instrumental Variables - Descriptions

Variable	Definition
Endogenous Variable	
Type A	Personalized Reward Campaign sent to the most loyal customer base
Instrumental Variables	
Time in Previous NPP	The number of days a household was not exposed to any campaign in the previous promotional period
Lag items per period	The total number of in-basket items purchased from the retailer in the previous period of the test phase
Baseline Visit Rate	The total number of visits for each household's first period divided by the duration in days of the respective period
Baseline Spend	The total revenue contribution to the retailer for each household's first period.
Pre Days Between	The sum of the number of days in the control period where a household did not make a purchase (Days 1-223)
Pre Visits	The sum of the number of visits in the control period (Days 1-223)
Pre Sales	The total revenue contribution to the retailer during the control period (Days 1-223)

Table 2.8: Instrument Relevance Test

Response: Campaign Type A	F-Statistic
Pre Sales	$F(1,1583) = 2.36$
Time In Previous NPP	$F(1,1583) = 6513.82$
Lag Items Per Period	$F(1,1583) = 11.51$
Baseline Spend	$F(1,1583) = 7.65$
Baseline Visit Rate	$F(1,1583) = 9.88$
Pre Days Between	$F(1,1583) = 17.25$
Joint ($H_0 : \pi_1 = \dots = \pi_6 = 0$)	$F(6,1583) = 1382.98$

Finally, we note that all of these additional exclusions are reasonably “strictly exogenous” to the count response “visits per period”. Out of these six additional exclusions, we note that four of them are completely time-invariant (either computed for days 1-223 only or computed based on the range of days for the household-specific first period), and almost completely assured to not have an extended influence in visits further in the subsequent test phase (periods 2-34). This is even further reaffirmed by the fact that the estimation window for assessing the promotional campaigns will not include Period 1 for all households where the starting date begins on at day 224 (thus, the first associated exposure effect for all households in Period 1 corresponds to pre-emptive exposure effect allocation). Beneficially, as well, our CRE/CF approach will allow for the estimation of these time-invariant instruments which will ultimately produce variability in Type A_{it}. Regarding the two time-varying covariates, we note that entire sample of 1,584 households in the test phase are amongst the most loyal customers, and will likely visit the retailer regardless of the amount of money spent per aggregated purchased items in a previous period or regardless of the days these households went without an exposure effect.

2.5.2 Promotional Modeling Results: Endogeneity Test

We now present the results from our CRE model with a control function approach to test whether or not the Type A campaign exposure effect is endogenous. Note, Type A is a 0/1 dummy variable as well in the periods that span the promotional phase. Regarding the necessary assumptions, it is reasonable to assume that once controlling for all of the exogenous covariates and unobserved heterogeneity, that the unbalancedness of our cross section of households differing for periods is indeed unrelated to the idiosyncratic errors of our panel.

We present the linear reduced form for Type A campaign in the following equation

$$\text{Type A}_{it} = \delta_2 + \mathbf{z}_{it}'\mathbf{\Lambda}_2 + \bar{\mathbf{z}}_i'\phi_2 + \nu_{it2}$$

where $\mathbf{z}_{it}$ = (Type B, Type C, Type AB, Type AC, Type BC, Type ABC, Period2,...,Period34, Lag Items per period, Time in Previous NPP, Sales per period, Control Visit Rate, Baseline Spend, Pre Days Between, Pre Sales), and $\bar{\mathbf{z}}_i$ contains the averages across t via the Mundlak device. From here, we compute $\hat{\nu}_{it2}$, the first stage residuals. These computed residuals, $\hat{\nu}_{it2}$ serve as a control function for the structural equation. Note, in this case, we also compute ($\bar{y}_{i2} = \overline{\text{Type A}_{it}}$), or the time-invariant average of the proposed endogenous campaign for inclusion into the structural equation.

We finally report the results from chosen covariates from our second stage pooled poisson quasi-maximum likelihood estimation in Table 2.9⁵⁴. First, we note that our variable “period” was transformed into time dummies and all time dummies variables in our structural equation are completely suppressed⁵⁵. Additionally, an intercept is estimated with this procedure, yet suppressed in the output, and the offset log(DiffDays) was included in the estimation as well. We only display a few of the covariates in $\bar{\mathbf{z}}_i$ as well as $\bar{y}_{i2}$. Regarding the endogeneity test of Type A, we note $\theta_1 \neq 0$ which gives evidence that our binary treatment “Type A” is indeed marginally endogenous as the coefficient is statistically significant at $\alpha = 0.10$, verifying our **H2**. If y_{it2} were to be continuous which could be written as linear reduced form provided proper instruments, then we could bootstrap our standard errors to correct for the generated regressor or control function [3]. However, as stated previously, since endogenous variable has a restrictive range (binary), the bootstrap correction is invalid for inference, and thus these results can only function as a pure test of idiosyncratic endogeneity. The homoscedastic-normal assumption of residual normality of the binary regressor Type A as well as its independence assumed from $\mathbf{z}_{it}$ is highly unlikely due to the residuals’ non-constant variance resulting from a linear probability model and its non-Gaussian error distribution. We seek to gain consistent estimates of our parameters via a nonlinear, IV-GMM approach in the next subsection.

2.5.3 IV-GMM Results

Our goal is to assess the effectiveness of the exposure effect of the campaigns in the promotional period relative to periods of reversion to no campaign (Type D) in the promotional phase. Recall, our promotional period begins at $T \geq 2$. As we invoked Wooldridge quasi-differencing transformation to eliminate the presumed multiplicative fixed effect, the initial condition requires that, for each household/cross-section, period 1 is removed (analogous to the traditional “first-difference”). Its removal precisely coincides to the scope of our research question: estimating effectiveness of personalized and targeted campaigns in the promotional phase ($T \geq 2$).

⁵⁴We do not report here the results of the first-stage estimation, yet can be found in the appendix.

⁵⁵We included time dummies and their corresponding means into both the first stage and second stage equations as suggested by ([3], [9], [19], [17]). Of note, 26 of the time dummies representing periods 2-34 are all statistically significant in the second stage pooled Poisson QMLE regression.

Table 2.9: Two Stage CRE/Control Function Test for Idiosncratic Endogeneity

	2nd Stage - Pooled Poisson QML (cluster-robust standard errors)
Response: Visits per period	
Type A	.0619143 (0.0804806)
Type AB	.6723852*** (.0316291)
Type ABC	.9180157** (.0605815)
Type AC	.6097588*** (.0420382)
Type B	.1863727*** (.0571307)
Type BC	.5861677*** (.0535147)
Type C	.0881119 (.0619476)
Sales Per Period	.0004617*** (.0001938)
Lag Items per Period Mean	.0017483*** (.0003784)
Pre Days Between Mean	-.0106137*** (.0015836)
Baseline Visit Rate Mean	.9524872*** (.1689464)
Baseline Spend Mean	.0000574*** (.0000017)
Pre Sales Mean	-0.0004348*** (.0000339)
Time in Previous NPP Mean	.0058673*** (.000749)
Type A Mean	-0.1093914 (.4285114)
$\hat{\nu}_{it2}$	.1355148* (.0818005)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Our moment evaluator algorithm utilizes the STATA software in conjunction with the GMM function for our estimation ([82]). Unique to our application is the necessity to incorporate the offset, $\log(\text{diff days})$, into our explanatory vector $\mathbf{x}_{it}$ such that the covariate interpretation can be articulated as rates per campaign duration day once the resulting coefficients are exponentiated. Assuming our offset, t_{it} is a covariate in $\mathbf{x}_{it}$, recall that the conditional moment conditions can be represented as

$$E[\pi_{it}(\boldsymbol{\beta})|\mathbf{z}_{it}] = E\left[\left(\frac{y_{it}}{\mu_{it}} - \frac{y_{i(t-1)}}{\mu_{i(t-1)}}\right)\middle|\mathbf{z}_{it}\right] = \mathbf{0} \quad i = 1, \dots, N \quad t = 1, \dots, T_i$$

where $\mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \ln(t_{it}))$, $\mu_{i(t-1)} = \exp(\mathbf{x}_{i(t-1)}'\boldsymbol{\beta} + \ln(t_{i(t-1)}))$. Therefore, the covariate vector $\mathbf{x}_{it}$ contains the offset variable $\ln(t_{it}) = \log(\text{diff days})$.

For all GMM estimators, we have:

- $\mathbf{x}_{it} = (\text{Type A, Type B, Type C, Type AB, Type ABC, Type BC, Type AC, Sales Per Period, Period})$
- $\mathbf{z}_{it} = (\text{L2.Type A}^{56} \dots, \text{Baseline VR, L1.Items Per Period, Baseline Spend, Pre Items, Pre Sales, Pre Days Between, Pre Visits, Type B, Type C, Type AB, Type ABC, Type BC, Type AC, Sales Per Period, Period})$

All variables in $\mathbf{x}_{it}$ are presumed/rendered exogenous with the exception of Type A in the moment conditions. Additionally, for parsimony and for estimation of the time component in this nonlinear IV-GMM estimation, we convert the variable period to continuous to represent the time effect trend as opposed to time dummies as, conceptually, there is no difference. All explanatory variables in $\mathbf{x}_{it}$ can be instrumented contemporaneously with themselves as strict exogeneity holds except for the endogenous Type A which can be instrumented with its 2 or more lagged specification of itself and/or the other external instruments in $\mathbf{z}_{it}$. The idea of using lags of the endogenous variable for the customized exposure effect in our application is intuitive. The receipt of customized campaign in the past is likely to be strongly associated with the receipt of a customized campaign in the contemporaneous period, but it's very unlikely to be correlated with the contemporaneous error term or unobservables at time t . In other words, the customized exposure effect may affect visits in a current period, but it is less likely that visits per period in a current period t can influence past periods of the customized exposure effect. Although the contemporaneous distribution of customized coupons is endogenous to contemporaneous visits per period, it is unlikely that past receipts of the customized exposure effects are subject to the same problem. In our instrument set is the use of a few time invariant instruments for the Type A endogenous exposure effect as well. Specifically, for each household, we have total control period purchased items, total control period spend, total baseline spend, total baseline visits, total control period visits, and total days of intermittent visits in the control period. The use of these time-invariant instruments for our time-varying exposure effect is permissible as essentially we instrument the mean of the time-varying Type A exposure effect while ignoring the variation over time. As our Type A endogenous

⁵⁶The “...” represents further lags of the endogenous Type A campaign.

exposure effect has sufficient variation between households over time, this poses no problem. Additionally, we do have the time-varying, one period lag of items per period for our contemporaneous endogenous Type A exposure effect as a higher degree of prior period item purchasing is likely to induce the receipt of a Type A campaign in the following period. We now present the results from our three feasibly efficient GMM estimators under assumptions of pure homoscedasticity, heteroscedastic-robust, and HAC robust weight matrices. We will compare these sets of parameter estimates to the Poisson fixed-effects estimator. We arbitrarily include this inconsistent Poisson fixed-effects just as a comparison. As seen previously, we have statistical evidence of its inconsistency due to the previous statistically significant endogeneity test for Type A. The results are presented in Table 2.10. We observe the results in 2.10. Regarding the sets of parameter estimates estimated by GMM (differing only by the weight matrix $\mathbf{W}$), we firstly note that we have 16 moment conditions (the size of $\mathbf{z}_{it}$) where we instrument Type A with 8 instrumental variables: Control VR, Items Per Period, Baseline Spend, Pre Items, Pre Sales, Pre Days Between, Pre Visits, and a five period lag of Type A, resulting in overidentification. Testing the overidentification moment conditions, we note the Hansen's J test statistics in all three cases. Under the null hypothesis, all three of our feasible GMM estimators seem to approximate a $\chi^2_{16-9} = \chi^2_7$ distribution. The p-values indicate that there is no issue with a correctly specified model as we have no evidence that our instruments are not truly orthogonal to the errors. We observe the difference in sample sizes for our feasible GMM estimators relative to the Poisson fixed effects estimator.

Table 2.10: Nonlinear-IV GMM - Promotional Period Estimation Results

	Poisson Fixed Effects $\hat{\beta}$	Exponential Feasible GMM $\hat{\beta}_{GMM}$	Exponential Robust Feasible GMM $\hat{\beta}_{ROGMM}$	Exponential HAC Feasible GMM $\hat{\beta}_{HACGMM}$
Response: Visits Per Period				
Type A	.2106952*** (.0132419)	.2042024*** (.0534658)	.1653864*** (.0550055)	.2186897*** (.0520153)
Type AB	.7915221*** (.0191563)	.7277908*** (.0469544)	.6891328*** (.0456058)	.7267136*** (.04326)
Type ABC	1.082939*** (.0402922)	1.243841*** (.1013631)	1.110758*** (.1107737)	1.163641*** (.0893478)
Type AC	.7597581*** (.0319428)	.8090906*** (.067551)	.7275532*** (.0680961)	.7887801*** (.0606943)
Type B	.3130157*** (.0188232)	.1448757*** (.030293)	.1465182*** (.0362061)	.1609322*** (.0355984)
Type BC	.7694571*** (.0480557)	1.024767*** (.1070309)	.8901577*** (.130973)	.9250788*** (.1014796)
Type C	.1874544*** (.0325578)	.4285271*** (.0809527)	.3290528*** (.0837108)	.3791257*** (.0706266)
Sales Per Period	.0000435*** (.00000853)	.0005508*** (.0000299)	.0005659*** (.0000439)	.0005638*** (.000047)
Period	-.0022905 (.0019216)	-.0487383*** (.0108926)	-.0405541*** (.0097699)	-.0443778*** (.005396)
N	15822	8439	8439	8439
Hansen's J	-	9.04799 (p = 0.2492)	4.17167 (p = 0.7598)	13.6746 (p = 0.0573)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Our choice of using a relatively non-conservative, five-period lag of Type A to mitigate the bias induced by the endogenous covariate causing such a sample size reduction. Therefore, as opposed to garnering estimates for the entire 1,584 households that spans the promotional period effects beginning at period 2, we estimate a model for 1,092 households beginning at period 6 (due to the five period lag). In other words, these parameter estimation corresponds to the most active subset of the households in the promotional periods (approximately $1092/1584 = 68\%$) of the active sample.

For our three GMM estimates, all coefficients are statistically significant and have similarly sized magnitudes. Unsurprisingly, the simultaneous exposure effect to two or three campaigns in the promotional cycle yield higher coefficients in all cases. This is perhaps due to an induced feeling of exclusivity, surprise, or loyalty of the households in the promotional program by the overlapping receipts of the combination of campaigns. In line with other literature, sales contribution to the retailer from households in the test phase seems to have a per-dollar practical insignificance on the rate of visits. It is also noteworthy that based on the Hansen’s J test statistic, the Robust GMM ($\hat{\beta}_{ROGMM}$) is lowest, which seemingly indicates that the specified moment conditions are met most appropriately in our multiplicative error regression specification when there is heteroscedasticity of unknown form given our instruments. For interpretation of our exposure effect, we choose to interpret the covariates corresponding to the heteroscedastic and autocorrelation consistent, feasibly efficient GMM estimator. Regarding our HAC GMM estimator, $\hat{\beta}_{HACGMM}$, the necessity of lag choice is essential to employ the Newey-West kernel for the heteroscedasticity and autocorrelation robust estimator. Using previous literature for a guided choice of the temporal lag ([79]), we choose a lag order of $m \approx \lfloor N^{1/4} \rfloor = 8$. We turn our attention for the results presented in Table 2.10 for coefficient interpretation. The coefficient corresponding to Type A is .2186897. Therefore, holding all other covariates constant, the rate of visitation per day for those who received the Type A exposure effect is approximately 24.44% ($\exp(.2186897) = 1.244445$) higher relative to no promotional campaign. The covariate corresponding to period indicates an approximate 4.34% ($\exp(-.0443778) = 0.9565925$) period decrease in visitation rate each additional period further in the test period, holding all else constant. Other coefficients can be interpreted similarly.

2.6 Discussion

Summary: In this chapter, we present an extension of work presented by [21] to create a fully robust, complete-case (CC) test of idiosyncratic endogeneity in the presence of heterogeneity endogeneity using a combined correlated random effects/control function approach for an exponential, conditional mean model for count responses in an unbalanced panel framework when the endogenous variable is of restricted range (binary). The robustness of this test manifests itself in its ability to counteract potential heteroscedasticity of unknown form, potential serial correlation, and model misspecification as it relates to counts that may potentially exhibit excessive variability or overdispersion. The flexibility of the CRE allows the inclusion

of exogenous, time-invariant instruments which will produce variation in the first-stage regression of the binary endogenous covariate. Via a linear probability model, we yield a generated regressor which serves as the control function in the test for idiosyncratic endogeneity. This approach completely isolates the potential endogeneity between a binary time-varying regressor and the unobservables exclusively as the time constant endogeneity is completely controlled for as we achieve an estimator that is completely impervious to any time-invariant endogeneity. To assess the finite sample performance of our endogeneity tests, we simulate three scenarios for panel structures that researchers will likely encounter in applied settings: an unbalanced panel with selection, an unbalanced panel without selection, and a balanced panel with distinct considerations for instrument strength (weak vs. strong). In all cases, we were able to produce high test rejections rates dependent upon the strength of the endogeneity induced. As these promising Monte Carlo results are highly dependent upon instrument strength, we note the importance of the relevancy condition of chosen instrumental variables for performing such a test.

As our endogenous form of the regressor of interest is of restricted range, this is only a test of pure idiosyncratic endogeneity in the presence of time-constant, heterogeneity endogeneity, and not an outright correction. A more advanced, yet flexible procedure is needed. Therefore, we rely on a nonlinear, instrumental variables, GMM estimation approach to produce three feasibly efficient IV-GMM estimators. In all three instances, our nonlinear moment conditions are met via the overidentifying restrictions test. In terms of our application, we were able to confidently verify our hypotheses, detecting endogeneity of the distribution of personalized campaigns in addition to increased number of visits relative to intermittent epochs of no exposures.

References

- [1] J. A. Hausman, B. H. Hall, and Z. Griliches, “Econometric models for count data with an application to the patents-r&d relationship,” 1984. [15](#), [16](#)
- [2] C. Gourieroux, A. Monfort, and A. Trognon, “Pseudo maximum likelihood methods: Theory,” *Econometrica: journal of the Econometric Society*, pp. 681–700, 1984. [16](#)
- [3] J. M. Wooldridge, *Econometric analysis of cross section and panel data*. MIT press, 2010. [16](#), [18](#), [21](#), [22](#), [23](#), [25](#), [27](#), [31](#), [32](#), [36](#), [56](#)
- [4] E. A. Ramalho, J. J. Ramalho, and L. M. Coelho, “Exponential regression of panel data fractional response models with an application to firm capital structure,” 2015. [16](#), [44](#), [45](#), [46](#)
- [5] P. D. Allison, *Fixed effects regression methods for longitudinal data using SAS*. Sas Institute, 2014. [16](#), [18](#)
- [6] J.-P. Boucher and M. Denuit, “Fixed versus random effects in poisson regression models for claim counts: A case study with motor insurance,” *ASTIN Bulletin: The Journal of the IAA*, vol. 36, no. 1, pp. 285–301, 2006. [17](#)
- [7] J. A. Hausman, “Specification tests in econometrics,” *Econometrica: Journal of the econometric society*, pp. 1251–1271, 1978. [17](#)
- [8] A. C. Cameron, P. K. Trivedi, *et al.*, “Count panel data,” 2013. [17](#), [32](#)
- [9] Y. Mundlak, “On the pooling of time series and cross section data,” *Econometrica: journal of the Econometric Society*, pp. 69–85, 1978. [18](#), [28](#), [31](#), [56](#)
- [10] J. M. Neuhaus and J. D. Kalbfleisch, “Between-and within-cluster covariate effects in the analysis of clustered data,” *Biometrics*, pp. 638–645, 1998. [18](#)
- [11] A. Elzinga and B. Gasperini, “Correlated random effects model: an application to the exchangeability of siblings and twins,” 2015. [18](#)
- [12] M. Arellano and S. Bond, “Some tests of specification for panel data: Monte carlo evidence and an application to employment equations,” *The review of economic studies*, vol. 58, no. 2, pp. 277–297, 1991. [18](#), [43](#), [46](#)
- [13] B. Crépon and E. Duguet, “Estimating the innovation function from patent numbers: Gmm on count panel data,” *Journal of Applied Econometrics*, vol. 12, no. 3, pp. 243–263, 1997. [18](#)
- [14] R. Blundell, R. Griffith, and F. Windmeijer, “Individual effects and dynamics in count data models,” *Journal of econometrics*, vol. 108, no. 1, pp. 113–131, 2002. [18](#)

- [15] P. Albarran, R. Carrasco, and J. M. Carro, “Estimation of dynamic nonlinear random effects models with unbalanced panels,” *Oxford Bulletin of Economics and Statistics*. [18](#)
- [16] J. M. Wooldridge, “A simple method for estimating unconditional heterogeneity distributions in correlated random effects models,” *Economics Letters*, vol. 113, no. 1, pp. 12–15, 2011. [18](#)
- [17] J. M. Wooldridge, “Correlated random effects models with unbalanced panels,” *Journal of Econometrics*, 2018. [18](#), [20](#), [27](#), [28](#), [31](#), [56](#)
- [18] I. Murtazashvili and J. M. Wooldridge, “Fixed effects instrumental variables estimation in correlated random coefficient panel data models,” *Journal of Econometrics*, vol. 142, no. 1, pp. 539–552, 2008. [19](#)
- [19] L. E. Papke and J. M. Wooldridge, “Panel data methods for fractional response variables with an application to test pass rates,” *Journal of Econometrics*, vol. 145, no. 1-2, pp. 121–133, 2008. [20](#), [56](#)
- [20] R. Joshi, J. M. Wooldridge, *et al.*, “Correlated random effects models with endogenous explanatory variables and unbalanced panels,” *Annals of Economics and Statistics*, no. 134, pp. 243–268, 2019. [20](#), [31](#)
- [21] W. Lin and J. M. Wooldridge, “Testing and correcting for endogeneity in nonlinear unobserved effects models,” 2017. [20](#), [26](#), [28](#), [31](#), [36](#), [37](#), [39](#), [41](#), [60](#)
- [22] J. B. Sande and M. Ghosh, “Endogeneity in survey research,” *International Journal of Research in Marketing*, 2018. [21](#)
- [23] I. Kasparis, “Detection of functional form misspecification in cointegrating relations,” *Econometric Theory*, vol. 24, no. 5, pp. 1373–1403, 2008. [21](#)
- [24] P. A. Jargowsky, “Omitted variable bias,” 2005. [21](#)
- [25] J. A. Clougherty, T. Duso, and J. Muck, “Correcting for self-selection based endogeneity in management research: Review, recommendations and simulations,” *Organizational Research Methods*, vol. 19, no. 2, pp. 286–347, 2016. [22](#), [38](#), [51](#)
- [26] J. J. Heckman, “The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models,” in *Annals of Economic and Social Measurement, Volume 5, number 4*, pp. 475–492, NBER, 1976. [22](#)
- [27] J. J. Heckman, “Sample selection bias as a specification error (with an application to the estimation of labor supply functions),” 1977. [22](#)
- [28] R. A. Berk, “An introduction to sample selection bias in sociological data,” *American Sociological Review*, pp. 386–398, 1983. [22](#)

- [29] S. Bushway, B. D. Johnson, and L. A. Slocum, “Is the magic still there? the use of the heckman two-step correction for selection bias in criminology,” *Journal of quantitative criminology*, vol. 23, no. 2, pp. 151–178, 2007. [22](#)
- [30] C. Winship and R. D. Mare, “Models for sample selection bias,” *Annual review of sociology*, vol. 18, no. 1, pp. 327–350, 1992. [22](#)
- [31] I. Fernández-Val and F. Vella, “Bias corrections for two-step fixed effects panel data estimators,” *Journal of Econometrics*, vol. 163, no. 2, pp. 144–162, 2011. [22](#), [24](#)
- [32] H. B. Nguyen, *Estimating Nonlinear Cross Section and Panel Data Models with Endogeneity and Heterogeneity*. Michigan State University. Economics, 2011. [22](#)
- [33] J. Antonakis, S. Bendahan, P. Jacquart, and R. Lalive, “Causality and endogeneity: Problems and solutions,” [22](#)
- [34] C. F. Baum, M. E. Schaffer, *et al.*, “Instrumental variables and gmm: Estimation and testing,” 2003. [23](#), [47](#), [49](#), [50](#)
- [35] J. J. Heckman, H. Ichimura, and P. Todd, “Matching as an econometric evaluation estimator,” *The review of economic studies*, vol. 65, no. 2, pp. 261–294, 1998. [23](#)
- [36] I. Petrakis, “The application of quasi-experimental methods in the context of marketing: a literature review,” 2016. [23](#)
- [37] J. V. Terza, A. Basu, and P. J. Rathouz, “Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling,” *Journal of health economics*, vol. 27, no. 3, pp. 531–543, 2008. [24](#)
- [38] S. Navarro, “Control functions,” in *Microeconometrics*, pp. 20–28, Springer, 2010. [24](#)
- [39] G. Bascle, “Controlling for endogeneity with instrumental variables in strategic management research,” *Strategic Organization*, vol. 6, no. 3, pp. 285–327, 2008. [24](#)
- [40] B. Hidayat and S. Pokhrel, “The selection of an appropriate count data model for modelling health insurance and health care demand: case of indonesia,” *International journal of environmental research and public health*, vol. 7, no. 1, pp. 9–27, 2010. [24](#)
- [41] Z. Guo and D. S. Small, “Control function instrumental variable estimation of nonlinear causal effect models,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3448–3482, 2016. [24](#), [36](#)
- [42] A. Semykina and J. M. Wooldridge, “Estimating panel data models in the presence of endogeneity and selection,” *Journal of Econometrics*, vol. 157, no. 2, pp. 375–380, 2010. [24](#), [25](#), [27](#), [31](#)

- [43] J. Terza, “Estimating endogenous treatment effects in retrospective data analysis,” *Value in Health*, vol. 2, no. 6, pp. 429–434, 1999. [24](#), [36](#)
- [44] H. Chernoff and N. Divinsky, “The computation of maximum-likelihood estimates of linear structural equations,” [24](#)
- [45] G. C. Chow, “Two methods of computing full-information maximum likelihood estimates in simultaneous stochastic equations,” *International Economic Review*, vol. 9, no. 1, pp. 100–112, 1968. [24](#)
- [46] J. Durbin, “Maximum likelihood estimation of the parameters of a system of simultaneous regression equations,” *Econometric Theory*, vol. 4, no. 1, pp. 159–170, 1988. [24](#)
- [47] J. Grogger, “A simple test for exogeneity in probit, logit, and poisson regression models,” *Economics letters*, vol. 33, no. 4, pp. 329–332, 1990. [24](#)
- [48] M. H. Seo and Y. Shin, “Dynamic panels with threshold effect and endogeneity,” *Journal of Econometrics*, vol. 195, no. 2, pp. 169–186, 2016. [24](#)
- [49] S. Ullah, P. Akhtar, and G. Zaefarian, “Dealing with endogeneity bias: The generalized method of moments (gmm) for panel data,” *Industrial Marketing Management*, vol. 71, pp. 69–78, 2018. [25](#)
- [50] J. Fan and Y. Liao, “Endogeneity in high dimensions,” *Annals of statistics*, vol. 42, no. 3, p. 872, 2014. [25](#)
- [51] R. Jäckle and O. Himmler, “Health and wages panel data estimates considering selection and endogeneity,” *Journal of Human Resources*, vol. 45, no. 2, pp. 364–406, 2010. [25](#)
- [52] J. M. Wooldridge, “Selection corrections for panel data models under conditional mean independence assumptions,” *Journal of econometrics*, vol. 68, no. 1, pp. 115–132, 1995. [25](#)
- [53] K. E. Staub and R. Winkelmann, *Essays in Microeconometrics*. PhD thesis, 2011. [26](#), [35](#)
- [54] A. C. Cameron and P. K. Trivedi, “Essentials of count data regression,” *A companion to theoretical econometrics*, vol. 331, 1999. [32](#)
- [55] J. S. Silva and S. Tenreyro, “The log of gravity,” *The Review of Economics and statistics*, vol. 88, no. 4, pp. 641–658, 2006. [32](#)
- [56] R. Phillips, A. S. Simsek, and G. VanRyzin, “Endogeneity and price sensitivity in customized pricing,” *Columbia University Center for Pricing and Revenue Management Working Paper*, vol. 4, 2012. [32](#)
- [57] J. M. Wooldridge, “Quasi-likelihood methods for count data,” *Handbook of Applied Econometrics Volume 2: Microeconomics*, pp. 321–368, 1999. [32](#), [33](#), [35](#), [36](#)

- [58] G. W. Imbens and J. M. Wooldridge, “Recent developments in the econometrics of program evaluation,” *Journal of economic literature*, vol. 47, no. 1, pp. 5–86, 2009. [36](#)
- [59] J. Mullahy, “Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior,” *Review of Economics and Statistics*, vol. 79, no. 4, pp. 586–593, 1997. [36](#), [44](#)
- [60] J. H. Stock and M. Yogo, “Testing for weak instruments in linear iv regression,” 2002. [37](#), [54](#)
- [61] B. E. Honoré and M. Kesina, “Estimation of some nonlinear panel data models with both time-varying and time-invariant explanatory variables,” *Journal of Business & Economic Statistics*, vol. 35, no. 4, pp. 543–558, 2017. [38](#)
- [62] L. P. Hansen, “Large sample properties of generalized method of moments estimators,” *Econometrica: Journal of the Econometric Society*, pp. 1029–1054, 1982. [43](#), [47](#), [49](#)
- [63] J. G. Montalvo, “Gmm estimation of count-panel-data models with fixed effects and predetermined instruments,” *Journal of Business & Economic Statistics*, vol. 15, no. 1, pp. 82–89, 1997. [43](#)
- [64] J. Wilde, “A note on gmm estimation of probit models with endogenous regressors,” *Statistical Papers*, vol. 49, no. 3, pp. 471–484, 2008. [43](#)
- [65] T. Amemiya, “The nonlinear two-stage least-squares estimator,” *Journal of econometrics*, vol. 2, no. 2, pp. 105–110, 1974. [43](#)
- [66] J. M. Wooldridge, “Distribution-free estimation of some nonlinear panel data models,” *Journal of Econometrics*, vol. 90, no. 1, pp. 77–97, 1999. [43](#)
- [67] D. Holtz-Eakin, W. Newey, and H. S. Rosen, “Estimating vector autoregressions with panel data,” *Econometrica: Journal of the Econometric Society*, pp. 1371–1395, 1988. [43](#), [46](#)
- [68] M. B. Wintoki, J. S. Linck, and J. M. Netter, “Endogeneity and the dynamics of internal corporate governance,” *Journal of Financial Economics*, vol. 105, no. 3, pp. 581–606, 2012. [43](#)
- [69] E. L. Schultz, D. T. Tan, and K. D. Walsh, “Endogeneity and the corporate governance-performance relation,” *Australian journal of Management*, vol. 35, no. 2, pp. 145–163, 2010. [43](#)
- [70] F. Windmeijer, “Gmm for panel count data models,” 2006. [44](#), [45](#), [49](#)
- [71] A. Romeu, “Expend: Gauss code for panel count-data models,” 2004. [44](#)
- [72] J. Kim and G. Marschke, “Labor mobility of scientists, technological diffusion, and the firm’s patenting decision,” *RAND Journal of Economics*, pp. 298–317, 2005. [44](#), [45](#), [46](#)
- [73] J. M. Wooldridge, “Some alternatives to the box-cox regression model,” *International Economic Review*, vol. 33, no. 4, pp. 935–955, 1992. [44](#)

- [74] J. M. Wooldridge, “Multiplicative panel data models without the strict exogeneity assumption,” *Econometric Theory*, vol. 13, no. 5, pp. 667–678, 1997. [45](#)
- [75] F. Windmeijer, “Moment conditions for fixed effects count data models with endogenous regressors,” *Economics Letters*, vol. 68, no. 1, pp. 21–24, 2000. [45](#)
- [76] J. Jacob and T. Osang, “Democracy and growth: A dynamic panel data study,” *The Singapore Economic Review*, 2018. [46](#)
- [77] G. Wawro, “Estimating dynamic panel data models in political science,” *Political Analysis*, vol. 10, no. 1, pp. 25–48, 2002. [46](#)
- [78] D. Roodman, “How to do xtabond2: An introduction to difference and system gmm in stata,” *The Stata Journal*, vol. 9, no. 1, pp. 86–136, 2009. [46](#)
- [79] W. H. Greene, *Econometric analysis*. Pearson Education India, 2003. [49](#), [60](#)
- [80] F. A. Windmeijer and J. M. Santos Silva, “Endogeneity in count data models: an application to demand for health care,” *Journal of applied econometrics*, vol. 12, no. 3, pp. 281–294, 1997. [49](#)
- [81] B. W. Brown and W. K. Newey, “Generalized method of moments, efficient bootstrapping, and improved inference,” *Journal of Business & Economic Statistics*, vol. 20, no. 4, pp. 507–517, 2002. [49](#)
- [82] StataCorp, *Stata 15 Base Reference Manual*. StataCorp, College Station, Texas, 2018. [50](#), [58](#)
- [83] R. Venkatesan and P. W. Farris, “Measuring and managing returns from retailer-customized coupon campaigns,” *Journal of marketing*, vol. 76, no. 1, pp. 76–94, 2012. [51](#), [54](#)
- [84] A. Baik, “Personalized coupon campaigns: Understanding and directing customers,” 2015. [54](#)

Appendix

Idsiyncratic Endogeneity Test - First Stage Regression

Table 2.11 shows the partial output of the first stage out of the two-stage estimation process of the CRE/CF test for endogeneity. For simplicity, we surpress the output corresponding to the time dummies as well as their corresponding averages.

Table 2.11: First Stage LPM

	1st Stage - Pooled LPM (cluster-robust standard errors)
Response: Type A	
Type B	-.6641697*** (.0082453)
Type AB	-.3544812*** (.0124678)
Type ABC	-.5600362*** (.0128869)
Type BC	-.3959399*** (.0099097)
Type C	-.6343642*** (.0126641)
Type AC	-.3823563*** (.0146179)
Sales Per Period	-.0000127 (9.33e-06)
Time in Previous NPP	.001606 (.0000903)
Pre Days Between	-.0000137 (.0000887)
Pre Sales	-6.90e-06*** (1.95e-06)
Baseline Spend	6.17e-06*** (1.13e-06)
Control Visit Rate	-.0052313 (.0105184)
Type B Mean	-.1218269*** (.0143384)
Type AB Mean	.0660886*** (.0211507)
Type ABC Mean	-.0508333 (.0386821)
Type BC Mean	-.1443107*** (.023305)
Type AC Mean	.2660258*** (.0254528)
Type C Mean	-.245397*** (.0241971)
Lag Items per Period	-.0001966*** (.0000216)
Lag Items per Period Mean	.0001642*** (.0000273)
Baseline Visit Rate Mean	.9524872*** (.1689464)
Baseline Spend Mean	.0000574*** (.0000017)
Time in Previous NPP Mean	-.0015792*** (.0000974)
Sales Per Period Mean	7.12e-06 (.0000118)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

GMM Consistency with at least one endogenous regressor

We can generally show consistency for GMM estimation with an endogenous regressor. As a general form of our GMM problem, consider the following model with a multiplicative error specification:

$$Y_{it} = \varphi(\mathbf{X}_i, \beta) \alpha_i r_{it}$$

with $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})'$, $\beta = (\beta_1, \dots, \beta_k)'$, α_i is the individual fixed effect, and $\mathbf{r}_i = (r_{i1}, \dots, r_{iT})'$ with $\varphi(\cdot)$ an assumed continuous function. There exist an endogenous regressor x_{it}^* in $\mathbf{X}_i$ (e.g. $x_{it}^* \subset \mathbf{X}_i$) such that $E[\mathbf{X}_i' \mathbf{r}_i] \neq \mathbf{0}$.

To eliminate the fixed effect term, α_i , we use Wooldridge's transformation which results in:

$$\frac{Y_{it}}{\varphi(\mathbf{X}_i, \beta)} - \frac{Y_{i(t-1)}}{\varphi(\mathbf{X}_{i-}, \beta)} = \alpha_i (r_{it} - r_{i(t-1)}) \equiv \pi_{it}(\beta)$$

As endogeneity is still present, through correlation between the regressor and the error, $\hat{\beta}$ is indeed inconsistent. Suppose, now, there exists a vector $\mathbf{Z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{iT-2})$ composed of exogenous regressors/instruments in the given model which can include lags of the endogenous variables in $\mathbf{X}_i$. As $\mathbf{Z}$ exists, then the appropriate orthogonality condition holds:

$$E[\mathbf{Z}_i' \pi_{it}(\beta)] = \mathbf{0}$$

As β is (implicitly) defined to be a K -dimensional vector of parameters, then we can compute the sample analog/complement of $E[\mathbf{Z}_i' \pi_{it}(\beta)] = \mathbf{0}$, the unconditional population moment condition. If $\mathbf{Z}_i \in Z \subset \mathbb{R}^L$, where $L > K$, ($L \geq K$), with $\varphi(\mathbf{Z}_i, \beta) = \mathbf{Z}_i' \pi_{it}(\beta)$, then:

$$\hat{\beta}_{GMM} = \arg \min_{\beta \in \Theta} \left[\frac{1}{N} \sum_{i=1}^N \varphi(\mathbf{Z}_i, \beta) \right]' \mathbf{W}_N \left[\frac{1}{N} \sum_{i=1}^N \varphi(\mathbf{Z}_i, \beta) \right]$$

Then in such a problem above, we can show that the estimator/solution $\hat{\beta}_{GMM}$ from above is a consistent estimator of β^* .

From above, we have $\mathbf{Z}_i \in Z \subset \mathbb{R}^L$, where $L > K$, ($L \geq K$). Let $\varphi(\mathbf{Z}_i, \beta) = \mathbf{Z}_i' \pi_{it}(\beta)$ with $E[\varphi(\mathbf{Z}_i, \beta^*)] = \mathbf{0}$. Let $F(\beta) = E \left[\varphi(\mathbf{Z}_i, \beta) \right]' \mathbf{\Omega} E \left[\varphi(\mathbf{Z}_i, \beta) \right]$ and $F_N(\beta) = E \left[\frac{1}{N} \sum_{i=1}^N \varphi(\mathbf{Z}_i, \beta) \right]' \mathbf{W}_N E \left[\frac{1}{N} \sum_{i=1}^N \varphi(\mathbf{Z}_i, \beta) \right]$ where $\mathbf{W}_N$ is a positive, semidefinite weight matrix.

Consider the following assumptions

A1: Θ is a closed and bounded set

A2: $\forall \mathbf{z} \in Z, \varphi(\mathbf{z}, \beta)$ is continuous on Θ

A3: $\forall \mathbf{z} \in Z, \varphi(\mathbf{z}, \beta) \neq \varphi(\mathbf{z}, \beta^*)$ if $\beta \neq \beta^*$

A4: $\mathbf{W}_N \xrightarrow{P} \mathbf{\Omega}$

A5: The (Weak) LLN is true

Given these assumptions are true, then $\hat{\beta}_{GMM} \xrightarrow{P} \beta^*$, with $\hat{\beta}_{GMM} = \arg \min_{\beta \in \Theta} F_N(\beta)$

Proof: Given A1 and A2 holds, then $\arg \min_{\beta \in \Theta} F_N(\beta)$ necessarily has a solution. Let $\hat{\beta}_{GMM} = \arg \min_{\beta \in \Theta} F_N(\beta)$. Given the definition, then $\beta^* \in \Theta$ is a solution to $\arg \min_{\beta \in \Theta} F(\beta)$. By A3 and because $\mathbf{\Omega}$ is positive definite matrix, then β^* must be a singularly unique solution to $\arg \min_{\beta \in \Theta} F(\beta)$. Finally using A4 and A5, $|F_N(\hat{\beta}_{GMM}) - F(\beta^*)| \xrightarrow{P} 0$, therefore $\hat{\beta}_{GMM} \xrightarrow{P} \beta^*$.

CHAPTER 3

Estimating Conditionally Serially Correlated Generalized Additive Mixed Models with Parametric Factor Smooths for Repeated Counts

Abstract

Both balanced and unbalanced longitudinal data structures are prone to exhibit a complex covariance structure where variability arising from repeated responses and conditional, within-subject temporal correlations can exist. Failure to properly account for a proper covariance structure in repeated measures can have severe consequences on the validity of statistical inference due to the distortion of standard errors. As this is well-studied in the generalized linear mixed model context, we seek to investigate conditionally serially correlated longitudinal models for count outcomes in which we eschew the constraint of a pure linear predictor to be strictly parametric and focus on generalized additive mixed models. Via two broad simulation designs, we investigate parameter performance of model components in which a portion of the predictor is estimated via nonparametric smooths that are allowed to vary across levels of a parametric factor covariate of interest using penalized quasi-likelihood (PQL) estimation. The retailer scanner-panel data is further analyzed in order to quantify a potentially nonlinear, stimulated effect of prior promotional period campaigns' on in-store visits during the direct subsequent period of non-promotion for households participating in the loyalty promotional program. We seek to extend the literature of semi-parametric mixed models via the simultaneous estimation of variance components, fixed parameters, and unknown factor smooths using low-rank, optimal thin plate regression splines for conditionally serially correlated longitudinal counts in cases of stationary and non-stationary panel structures to identify the functional form of prolonged stimulation of a potential carryover effect in a marketing promotional program.

3.1 Introduction

In this chapter, we seek to exploit the flexibility offered by generalized additive mixed models (GAMMs) which augments the traditional generalized linear mixed model (GLMM) by utilizing smooth functions of covariates with consideration for factor smooth interactions. These smooths are spline functions of explanatory covariates which themselves can be constructed as a linear combination of spline bases. Any spline-based function's interaction with a factor smooth enables the ability to assess the effect of a smoothed covariate across levels of a factor in which each level's function is learned non-parametrically from the data. Additionally, as in most cases of any models fit to longitudinal data, the consideration for the complex covariance structures that are likely associated with observing a cross-section of interest repeatedly over time is crucial. From a longitudinal modeling perspective, the among-individual variability across the subjects of a cross-section, typically modeled by a random effect, and the likely existence of time-dependent, within-level correlation in a subject, typically modeled with a temporal parameter, are needed in order to mitigate potential confounding effects ([1]). As in the traditional generalized linear mixed model, the response measurements within a subject are likely more homogenized than between subjects, requiring the use of a random effect to account for the portion of variability surrounding conduct influencing the response measurements occurring among subjects, on average ([2]). However, the inclusion of this style of random effect only presumes that the repeated response measurements are conditionally independent. In theory, a portion of an observed, cross-sectional unit's behavior could be affected by a temporal, underlying stochastic process within the unit where there exists significant correlation between responses. This phenomenon is known as serial correlation and can occur conditional upon the random effect. The joint omission of the variance components needed to properly account for serially correlated observations given the random effects leads to biases of fixed effects of interest due to the untrustworthy precision of estimating associated standard errors as well as statistical inference invalidity due to the distortion of p-values and confidence interval construction attributed to inflated Type I error rates ([3],[4], [2], [5]). Conversely, accounting for these variance components leads to a covariance structure that aids in correct inference in the fixed, parametric portion of our model structure in addition to proper estimation of the variance in the response of repeated measures. However, to what extent does the inclusion of a factor smooth specification of a generalized additive mixed model complicate the estimation of procedure of variance components corresponding to serial correlation given the random effect as well as any strictly parametric effects? Moreover, what utility does the conversion of the extended GAMM to an equivalent GLMM form offer in the presence of estimation fixed effects, variance components, and smoothing functions? Is variance component estimation hindered greatly by balanced vs. unbalanced longitudinal structures? We offer an exploration into the simultaneous or joint estimation of the unknown smooth functions, associated factor smooth interactions, any strictly parametric fixed effects, an arbitrary random effect, and a specified temporal correlation parameter conditional upon the random effect.

Continuing the analysis from Chapter 2, we are specifically interested in potential nonlinearity of the prior promotional campaign’s exposure effect on our response over the weeks spanning the first, intermittent epoch of non-promotion. We utilize these semi-parametric models that can isolate these nonlinearities while controlling for parametric components and the variance components related to the covariance structure. In addition to the individual heterogeneity or interdependence of visiting behavior that can be accounted for by random effects, there is likely an underlying time-varying, temporal process attributing to random variation operating within the households conditional upon the random effect.

As we still have a longitudinal structure, we can utilize mixed models to employ a random effects structure that accounts for hypothesized correlated responses within our cross-section. However, as previously stated, we allow for a covariate in our linear predictor to be estimated non-parametrically via a smooth function to vary smoothly over levels of a factor of interest. The smooth functions can be estimated in a variety of different ways such as cubic splines, cyclic splines, B-splines, P-splines, and thin-plate regression splines ([6]). We focus primarily on the utility of employing thin-plate regression splines. Thin-plate regression splines offer an invaluable utility relative to the other aforementioned splines. The degree of subjectivity of choosing the bases are not particularly clear (B-spline vs. P-spline). Such splines require the choice of manually selecting knot locations. The knots add uncertainty induced by the subjective choice in the fitting process of any GAMM of interest, whereas the knot-free, thin-plate regression splines circumvent this issue ([7]).

Before delving into the specifics regarding the estimation process for assessing the prior promotional campaign exposures’ hypothesized lasting impacting into the first non-promotional time interval, we offer a brief explanation regarding the change of model form. In the previous section, we utilized an unobserved effects panel data model to assess the campaign exposure effects on visit rate relative to periods of non-promotion within the test cycle. Here, we posit that it suffices to rely on methods as it relates to semi-parametric mixed effects models instead of the unobserved fixed effects specification in the previous chapter. In the specific GAMM model of interest for the first intermittent cycle of promotion beyond the baseline, we are able to account for repeated measures of households by specifying a random effect term into our model under the assumption that there is no correlation between the included, parametric effects and any of the model’s unobservables. This assumption presumes that the covariates defining the parametric portion of our model does not suffer from an idiosyncratic endogeneity bias. In the previous chapter, we did empirically show that the personalized rewards were indeed marginally endogenous in the promotional period relative to periods of non-promotion. However, in the case of these models and the epoch of interest in this application, we specify the type of campaign received in the prior promotional period in the parametric portion of our models, but here it is treated as a time-constant, strictly parametric effect that is not likely to be correlated distinctly with time-varying unobservables that span the immediate non-promotional period’s epochs (or the first washout period after the first promotional period). Therefore, the exogeneity of regressors in our parametric part of linear predictor holds by assumption ([8]). Thus, we can utilize random effects here as an

avenue to account for repeated measures or clustering of households' visiting activity over weeks spanning the non-promotional periods with our generalized additive mixed models. Additionally, there likely exists a correlation component for households' visiting behavior conditional upon the random effect, requiring an additional parameter that is needed to capture this temporal structure. As will be formally articulated in the application, we expect that a previous campaign's exposure effect on household visiting behavior to attenuate in a potentially nonlinear manner across the weeks that span the non-promotional period, while presuming that temporal correlations exist among the visits per week of the cross-sectional household i over the time frame. To capture this potential nonlinear effect in the response over time given the prior promotional campaign, we apply generalized additive mixed models with factor smooth interactions to estimate the effects of the parametric component of prior campaign types over time on the weekly visit rate ratio over the baseline visiting behavior. Logically, we expect that there is potential stimulation of the prior campaigns' exposure effects that indicate an increased visits beyond the visiting behavior observed in the baseline (before any campaigns were distributed) that exhibits nonlinearity over time.

This chapter has three primary focal points which contribute to the literature. Particularly in marketing and business applications, generalized additive mixed models are not fully explored in terms of a standard modeling structure in the current literature, and factor smooth interactions are not often utilized in exploring the nonlinearity that can exist in parametric effects over time. Secondly, we highlight the process of recasting the generalized additive mixed model (GAMM) with a factor smooth into a standard generalized linear mixed model (GLMM) for convenient estimation utilizing a penalized quasi-likelihood (PQL) approach in which the parameter that controls the bias/variance tradeoff for the inclusion of smoothing function are concurrently estimated along with the variance components. Thirdly, and most importantly, we conduct simulation exercises in order to utilize generalized additive mixed models for the estimation conditionally serially correlated counts that are applied in two scenarios: inclusion of time-invariant covariates (stationary) exclusively as well as the allowance for time-varying covariates (non-stationary) in both balanced and unbalanced designs for longitudinal count responses. These two simulation designs will enable us to study how well the variance components and the parameter associated with the fixed effect are estimated while assessing the factor smooths' approximation to known, designated "shapes", or test functions, utilizing the optimality associated with thin plate regression splines. To our knowledge, this chapter's investigation is the first foray into observing the finite sample performance of the simultaneous estimation of assessing the joint estimation of factor smooths, fixed effect parameters, and covariance parameter estimation of GAMMs exclusively for conditionally serially correlated counts. We also consider cases in which our GAMM models are essentially misspecified where the random effect's variance is set to zero or the autoregressive correlation coefficient is null. The remainder of this chapter is outlined as follows: Section 3.2 will briefly highlight the background of estimating smooth function for a non-parametric regression model as well as an overview of popularly utilized splines including the thin-plate regression spline. Section 3.3 highlights the formal specification of a generalized additive mixed model (GAMM) as well as the process of its conversion into

an over-parameterized generalized linear mixed model. Section 3.4 highlights the penalized quasi-likelihood (PQL) estimation process of the over-parameterized GLMM as well as commonly used temporal correlation structures. Section 3.5 presents two simulation designs of conditional serially first-order correlated models for counts from both stationary and non-stationary perspectives. We conclude with Section 3.6 with an application that utilizes these GAMM models for estimating weekly visit rates in an intermittent promotional cycle utilizing the quasi-experiment data presented in Chapter 2. To aid in our analyses, we explicitly rely on the **mgcv** package (with a host of other subfunctions) in R ([9]).

3.2 Nonparametric Curve Fitting: A single covariate

Consider the regression model

$$y_i = g(x_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad (3.1)$$

where y_i represents the response for the cross-sectional unit i , $g(\cdot)$ is a smooth function of a covariate x_i , and ϵ_i is an error term assume iid Gaussian random variables. As $g(\cdot)$ has an unknown structure, there are a variety of ways or methods to represent “curvature” (nonlinearity) in model (2.1) including, but not limited to: locally weighted scatterplot smoothing (LOWESS), moving averages, polynomial regressions, fractional polynomials, and regression splines. Of particular interest to this manuscript is the notion of *penalized* regression splines which are simply defined as piecewise polynomials of an arbitrary degree. The segments or points that join polynomials are known as knots ([10]). Additionally, for penalized regression splines, there is a Lagrangian penalty associated with finding the best fit to the data. I highlight two brief, but essential spline forms in the following two subsections.

3.2.1 Regression Splines

As a brief illustration to highlight the nonparametric of the model (3.1), the function of unknown form $g(\cdot)$ can be estimated via a spline. For example, a polynomial of degree p can be written for $g(x)$ as:

$$g(x) = \sum_{k=1}^K b_k (x_i - t_k)_+^p + \beta_0 + \beta_1 x_i + \dots + \beta_p x_i^p \quad (3.2)$$

where $(x_i - t_k)_+^p \equiv \max\{(x_i - t_k)^p, 0\}$, $(t_1, \dots, t_k)$ are knots, and $\beta = (\beta_0, \dots, \beta_p)'$ with $\mathbf{b} = (b_1, \dots, b_k)'$ are coefficients. Given this estimator, we can reformulate (3.1) using notation for a matrix as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \quad (3.3)$$

where $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{X} = (1, x_i, \dots, x_i^p)$, $\mathbf{Z} = ((x_i - t_1)^p, \dots, (x_i - t_K)^p)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$. Thus, equation (3.3) can be estimated via OLS to obtain a prediction for $\mathbf{y}$.

3.2.2 Penalized Splines

As seen previously, while regression splines are simple computationally, they are highly unpopular due to the associated subjectivity involved with the knots (both in terms of the number chosen and their placement). Instances in which knots are non-conservatively chosen can result in a large degree of overfitting¹ ([11], [12]). An approach that mitigates the overfitting issue would be to employ penalized splines. Explicitly, the overfitting issue can be avoided via a mandated constraint imposed on the spline coefficients ([9]).

The minimization problem can be considered for a penalized spline from (3.3) as

$$\min_{\beta, \nu} ||\epsilon|| \quad \text{s.t.} \quad \sum_{k=1}^K b_k^2 \leq a \quad (3.4)$$

where $\epsilon = \mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\nu$, b_k is the spline coefficients, and a is some arbitrary constant. An equivalent analog the minimization problem is

$$\begin{aligned} P(\beta, \nu) &= ||\epsilon||^2 + \lambda \nu' \nu \\ &= ||\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\nu||^2 + \lambda \nu' \nu \\ &= ||\mathbf{y} - \mathbf{A}\theta||^2 + \lambda \theta' \mathbf{D}\theta \end{aligned} \quad (3.5)$$

where $\mathbf{A} = [\mathbf{X}, \mathbf{Z}]$, $\theta = (\beta', \nu')'$, and $\mathbf{D} = \text{blockdiag}(\mathbf{0}, I_K)$. The parameter λ is a Langrangian term or smoothing parameter that mitigates the smoothness of the fitted curvature. Essentially, in a penalized spline setting, it controls the tradeoff between variance and bias. Smaller values of λ results in a curve that has less bias yet higher variability (as the curve fits closer to the data, hence “overfitting”), whereas the opposite (higher values) lead to more bias yet less variability. This parameter be chosen from a variety of different techniques including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or Mallows Cp criterion, among others ([9]). Each estimation techinque for the Langragian is dependent in its effectiveness, for example, on known versus unknown scale parameter for non-Gaussian responses or correlated data ([11]).

However, the primary benefit offered by the penalized spline setting in general is its direct connection to mixed models with either Gaussian and non-Gaussian responses. This link enables us to subsume the smoothing parameter in a mixed model framework which is easily estimable with software. Next, we turn to the spline of choice considered in this chapter: thin plate regression splines.

¹Overfitting is the phenomenon in which a function is fit too closely to a set of data points.

3.2.3 Thin Plate Regression Splines

There are an abundance of bases that can be chosen for fitting of the nonparametric component, but we focus on thin plate regression splines² as it is advantageous over competitors in two ways: a certain degree of ‘optimality’ achieved by simultaneously mitigating the conflicting issues of the construction of estimated smooth functions and finding functions that best suits the resulting smooth objective function ([13]). Clear disadvantages of this procedure, however, is the computational cost relative to choosing knots in other bases and the lack of invariance to rescaling of covariates. This method also is suitable for any number of predictor covariates and necessitates no a priori knowledge of the data’s functional form.

We outline the theory of thin plate splines here and discuss the thin plate regression spline approximation technique used in the **mgcv** package. For example, to estimate a smooth function $t(\mathbf{x})$ from data $(y_i, \mathbf{x}_i)$, such that $y_i = t(\mathbf{x}_i) + \epsilon_i$, the thin plate spline, first proposed by [14], estimates $t(\cdot)$, a functional smooth of a covariate, by finding the estimated function $\hat{c}$ which minimizes:

$$\|\mathbf{y} - \mathbf{c}\|^2 + \lambda J_{md}(c)^3 \quad (3.6)$$

where $\mathbf{c} = [c(\mathbf{x}_1), \dots, c(\mathbf{x}_n)]'$ represents a vector of all unknown smoothing functions, and J_{md} is the penalty function which measures the nonlinearity or degree of “wiggleness” or curvature of c at $\mathbf{x}_i$. The sign of this functional states whether the curvature is convex or concave, which is then squared and integrated over all $\mathbf{x}_i$ for assessing the average curvature. The indices m and d corresponds to the dimension of J , with m being the penalty order derivative, d corresponding to the size of the vector $\mathbf{x}_i$, and $(\nu_1, \dots, \nu_d)$ are terms required to necessitate that all specified derivatives are taken with respect to all parameters in the combination. In the general case, the penalty function has the explicit form of

$$J_{md} = \int \dots \int_{\mathcal{R}^d} \sum_{s_1 + \dots + s_d = m} \frac{m!}{s_1! \dots s_d!} \left(\frac{\partial^m c}{\partial x_1^{s_1} \dots \partial x_d^{s_d}} \right)^2 dx_1 \dots dx_d \quad (3.7)$$

In the case of one smoothed covariate in which there is an arbitrary penalization of the second degree derivative, then $d = 1$, $m = 2$, and two smooth covariates of the same degree, (3.7) takes the two forms

$$J_{21} = \int \left(\frac{\partial^2 c}{\partial x_1^2} \right)^2, \quad J_{22} = \int \int \left(\frac{\partial^2 c}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 c}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 c}{\partial x_2^2} \right)^2$$

Thus, the J_{md} term employs a penalized matrix consisting of terms based on the squared derivatives of basis functions. Provided that $2m > d^4$, minimizing (3.6) has the expression⁵ ([15], [14])

²A spline is a sum of weighted basis functions, and its type impacts the form and properties.

³This can easily be seen as the penalized least squares form where we use $c(\mathbf{x}_i)$ to predict y_i .

⁴Using the **mgcv** package, by default, “visually smooth” results/functions are chosen to the smallest value which satisfies $2m > d + 1$.

⁵Generalizable to any number of dimensions, d , and any order of differential, m .

$$c(x) = \sum_{j=1}^M a_j \phi_j(\mathbf{x}) + \sum_{i=1}^n \delta_i \eta_{md}(l_i)$$

where $l_i = \|\mathbf{x} - \mathbf{x}_i\|$, $\mathbf{a}$ and $\boldsymbol{\delta}$ are estimable parameters⁶, $M = \binom{m+d-1}{d}$ are the total number of functions, ϕ_j , that span the linear polynomials in $\mathcal{R}^d$ for which $J_{md} = 0$. In other words, the ϕ_j receive zero penalty and span the “null space” of J_{md} . Lastly, η_{md} represents the remaining bases, or the radial basis functions, defined as a piecewise function, for d_{even} and d_{odd} for a given m, d choice as ([9])

$$\eta_{md}(l) = \begin{cases} l^{2m-d} \frac{\Gamma(\frac{d}{2} - m)}{2^{2m} \pi^{\frac{d}{2}} (m-1)!} & d_{\text{odd}} \\ l^{2m-d} \log(l) \frac{-1^{m+1+\frac{d}{2}}}{2^{2m-1} \pi^{\frac{d}{2}} (m-1)! (m - \frac{d}{2})!} & d_{\text{even}} \end{cases}$$

Therefore, the function c is now partitioned, for a thin plate spline, into two distinct “parts”: M polynomials that affect the entire data (ϕ_j s) and a set of radial basis functions ($\eta_{md}(\cdot)$ s).

In terms of constructing the minimization problem along the lines of (3.4), we define the design matrix for a thin plate spline as: $\mathbf{X} = X_{ij} = \phi_j(x)$ and $\mathbf{Z} \equiv Z_{ij} = \eta_{md}(l_i)$ with associated coefficient vectors $\mathbf{a}$, and $\boldsymbol{\delta}$. The minimization problem for a thin plate spline basis is then constructed as follows

$$\min_{\mathbf{a}, \boldsymbol{\delta}} \|\mathbf{y} - \mathbf{X}\mathbf{a} - \mathbf{Z}\boldsymbol{\delta}\|^2 + \lambda \boldsymbol{\delta}' \mathbf{Z} \boldsymbol{\delta} \quad \text{s.t.} \quad \mathbf{X}' \boldsymbol{\delta} = \mathbf{0} \quad (3.8)$$

Notice the similarity in form to (3.5).

For practitioners, employing a thin plate smoothing spline is generally computationally prohibitive for parameters of interests particularly in higher dimensions as it requires estimating as many parameters as unique data points exist in applied settings (a computational cost of $O(n^3)$). An efficient approximation, proposed by ([7],[9]), to thin plate splines are *thin plate regression splines*, utilizing a low-rank (significantly fewer coefficients than data that exists) approximation that involves a truncated eigenvalue decomposition of the penalized basis (the basis for the parameter space of $\boldsymbol{\delta}$) while simultaneously leaving the unpenalized basis unaltered⁷. In essence, a spectral decomposition is utilized on $\mathbf{Z}$ as $\mathbf{Z} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}$, where the eigenvalues composing $\mathbf{Z}$ are sorted so that $|\Lambda_{i-1,i-1}| \leq |\Lambda_{i,i}|$ where $\mathbf{Q}$ are associated eigenvectors. Defining $\mathbf{Q}_k$ to be a matrix consisting of the first columns of $\mathbf{Q}$ of size k , with $\boldsymbol{\Lambda}_k$ defined to be the upper left, symmetrical submatrix of size $k \times k$ of $\boldsymbol{\Lambda}$. Regulating the coefficient vector $\boldsymbol{\delta}$ to the column space of $\mathbf{Q}$ as $\boldsymbol{\delta} = \mathbf{Q}_k \boldsymbol{\delta}_k$ leads to the constrained minimization problem:

$$\min_{\mathbf{a}, \boldsymbol{\delta}_k} \|\mathbf{y} - \mathbf{X}\mathbf{a} - \mathbf{Q}_k \boldsymbol{\Lambda}_k \boldsymbol{\delta}_k\|^2 + \lambda \boldsymbol{\delta}_k' \mathbf{Z}_k \boldsymbol{\delta}_k \quad \text{s.t.} \quad \mathbf{X}' \mathbf{Q}_k \boldsymbol{\delta}_k = \mathbf{0}$$

⁶The parameter vector $\boldsymbol{\delta}$ is constrained such that $\mathbf{X}' \boldsymbol{\delta} = \mathbf{0}$.

⁷The unpenalized basis for the parameter space for $\mathbf{a}$ is left untouched in the low rank approximation as $\phi_j(x_i)$ spans the kernel of the penalty measure.

respecting parameter vectors $\mathbf{a}$ and $\boldsymbol{\delta}_k$. Moreover, the constrained minimization problem can be recasted into an unconstrained minimization problem by exerting a QR decomposition on $\mathbf{Q}_k' \mathbf{X}$ to obtain an orthogonal column basis $\mathbf{L}_k$ such that $\mathbf{X}' \mathbf{Q}_k \mathbf{L}_k = \mathbf{0}$ holds true. Again, restricting the column space $\boldsymbol{\delta}_k$ as $\boldsymbol{\delta}_k = \mathbf{L}_k \hat{\boldsymbol{\delta}}$ yields the unconstrained, “rank k ”, minimization problem:

$$\min_{\mathbf{a}, \boldsymbol{\delta}} \|\mathbf{y} - \mathbf{X}\mathbf{a} - \mathbf{Q}_k \boldsymbol{\Lambda}_k \mathbf{L}_k \hat{\boldsymbol{\delta}}\|^2 + \lambda \hat{\boldsymbol{\delta}}' \mathbf{L}_k' \boldsymbol{\Lambda}_k \mathbf{L}_k \hat{\boldsymbol{\delta}} \quad (3.9)$$

where $k < n$ ⁸. The computational cost is decreased to $O(k^3)$ to minimize (3.9). Further details can be found in [7] and [9].

It is possible to incorporate the non-parametric smoothing function (or multiple smoothing functions) into a larger, generalized linear mixed model setting. Specifically, in (3.4), the constraint $\sum_{k=1}^K b_k^2 \leq a$ in the penalized spline setting is equivalent to assuming $b_k \sim N(0, \sigma_b^2)$. In this case, given finite variance (e.g. $\sigma_b^2 < \infty$), the spline coefficients (the b_k s) shift toward the mean resulting approximating smoother fits. Therefore, the penalized spline regression problem can be absorbed into a traditional mixed model framework for both Gaussian and non-Gaussian responses where now a portion of the predictor is specified in terms of smoothing functions, yielding a (generalized) additive model. In this setting, the smoothing parameter λ associated with the non-linear component can now be estimated “automatically” as a variance component via maximum likelihood, restricted maximum likelihood, or penalized quasi-likelihood estimation. The non-parametric smoothing function may also be combined multiplicatively with a known covariate as well, resulting in a factor smooth interaction. Random effects may also be employed in order to account for correlation between observations likely present in longitudinal studies. Therefore, the model construction using strictly linear covariates with associated parametric coefficients, a random effects design matrix with associated coefficients for each cross-sectional unit i , and unknown smoothing functions of covariates, additively, are known as generalized additive mixed models (GAMMs).

3.3 Generalized Additive Mixed Model

The generalized additive [mixed] model (GA[M]M) as a regression modeling method has been explored in a myriad of domains since the turn of this century such as: agriculture ([16], [17]), biodiversity ([18], [19]), life sciences and biomedicine ([20], [21]), and demography ([22]). Extending the generalized additive usage beyond univariate responses have also been explored ([23], [24]). This model is simply a generalized linear mixed model (GLMM) in which an additive portion of predictor can be modeled non-parameterically utilizing smooth functions of covariates ([25]). A GAMM’s form mirrors the GLMM in terms of parameterization requiring that the response variable follows a probability distribution. The non-parametric functions’ covariates in the predictor in addition to the mandated probability distribution for the response defines the semi-parametric identity of this general class of models. The structure of this model form allows both the

⁸As $\mathbf{Z}$ is a symmetric matrix of size $(n \times n)$, then a submatrix for the approximation is $\hat{\mathbf{Z}}_k$ of size $(n \times k)$.

estimation of non-parametric smoothing functions and parametric coefficients in unison, and the response variable of interest can be predicted by summing up the smoothing function as well as the fixed effects (defining the ‘additivity’ nomenclature in the acronym ‘GAMM’). This semi-parametric model formulation does not force us to completely abandon the linearity that presumably still exists between certain covariates of interest and the conditional mean of our response variable (via a link function), but broadens the flexibility of GLMMs in terms of adding non-parametric smoothers. This model grants several advantages such as ease of interpretation, flexible predictor functions that allows uncovering of hidden patterns in the data, and regularization of predictor functions that helps to circumvent overfitting via functional penalties ([9]). Additionally, these models can incorporate a form of interactions called factor smooths. The factor smooth interactions we will investigate measure the interactive effect by including both a categorical and continuous covariate for separate smooth functions.

As previously stated, these models are an intuitive extension of the traditional generalized additive models that has the added benefit of incorporating random effects ([26]) as well as its utility in modeling clustered, correlated observations from residuals found most often in longitudinal data structures as well as spatio-temporal data structures ([27], [9], [28]). In the case in which a generalized additive models omits the individual random effect, there becomes a risk of estimation of parameters in two ways: the likelihood of omitting clear correlation that exist in the measurement of our dependent variables within subjects as well as the independence assumption of residuals as well ([5]). Consequences for omitting a random effect specification could result in underestimation of associated standard errors with our parameter estimates which leads to overestimation in the precision of our estimates. We now turn the model form with associated factor smooths.

3.3.1 Generalized Additive Mixed Model formulation with Parametric Factor Smooths

We consider modeling our generalized additive models with random effect consideration given a longitudinal data structure. We focus on an augmented semi-parametric model form that allows for the functional form of a (continuous) covariate’s effect on the conditional mean to vary according to levels of a factor covariate nonlinearly. The response variable, y_{ij} , denotes the cross-sectional unit i in cluster j which is assumed to (conditionally) belong to a member of the exponential family⁹. The GAMM of interest is

$$H(E[y_{ij}|\mathbf{x}_{ij}^*, \mathbf{z}_{ijR}, \boldsymbol{\nu}_i]) = H(\mu_{ij}) = \boldsymbol{\eta} = \mathbf{x}_{ij}^{*'}\boldsymbol{\vartheta} + f_1(w_{1ij})d_{ij} + \dots + \mathbf{z}_{ijR}'\boldsymbol{\nu}_i \quad i = 1, \dots, N \quad j = 1, \dots, n_i \quad (3.10)$$

⁹The exponential family density can be expressed as $f(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ijR}, \boldsymbol{\nu}_i) = \exp\left(\frac{y_{ij}\xi_{ij}^{\nu_i} - \zeta(\xi_{ij}^{\nu_i})}{\phi\delta_{ij}^{-1}} + c(y_{ij}, \phi)\right)$ where $\zeta(\cdot)$ and $c(\cdot, \cdot)$ are arbitrary, known functions, δ_{ij} is an known, prior weight (e.g. the denominator in the binomial distribution), ϕ is the dispersion parameter, and ξ_{ij} is the natural parameter. Also, note that the distribution assumption for the response may be relaxed and specified exclusively via a mean-variance relationship for y_{ij} for a quasi-likelihood approach.

where $E[y_{ij}|\mathbf{x}_{ij}^*, \mathbf{z}_{ij}, \boldsymbol{\nu}_i] = \mu_{ij}$ is the conditional expected value of the response variable given all covariates considered and random effects, $H(\cdot)$ is the specified, differentiable link function that is either strictly decreasing or increasing (monotonic), $\boldsymbol{\eta}$ represents the full additive predictor, $\mathbf{x}_{ij}^* \equiv (1, x_{1ij}, \dots, x_{pij})'$ is the (time-varying) covariate vector of size p of cross-section i in cluster j whose effects are assumed to be parametric (with the inclusion of an intercept), $\boldsymbol{\vartheta}$ corresponds to the associated parametric parameter vector, f_1 is an unknown, smooth, twice-differentiable, and centered¹⁰ function of a predictor, denoted as a weighted sum of basis¹¹ functions as: $f_j(x_j) = \sum_{i=1}^{K(j)} b_i^{(j)}(x_j)\beta_i^{(j)}$, where $b_i^{(j)}(x)$ is the i^{th} basis function for covariate j with an associated, estimable weight coefficient $\beta_i^{(j)}$, d_{ij} ¹² is a parametric, linear, factor covariate that interacts with a smooth function $f_1(\cdot)$ that belongs in $\mathbf{x}_{ij}^*$, ... specifically represents remaining (and allowable) twice-differentiable, smooth functions of predictors $f(\cdot)$, w_{1ij} is a covariate that is presumed to have a non-linear effect on the response and not necessarily a part of $\mathbf{x}_{ij}^*$, $\boldsymbol{\nu}_i$ is a vector of Gaussian assumed random effect coefficients (e.g. $\boldsymbol{\nu}_i \sim N(0, \mathbf{D}(\boldsymbol{\rho}))$ where $\mathbf{D}(\boldsymbol{\rho})$ is an assumed diagonal, positive definite covariance matrix of that depends on the parameter vector $\boldsymbol{\rho}$ ($\boldsymbol{\rho}$ is a unknown parameter vector of variance components for the covariance matrix associated with the random effects of size $q \times 1$), $\mathbf{z}_{ijR} \equiv (z_{1ijR}, \dots, z_{qijR})'$ is the random effects model matrix, and $H(\cdot)$ is the link function with its existing inverse $H^{-1}(\cdot)$.

The means are given as μ_{ij} with the conditional variances $\text{Var}(y_{ij}|\boldsymbol{\nu}_i) = \sqrt{\mathbf{C}}\mathbf{R}_i(\boldsymbol{\rho})\sqrt{\mathbf{C}}$. The conditional variance is a function of $\mathbf{C}$, which is a diagonal matrix that has the variance function $\phi v(\mu_{ij})$ of the model¹³, and $\mathbf{R}_i(\boldsymbol{\rho})$ is a block-diagonal correlation matrix that depends upon a parameter vector $\boldsymbol{\rho}$ for each subject in this repeated measure context.

We can express (3.10) more simply, where we group observations into a single cluster where $\mathbf{X}_i' = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})$ and $\mathbf{Z}_{iR}' = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})$ resulting in

$$H(E[\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_{iR}, \boldsymbol{\nu}_i]) = H(\boldsymbol{\mu}_i^\nu) = \boldsymbol{\eta} = \mathbf{X}_i^* \boldsymbol{\vartheta} + f_1(\mathbf{w}_{1i})\mathbf{d}_i + \dots + \mathbf{Z}_{iR} \boldsymbol{\nu}_i \quad i = 1, \dots, N \quad (3.11)$$

with definitions remaining as before¹⁴ with the exception that $\mathbf{X}_i^*$ is the explicit i^{th} row of the model matrix for the parametric covariates in our model. The previous equations given represents the relationship between the linear predictor and the conditional expected value of the outcome (modeled as a sum of potentially nonlinear functions ($f_i(\cdot)$) of covariates of interest in addition to the strictly parametric component and a vector of random effects). Additionally, if it is assumed that there exists additional temporal correlation structure conditional upon the correlation accounted for by the random effects due to the repeated observations of the cross-section, the longitudinal responses themselves are thus conditionally

¹⁰These unknown smooth functions are subject to centering constraints such that a smooth function $f(\cdot)$ can be partitioned by levels of a factor variable such that the data for each level of the factor has the same mean or constant/intercept term in the model.

¹¹A basis is a set of functions that collectively span a space of smooths that contains the true $f(x_j)$ or a close approximation to it.

¹²Note, d_{ij} is an assumed covariate in $\mathbf{x}_{ij}^*$.

¹³The variance functions contained in $\mathbf{C}$ explicitly expresses the variances as a function of the mean. Note, ϕ is the dispersion parameter, multiplicative to $v(\cdot)$, which is the variance function.

¹⁴In matrix notation, we can write $H(E[\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{u}]) = H(\boldsymbol{\mu}^\nu) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + f_1(w_{1ij})\mathbf{d} + \mathbf{Z}_R \mathbf{u}$.

serially correlated given the random effect. The inclusion of a temporal structural warrants an additional correlation parameter using an AR(1) correlation model is

$$\text{Corr}[Y_{ij}, Y_{ik} | \boldsymbol{\nu}_i] = \rho^{k-j} \quad \forall (j, k) \in (1, \dots, n_i), \quad k \geq j$$

Beyond this correlation AR(1) model, there exists a variety of other extensions such as linear spatial correlation, compound symmetry correlation, autoregressive moving average process (ARMA), first-order, moving average correlation model MA(1) ([29]). In a pure linear mixed model (LMM) setting, [30] previously focused on the “conditional independence” case, where $\mathbf{y}_i$ are independent normal variates with a marginal variance as: $\text{Var}(\mathbf{y}_i) = \mathbf{V} = \sigma^2 \mathbf{I} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' \equiv \mathbf{R}_i(\boldsymbol{\rho}) + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' \equiv \mathbf{R}_i + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i'$. Another subsequent work, [31], extended the linear mixed model to allow for an unbalanced longitudinal structure where within-individual errors follows an AR(1) time series process. We focus specifically on the simplicity of a conditionally, temporal correlation structure offered by a first-order, autoregressive AR(1) model for counts in the proposed GAMM after first recasting the model into an over-parametrized GLMM.

3.3.2 Recasting GAMM to Penalized GLMM with parametric factor smooth

Following similarly to [32], we now outline the process of recasting our GAMM in the previous section into a GLMM, which involves partitioning each smoothing function into two components: “unpenalized” and “penalized”. The process ultimately will allow these two components to be subsumed into the traditional fixed effects and the random effects matrices of a traditional GLMM. Recall, the nonparametric component of the GAMM in the previous section is represented as

$$f_j(x_j) = \sum_{i=1}^{K_j} b_i^{(j)}(x_j) \beta_i^{(j)}$$

where the basis is the component $b_i^{(j)}$ for the j^{th} covariate. For each nonlinear smoother for covariate j , $f_j(x_j)$, there is an associated penalty, $J(c) = \boldsymbol{\beta}' \mathbf{S} \boldsymbol{\beta}$, which is a measure of the wiggleness used to circumvent overfitting ([7], [33]). With an arbitrary basis, we can define a model matrix as $\ddot{\mathbf{X}}_{\mathbf{j}}$ for each unknown smooth corresponding to a covariate j . A vector $\mathbf{f}_{\mathbf{j}}$ is defined such that $\mathbf{f}_{\mathbf{j}_i} = f_j(x_{ji})$ with $\ddot{\boldsymbol{\beta}}_{\mathbf{j}} = [\beta_{j1}, \dots, \beta_{jK_j}]'$. Thus, $\mathbf{f}_{\mathbf{j}} = \ddot{\mathbf{X}}_{\mathbf{j},ik} = \ddot{\mathbf{X}}_{\mathbf{j}} \ddot{\boldsymbol{\beta}}_{\mathbf{j}}$ with $\ddot{\mathbf{X}}_{\mathbf{j},ik} \equiv b_{jk}(x_{ji})$ ([32],[9]). To ensure identifiability¹⁵ of the model presented in (3.11), centering constraints are employed such that the sum of the components of $\mathbf{f}_{\mathbf{j}}$ are zero. This can be expressed as

$$\mathbf{1}' \ddot{\mathbf{X}}_{\mathbf{j}} \ddot{\boldsymbol{\beta}}_{\mathbf{j}} = 0$$

¹⁵In (2.11), the smooth functions are confounded with the intercept. Thus, the model is only estimable with proper constraints on $\mathbf{f}_{\mathbf{j}}$.

The constraint imposed imply a QR factorization¹⁶, which requires constructing a matrix $\mathbf{P}$, with $K_j - 1$ orthogonal columns, and such that $\mathbf{1}' \ddot{\mathbf{X}}_j \mathbf{P} = \mathbf{0}$ holds.

The unknown smooth is now redefined, for the $K_j - 1$ parameters from the matrix $\mathbf{P}$, as β_j where $\ddot{\beta}_j \equiv \mathbf{P} \beta_j$. Therefore, the adjusted model matrix for term j is $\mathbf{X}_j = \ddot{\mathbf{X}}_j \mathbf{P}$ such that $\mathbf{f}_j \equiv \mathbf{X}_j \beta_j$ satisfies the centering constraint. In other words, $\mathbf{f}_j$ is now orthogonal to the intercept which, for inference, results in the smallest width for confidence intervals.

As the model matrices satisfies the centering constraint, we have the corrected, identifiable model form for (3.11) as

$$H(\mu) = \mathbf{X}_i^* + (\mathbf{W}_1 \beta_1) \mathbf{d}_i + \dots + \mathbf{Z}_{iR} \boldsymbol{\nu}_i \quad (3.12)$$

For the inclusion of the interactive effect, $\mathbf{d}_i$, in (3.12), each row of the model matrix $\mathbf{W}_1$ is multiplied by the levels of d_i ¹⁷. This requires that we define $\text{diag}(d_i)$, a diagonal matrix with d_i as the i^{th} position on the diagonal, and 0 otherwise, multiplied to $\mathbf{W}_1$ as $\mathbf{W}_1 \text{diag}(\mathbf{d}_i)$. Thus

$$H(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_{iR} \boldsymbol{\nu}_i \quad (3.13)$$

with $\mathbf{X} = [\mathbf{X}^*: \mathbf{W}_1 \text{diag}(\mathbf{d}): \dots]$ and $\boldsymbol{\beta}' = [\boldsymbol{\vartheta}', \beta_1, \dots]$. Note, $\mathbf{Z}_{iR} \boldsymbol{\nu}_i$ from (3.11) remains unchanged completely and has the same definition as before.

To circumvent overfitting due to the basis dimensions associated with the unknown smooths, a penalty is employed for the departure from smoothness in the data fitting process, mitigating the weight given to the smoothing penalties. Adjusted for the centering parameterization results in $\beta_j' \tilde{\mathbf{S}}_j \mathbf{B}_j$ with $\tilde{\mathbf{S}}_j = \mathbf{P}' \ddot{\mathbf{S}} \mathbf{P}$ ¹⁸. Reforming the parameterized penalty in terms of the complete coefficient vector $\boldsymbol{\beta}$, then we have $\boldsymbol{\beta}' \mathbf{S}_j \boldsymbol{\beta}$ with $\mathbf{S} \equiv \tilde{\mathbf{S}}$ ¹⁹. Therefore, the model (3.13) is now

$$H(\boldsymbol{\mu}) = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z}_R \boldsymbol{\nu}_i + \lambda \boldsymbol{\beta}' \mathbf{S}_j \boldsymbol{\beta} \quad (3.14)$$

For the smoothing component, it is necessary to identify the fixed and random matrices of (3.14) via another re-parameterization. Therefore, the components of the smooth in of $\mathbf{X}$ can be split into β_f for fixed effects and $\boldsymbol{\nu}_r$ for random effects. For each smooth, it is beneficial to give a Bayesian interpretation to the approach by specifying a prior distribution on the parameter vector $\boldsymbol{\beta}$ associated with the degree of smoothness associated with $f(\cdot)$. That is, a prior for an arbitrary smooth function as

¹⁶The QR factorization requires an arbitrary matrix $\mathbf{A}$, it can be decomposed into a product $\mathbf{A} = \mathbf{Q}\mathbf{O}$ such that $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{O}$ is an upper triangular matrix

¹⁷In literature, this term is also deemed as a “varying coefficient” in which a coefficient in a standard generalized linear model is allowed to smoothly vary with an additional covariate

¹⁸Note, $\tilde{\mathbf{S}}$ is an arbitrary matrix of coefficients that may be a weighted sum of simpler matrices of known coefficients with estimable weights [9].

¹⁹A matrix of coefficients $\tilde{\mathbf{S}}$ is inundated with additional $\mathbf{0}$ such that $\boldsymbol{\beta}' \mathbf{S}_j \boldsymbol{\beta} = \boldsymbol{\beta}' \tilde{\mathbf{S}}_j \boldsymbol{\beta}$.

$$\mathbf{f}_\beta(\beta) \propto e^{\frac{-\lambda \beta' \mathbf{S} \beta}{2}}$$

However, the prior function itself is considered improper given $\mathbf{S}$ contains insufficient information in the data for model estimation (or has explicit rank deficiency). Therefore, to re-parameterize for the new parameters, a proper prior is needed for the random effect component as well as an improper prior for the fixed effects which generally can be achieved via the use of Singular Value Decomposition²⁰ (SVD) of the penalty matrix $\mathbf{S} = \mathbf{\Lambda}' \mathbf{K} \mathbf{\Lambda}$ ([9], [32]). The matrix $\mathbf{K}$ contain descending eigenvalues only on the principal diagonal where $\mathbf{\Lambda}$ are orthogonal matrices with the eigenvectors of $\mathbf{S}$. Submatrices from $\mathbf{K}$ are retrieved as

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_+ & \mathbf{0} \\ \mathbf{0} & \ddot{\mathbf{K}} \end{bmatrix} \quad (3.15)$$

where $\mathbf{K}_+$ is the smallest such submatrix containing strictly positive eigenvalues and $\ddot{\mathbf{K}}$ contains the remaining ordered eigenvalues. Thus, the new parameter or coefficient vector is separated into the fixed and random components as $(\mathbf{b}', \beta')' = \mathbf{\Lambda}' \beta$. Substituting this augmentation into the penalty term in (3.14) yields

$$\begin{aligned} \beta' \mathbf{S} \beta &= \beta' \mathbf{\Lambda} \mathbf{K} \mathbf{\Lambda}' \beta \\ &= (\mathbf{b}_r, \beta_f') \mathbf{K} (\mathbf{b}_r', \beta_f')' \\ &= (\mathbf{b}_r, \beta_f') [\mathbf{K}_+ : \ddot{\mathbf{K}}] (\mathbf{b}_r', \beta_f')' \\ &\equiv \mathbf{b}_r' \mathbf{K}_+ \mathbf{b}_r \end{aligned}$$

Given this result, it is clear that the strictly positive eigenvalues associated with $\mathbf{K}_+$ penalizes the random coefficients with no penalty applied to the fixed coefficients²¹. The random coefficients now has the distribution $\mathbf{b}_r \sim N(\mathbf{0}, \mathbf{K}_+^{-1}/\lambda)$. The resulting eigenvector matrix from the previous SVD on the matrix of known coefficients $\mathbf{S}$, $\mathbf{\Lambda}$, can be split such that $\mathbf{\Lambda} = [\mathbf{\Lambda}_r : \mathbf{\Lambda}_f]$, where we can define $\mathbf{X}_f = \mathbf{X}^f \mathbf{\Lambda}_f$ and $\mathbf{X}_r = \mathbf{X}^f \mathbf{\Lambda}_r$. The general linear mixed model representation for a smooth function f is

$$\mathbf{X}_f \beta_f + \mathbf{X}_r \mathbf{b}_r \quad (3.16)$$

However, for estimation utilizing the **mgcv** package, a further parameterization is needed such that $\mathbf{b} = \mathbf{b}_r \mathbf{K}_+^{1/2}$ and $\mathbf{Z}^\dagger = \mathbf{X}_r \mathbf{K}_+^{-1/2}$ ²² ([9]). The final resulting form of (3.16) leads to

²⁰The singular value decomposition of a matrix B is a factorization process of B into the product of three matrices, (e.g. $B = UDV'$) where the columns of U and V are orthonormal and the matrix D is diagonal with real, positive elements.

²¹This result also implies a new proper prior now for $\mathbf{f}_{\mathbf{b}_r}$ as: $\mathbf{f}_{\mathbf{b}_r}(\mathbf{b}_r) \propto e^{\frac{-\lambda \mathbf{b}_r' \mathbf{K}_+ \mathbf{b}_r}{2}}$.

²²Note, $\mathbf{Z}^\dagger$ is, now reparameterized, the random effects design matrix associated with the given functions associated with the smooths.

$$\mathbf{X}_f \boldsymbol{\beta}_f + \mathbf{Z}^\dagger \mathbf{b} \quad \mathbf{b} \sim N(0, \boldsymbol{\omega} \mathbf{I}) \quad (3.17)$$

with $(\mathbf{Z}^\dagger, \mathbf{b})$ defined above, and $\boldsymbol{\omega} = \lambda_j^{-1}$ ²³. The inclusion of (2.17) into a standard generalized linear mixed model (GLMM) requires appending the columns of $\mathbf{X}_f$ and $\mathbf{Z}^\dagger$ to the fixed and random model matrices, respectively, as well as specifying a chosen distribution for the random effects.

Therefore, in matrix notation, the GAMM with arbitrary basis functions for the smooths $f(\cdot)$ with a parametric factor smooth re-casted as a GLMM can be written as

$$H(E[\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\nu}]) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\nu} \quad (3.18)$$

where $\mathbf{X} = (\mathbf{X}^* : \mathbf{W}_1 \text{diag}(\mathbf{d}) : \dots \mathbf{X}_f :)$, $\boldsymbol{\beta} = (\boldsymbol{\vartheta}', \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_f)'$, $\mathbf{Z} = (\mathbf{Z}^\dagger, \text{blockdiag}(\mathbf{Z}_R))$, $\boldsymbol{\nu} = (\mathbf{b}', \boldsymbol{\nu}'_i)'$, and $\boldsymbol{\nu} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}, \boldsymbol{\varrho}})$ where $\boldsymbol{\Sigma}_{\boldsymbol{\omega}, \boldsymbol{\varrho}} = \text{Cov}(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{\omega} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}(\boldsymbol{\varrho}) \end{bmatrix}$.

From above, the model form contains:

1. $\mathbf{X}$ - The full fixed effects design matrix including: the fixed effects associated with all parametric terms included in the model, the multiplied diagonal matrix of the factor $\mathbf{d}$ to the model matrix of the smoothed covariate $\mathbf{w}_{1i}$, and the (generic) remaining basis matrices for all remaining smoothing functions (or the column vectors of the unpenalized portion of each smooth function);
2. $\boldsymbol{\beta}$ - All the coefficients for the parametric portion of the GLMM as well as the unpenalized coefficients from each smooth.
3. $\mathbf{Z}$ - The entire random effects design matrices including the standard design matrix of a GLMM and the random effects design matrix associated with each smooth.
4. $\boldsymbol{\nu}$ - All random effect coefficients ($\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\omega} \mathbf{I})$, $\boldsymbol{\nu} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}, \boldsymbol{\varrho}})$).

A thin plate regression spline is a trivial to define the terms in (3.18) as with any other basis. We now turn our attention to estimation via the penalized quasi-likelihood approach.

3.4 Penalized Quasi-Likelihood (PQL) Estimation for Re-Casted GAMM

We briefly highlight the penalized likelihood approach ([34], [35], [36], [37]) for estimation of parametric effects (regression coefficients) and variance components (which includes the estimation of λ as a mixed model component) for this overparameterized GLMM. As the penalized quasi-likelihood approach has been well-evaluated in traditional GLMMs ([38], [27], [39]), there has not been a thorough evaluation of this

²³The vector of $\boldsymbol{\omega}$ contains all inverses of the smoothing parameters for each smooth, $\boldsymbol{\lambda}^{-1}$ as: $(\lambda_1^{-1}, \dots)$.

procedure for estimating the variance components for performance of a GAMM particularly for approximate inference for longitudinal count responses with a complex covariance structure. This is a popular technique for estimation as it does not need the full likelihood to be known; yet, the technique can be employed when there is complete knowledge about the suspected relationship between the means and variances of a response of interest²⁴. Another useful benefit of the PQL approach is its implementation ease by fitting linear mixed model to an adjusted response vector ([39]). As the variance components (and smoothing parameter subsumed within the vector of variance component) are mainly of interest, a penalty is imposed in order to evade of arbitrary values attributed to the random effect and mandate that the random effects approach zero ([40], [41]). This method implements a modified Laplace approximation of the marginal log-likelihood (resulting in a penalized marginal log-likelihood) wherein the random effects are biased towards zero in the approximation (where the “penalization” comes from in the PQL acronym). This method is advantageous as it is a pliant fitting process and computationally efficient. Also, advantageously, is that there exists a level of robustness against misspecified models as the exact distribution of interest is not required (as with maximum likelihood). Additionally, PQL estimators are certainly beneficial for accommodating complex, hierarchical covariance structures. This estimation procedure has clear disadvantages, however, as standard likelihood tests cannot be performed as its based on a quasi-likelihood and inference is only approximate. Additionally, as proved by [42], the PQL estimation procedure produces fixed parameter estimates that can be inconsistent due to Laplace approximation’s induced bias of the likelihood function, particularly for binary outcomes. However, as shown by [38], second-order (higher order) approximations biases may be reduced to have practical utility from an applied researcher’s perspective. Nevertheless, we proceed with this estimation technique. The parameters can be estimated by maximizing the integrated quasi-likelihood function by removing the random effects via integration as

$$L(\beta, \Sigma_{\omega, \boldsymbol{\varrho}}) \propto \frac{1}{|\Sigma_{\omega, \boldsymbol{\varrho}}|^{\frac{1}{2}}} \int \exp \left[-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{n_i} D_{ij}(y_{ij}, \mu_{ij}^{\nu_i}) - \frac{1}{2} (\boldsymbol{\nu}_i' \Sigma_{\omega, \boldsymbol{\varrho}}^{-1} \boldsymbol{\nu}_i) \right] d\boldsymbol{\nu}_i \quad (3.19)$$

with $D_{ij}(y, \mu) = -2\delta_{ij} \int_y^{\mu^{\nu_i}} \frac{y-u}{\phi v(u)} du$ is a quasi-deviance fit measure²⁵.

As there is no closed form solution to (3.19), the integral associated with (3.19) must be approximated (or numerically evaluated) which most conveniently is done by deriving a penalized quasi-likelihood via a Laplace approximation as

$$\text{QL}_{\text{pen}}(\beta, \Sigma_{\omega, \boldsymbol{\varrho}}) = - \left[\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^{n_i} D_{ij}(y_{ij}, \mu_{ij}^{\nu_i}) + \frac{1}{2} (\boldsymbol{\nu}_i' \Sigma_{\omega, \boldsymbol{\varrho}}^{-1} \boldsymbol{\nu}_i) \right] \quad (3.20)$$

²⁴Therefore the mean-variance relationship for the conditional density need to be known in order to employ this technique.

²⁵Note $\delta_{ij} = 1$ (a prior weight) which is applicable for the assumed distribution and of the response in our application. Additionally, this term can also be considered the conditional log-quasilikelihood of β given $\boldsymbol{\nu}_i$ [38]. Also, if, conditional on $\boldsymbol{\nu}$, the observations are drawn from a linear exponential family, then the quasi deviance is essentially the (scaled) difference $2(\ell(y; y, \phi) - \ell(y; \mu, \phi))$.

We highlight the estimation equations for the parameters and variance components. The PQL approach for estimation involves estimating a generalized linear mixed model by writing linear Taylor expansion for the conditional mean evaluated at current estimates of the fixed and random effects $\hat{\beta}$ and $\hat{\nu}$.

Let $\mathbf{Y}_i^* = (y_{i1}^*, \dots, y_{ij}^*)$ be the contemporary or surrogate response vector defined as:

$$y_{ij}^* = H'(\hat{\mu}_{ij}^\nu)(y_{ij} - \hat{\mu}_{ij}^\nu) + H(\hat{\mu}_{ij}^\nu)$$

for $j = 1, \dots, n_i$.

Then, the distribution of the working vector (in complete matrix notation) can be described as a traditional linear mixed model as

$$\begin{aligned} \mathbf{Y}^* &= \mathbf{X}\beta + \mathbf{Z}\nu + \epsilon \\ &\equiv \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\nu} + \hat{\Delta}^{-1}(\mathbf{Y} - H^{-1}(\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\nu})) \end{aligned} \quad (3.21)$$

with $\hat{\Delta} \equiv \frac{\partial H^{-1}(\eta)}{\partial \eta}$ as a diagonal matrix defined at the parameter estimates $(\hat{\beta}, \hat{\nu})$. Given the covariance specification of the surrogate/response vector, the unconditional variance of the error of the traditional linear mixed model is given as

$$\text{Var}(\epsilon) = \sigma^2 \hat{\Delta}^{-1} \sqrt{\mathbf{C}} \mathbf{R}(\rho) \sqrt{\mathbf{C}} \hat{\Delta}^{-1}$$

This $\mathbf{R}(\rho)$ matrix (the ‘‘R-side’’ covariance structure) is the variance-covariance matrix of the surrogate responses that contains elements on the off-diagonal that can be modeled with temporal (or more advanced) correlation models, conditional on the random effects. The parameter vector ρ contain the parameters for any autocorrelation model. Additionally, the marginal variance of the surrogate response vector is given as

$$\mathbf{V} = \mathbf{V}(\theta) = \sigma^2 \hat{\Delta}^{-1} \sqrt{\mathbf{C}} \mathbf{R}(\rho) \sqrt{\mathbf{C}} \hat{\Delta}^{-1} + \mathbf{Z} \Sigma_{\omega, \theta} \mathbf{Z} \quad (3.22)$$

where $\theta = (\phi, \omega', \rho', \sigma^2)'$.

Laplace’s approximation of (3.19) is utilized due to the likely intractability due to potential high dimensional integration. This is accomplished by employing a quadratic augmentation of the exponent in the integrand of (3.19) in the neighborhood of the mode and approximating $D_{ij}(\cdot)$ via the Pearson chi-squared test statistic. Once done, the approximate marginal log-likelihood is

$$\ell(\mathbf{y}^*; \theta) \approx -\frac{1}{2}(\mathbf{y}^* - \mathbf{X}\hat{\beta})\mathbf{V}(\theta)^{-1}(\mathbf{y}^* - \mathbf{X}\hat{\beta}) - \frac{1}{2}\log|\mathbf{V}(\theta)| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{X}| \quad (3.23)$$

which directly corresponds to the REML version of the surrogate vector with θ as the parameter vector of variance components and ν as a random effect ([27], [34]). Minimizing (3.23) with respect to θ results in

$\hat{\boldsymbol{\theta}}$, or the REML estimates for the variance components which are typically unbiased relative to the analog estimates computed by maximum likelihood (ML) estimates due to the unavailability of degrees of freedom from estimating the mean function. The lack of availability of degrees of freedom in the ML case results in underestimation of the covariance matrix parameters ([43]).

Once computing the REML estimates of the variance components, $\hat{\boldsymbol{\theta}}$, from the minimization of (3.23), the estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are formed from

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{y}^* \quad (3.24)$$

$$\hat{\boldsymbol{\nu}} = \boldsymbol{\Sigma}_{\omega, \hat{\boldsymbol{\theta}}} \mathbf{Z}'\mathbf{V}(\hat{\boldsymbol{\theta}})^{-1}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (3.25)$$

The outer-iteration algorithm proposed by [35] is as follows: Given initial values of $(\hat{\boldsymbol{\beta}}^{[0]}, \hat{\boldsymbol{\nu}}^{[0]})$ and our recasted GAMM in (3.18), with a specified link function $H(\cdot)$, the variance functions $c(\mu)$ of the model from the diagonalized matrix $\mathbf{C}$, the between-subject covariance structure $\boldsymbol{\Sigma}_{\omega, \boldsymbol{\theta}}$, and the “residual” covariance structure for an autocorrelation (temporal) model $\mathbf{R}(\boldsymbol{\rho})$. The parameter estimates are iterated to convergence as follows

1. Create the surrogate response $\mathbf{y}^{*[k+1]}$ from (3.21), given the current k^{th} iterations $(\tilde{\boldsymbol{\beta}}^{[k]}, \tilde{\boldsymbol{\nu}}^{[k]})$.
2. Estimate $\hat{\boldsymbol{\theta}}^{[k+1]}$ using $\mathbf{y}^{*[k+1]}$ from (3.23) (minimizing the REML log-likelihood for $\hat{\boldsymbol{\theta}}$).
3. Estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\nu}})$ employing the forms from (3.24, 3.25). Afterwards, the parameters are updated setting $\tilde{\boldsymbol{\beta}}^{[k+1]} = \hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\nu}}^{[k+1]} = \hat{\boldsymbol{\nu}}$.
4. Set $k = k + 1$ and repeat steps until convergence.

This algorithm essentially describes the estimation procedure from the **mgcv** package for estimating the GAMMs utilizing the function `gammPQL()`, which is a slightly approximated version of the `glmmPQL()` function from the **MASS** package ([44], [45]).

Regarding inference of smoothing components, specifically the computational aspect of creating confidence intervals for the smoothing terms, we follow ([7], [9], [33]). Let $\underline{\boldsymbol{\beta}}$ contain the fixed and random effects for all smoothed terms. Defining $\underline{\mathbf{X}}$ represent the associated design matrix associated with the smoothed components as well as $\underline{\mathbf{Z}}$ be the random effects design matrix excluding the columns associated with the smooths, a covariance matrix can be constructed as

$$\mathbf{V} = \underline{\mathbf{Z}}' \underline{\boldsymbol{\Sigma}}_{\omega, \boldsymbol{\theta}} \underline{\mathbf{Z}} + \sigma^2 \hat{\boldsymbol{\Delta}}^{-1} \sqrt{\mathbf{C}} \mathbf{R}(\boldsymbol{\rho}) \sqrt{\mathbf{C}} \hat{\boldsymbol{\Delta}}^{-1}$$

where $\underline{\boldsymbol{\Sigma}}_{\omega, \boldsymbol{\theta}}$ is the covariance matrix associated with the random effects associated with $\underline{\mathbf{Z}}$. Then,

$$\underline{\boldsymbol{\beta}} \sim N(\tilde{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \mathbf{S})^{-1})$$

with $\mathbf{S} = \sigma^{-2} \sum \lambda_i \mathbf{S}_i$. Moreover, the effective degrees of freedom (EDF) per smoothing term element in $\boldsymbol{\beta}$ can be computed from the leading diagonal of

$$\mathbf{E} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X} + \mathbf{S})^{-1} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})$$

3.4.1 Temporal Serial Correlation Structure

As discussed in the previous section, there may be a need to attribute a specified covariance structure for the residual matrix. We highlight two popular temporal structures ([46]).

1. Diagonal - In the case of a diagonal structure, we assume equal variance and the residuals associated with the same subject are assumed to be uncorrelated.

$$\mathbf{R}_i = \text{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

2. AR(1) - First order, autoregressive structure with homogeneous variances. In this case, repeated measures can exhibit serial correlation where this temporal structure defines within-subject correlated observations at adjacent times. Observations further in time have correlations that exponentially decrease.

$$\mathbf{R}_i = \text{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} \sigma^2 & \sigma^2 \rho & \sigma^2 \rho^2 & \sigma^2 \rho^3 \\ \sigma^2 \rho & \sigma^2 & \sigma^2 \rho & \sigma^2 \rho^2 \\ \sigma^2 \rho^2 & \sigma^2 \rho & \sigma^2 & \sigma^2 \rho \\ \sigma^2 \rho^3 & \sigma^2 \rho^2 & \sigma^2 \rho & \sigma^2 \end{bmatrix}$$

Given the setup for the GAMM model and the likely temporal correlation structure given the random effect, we now discuss the empirical implications of estimation when having such a complex covariance structure. We explore the longitudinal counts that may be temporally correlated for a subject in cross-section i conditional upon the random effect. In the next section, we highlight two simulation designs that highlights joint parameter performance for the estimation of GAMM models in terms of: a specified fixed effects parameter, all model variance parameters, and the use of thin plate regression splines to estimate smooth functions associated with levels of a parametric factor covariate.

3.5 Conditionally Correlated AR(1) Models for Stationary and Nonstationary Count Data with random effects - Simulation Designs

With suspected complex covariance structures, much care should be taken into account when interest lies on modeling the response as a nonlinear function of covariates of interest. Failure to properly account for variance components that exist in longitudinal structures distorts inference on the parametric components of the predictor in a GAMM may even contribute to convergence issues. As these issues are well studied in traditional GLMM settings, we take particular interest in a variety of variance component parameter settings as well as fixed effects in the presence of factor smooths utilizing GAMMs for conditionally serially correlated counts for estimation. This leads to a variety of practical questions:

- With the inclusion of factor smooths as part of the linear predictor, to what extent can GAMM models effectively estimate variance components when they truly exist in a longitudinal setting?
- To what extent is utilizing GAMM models effective where misspecification occurs in extreme cases of variance components where either the simple random effect's variance is zero or the temporal autocorrelation parameter is zero given the random effect?
- To what extent are any fixed effects properly estimated in the joint estimation of the random effect, autocorrelation conditional upon the random effect, unknown smooth functions, and fixed effects in both properly specified and misspecified cases?
- Are these GAMM models' joint estimation of variance components, parametric regression coefficients, and unknown smooths functions strongly affected in either case when the panels are balanced versus unbalanced?
- How reliable is the penalized quasi-likelihood approach for estimation?

We seek to answer these questions by evaluating the bias in finite samples of the resulting PQL estimator of the variance components of interest in these two simulation designs for conditionally correlated AR(1) models for longitudinal counts. In understanding the general asymptotic bias of variance components of interest, following [39], we note two regularity conditions that must hold: (I.) The design matrix corresponding to $\mathbf{X}$ must contain an intercept, and (II.) the standard random effects design matrices²⁶ only contains zeros and ones. In our case, however, a caveat must be extended. The standard random effects design matrices that we refer to here that only contains zeros and ones refer to the blockdiag($\mathbf{Z}_R$) in $\mathbf{Z}$ as we still must include the design matrix associated with each smooth. Also, of note, as we are interested in the asymptotic bias of

²⁶Note, the standard random effects design matrices in the simulation designs that follow will actually be omitted in the data generating processes that follow.

the variance components in the following simulation exercises, we need not necessarily use REML to adjust for the loss of degrees of freedom from estimating the strictly parametric covariates (e.g. the maximum likelihood estimating equations could be use), but we continue with the REML version regardless²⁷.

In order to assess a generalized additive mixed models with a temporal AR(1) correlation structure conditional on a random effect in longitudinal settings, we perform a simulation study under a range of models and data. As count responses in longitudinal studies are our primary focus, we generate cluster-grouped data while altering the number of subjects (N), the number of repeated observations associated with each subject (n_i), parameters ($\{\sigma_\nu^2, \rho\}$), and the smoothing function f . Regarding the model parameters, σ_ν^2 is the variance of the random effects, and ρ is the parameter of serial correlation for a temporal, first-order autoregressive process conditional upon the random effect. Note, the β_1 coefficient of the GAMM remains unchanged throughout the simulations in all cases. The parameters are summarized in Table 3.1.

The GAMM model with a parametric factor smooth considered for this analysis is represented as

$$\log(E[y_{ij}|\mathbf{x}_{ij}, \nu_i]) = \beta_1 x_{1ij} + f_k(x_{2ij})x_{1ij} + \nu_i^{28} \quad (3.26)$$

where $\nu_i \sim N(0, \sigma_\nu^2)$, $k = \{1, 2\}$ corresponds to test functions considered in the semi-parametric models, x_{1ij} corresponds to the non-baseline for the binary, parametric component of model, β_1 is the fixed effect coefficient of interest associated with the non-baseline covariate, ν_i represent the random effects, and $\log(\cdot)$ is the link function in all designs and cases considered. As our interests lie in analyzing and/or capturing variance components' estimates for correlated counts in a longitudinal format given factor smooths, we follow the data generating processes (DGPs), outlined by [47], given considerations for time-independent and time-dependent covariates. When all covariates utilized in the model are time-independent, then it possible to construct correlated, Poisson counts that have a stationary correlation structure in which the repeated responses may still be correlated across time, conditional on the random effect. This represents a conditionally serially correlated stationary model for repeated counts. In the case in which the model includes any time-dependent (time-varying) covariates, a conditionally serially correlated non-stationary data generating process is considered. In this latter case, a stationary correlation structure may be inappropriate, as it can degrade the temporal correlation structure to the point of being unidentified, and a non-stationary correlation structure is warranted.

Table 3.1: Parameters of Simulation Designs

Parameters	Values of Parameters
N (Number of Subjects)	(200, 300)
n_i (Number of repeated Observations of subject i)	(25-40, 40)
σ_ν^2 (Random Effect Variance)	(0, 0.25, 0.5)
ρ (First-Order, Autoregressive Coefficient)	(0, 0.25, 0.5)

²⁷We will use the REML estimating equations in our applications.

²⁸Design matrix associated with the random effect is dropped for notational convenience.

These two conditions provide the basis of our two simulation designs considered. In both cases, the repeated counts have an additional temporal correlation structure, conditional on the individual random effect. The stationary and non-stationary conditional AR(1) correlation structure for correlated Poisson counts form Simulation Design I. and Simulation Design II, respectively. Explicitly, also, the varying of $n_i = 25 - 40$ vs. $n_i = 40$ in Table 3.1 represents the unbalanced and balanced cases of the longitudinal design. In the former case (unbalanced), $n_i = 25 - 40$ are generated randomly such that each subject within the cross-section can have at least 25 observations, yet at most 40 observations.

Following [47], we first outline Simulation Design I. The stationary, AR(1) Poisson DGP is our first structure of interest.

Let $y_{i1}|\nu_i \sim \text{Pois}(\check{\mu}_i)$ where $\check{\mu}_i|\nu_i = \exp(\beta_1\check{x}_{1i} + f_k(\check{x}_{2i})\check{x}_{1i} + \nu_i)$. The remaining longitudinal Poisson counts for y_{ij} , $j \geq 2$, is constructed as:

$$y_{ij}|\nu_i = \rho \circ (y_{i(j-1)})|\nu_i + q_{ij}|\nu_i \quad j = 2, \dots, n_i,^{29} \quad (3.27)$$

where $q_{ij}|\nu_i \sim \text{Pois}(\check{\mu}_i(1 - \rho))$. Regarding the covariate associated with the model (3.26) for our first simulation design, we only consider time-independent covariates, where we simulate $x_{1ij} \equiv \check{x}_{1i} \sim \text{Bern}(p)$ with $p = 0.6$. Thus, $x_{1ij} \equiv \check{x}_{1i}$ represents a binary 0,1 factor variable that is time-independent generated within-cluster. The factor structure is needed in order to assess the parametric factor smooths. For the smoothed covariate in (3.26), we construct $x_{2ij} \equiv \check{x}_{2i} = i/N$, a deterministic sequence, restricting this variable to be a continuous covariate that does not change over the time point j for any cross-sectional unit i .

We now consider the Simulation Design II. which can include time-dependent covariates (non-stationary structure). The AR(1) Poisson DGP for the non-stationary case follows similarly to Design I. as:

$$y_{ij}|\nu_i = \rho \circ (y_{i(j-1)})|\nu_i + q_{ij}|\nu_i \quad j = 2, \dots, n_i^{30} \quad (3.28)$$

where $q_{ij}|\nu_i \sim \text{Pois}(\mu_{ij} - \rho\mu_{i(j-1)})$ where $y_{i1}|\nu_i \sim \text{Pois}(\mu_{i1})$, as before. The conditional mean is defined as $\mu_{ij}|\nu_i = \exp(\beta_1\check{x}_{1i} + f_k(x_{2ij})\check{x}_{1i} + \nu_i)$, with $k = \{1, 2\}$. In this case, x_{2ij} is a continuous covariate that increments by $1*0.001$ across time points in the cross-sectional unit i (e.g. $1 \leq j \leq n_i$, time-varying). We will report the means, variances, and correlations corresponding to our DGPs in the appendix.

²⁹The term $\rho \circ (y_{i(j-1)})$ denotes a “binomial-thinning” parameter. For simulation in my setup, I simulate Bernoulli random variables from the “rbinom” function from the **stats** package that takes the appropriate number of observations to simulate, the size based on the lagged response, and the probability of success which equals ρ . More details can be found in the appendix as well as in [48], [49], and [47]. Also, ρ is constrained to lie between $0 \leq \rho \leq 1$.

³⁰Here, ρ is constrained to satisfy $0 \leq \rho \leq \min\left[\frac{\mu_{ij}}{\mu_{i(j-1)}}, 1\right]$, following [49].

Regarding the smoothing functions, $f(\check{x}_{2i})$ or $f(x_{2ij})$ depending on the simulation designs, two functions for testing are considered for estimating the smooth effect:

1. Polynomial: $f_1(x) = 10(10x)^3(1-x)^{10} + 0.2x^{11}(10(1-x))^6$

2. Linear: $f_2(x) = -3x$

where $x \equiv \check{x}_{2i}$ or $x \equiv x_{2ij}$. For both designs and in all cases, the total number of replications performed are $R = 500$.

3.5.1 Performance of Parameters

The principal goal of this simulation study is to assess how well these semi-parametric GAMMs jointly estimate the parameters of interest including the fixed effect parametric coefficient β_1 , the variance of the random effect σ_ν^2 , as well as the first-order, conditional autoregressive component ρ . Additionally, we will assess how well the estimated function $\hat{f}()$ approximates the true function $f()$.

We present a few performance indicators. For all parameter estimates of interest, $\{\hat{\beta}_1, \hat{\sigma}_\nu^2, \hat{\rho}\}$, we will compute the difference between a specified true value and the mean of the corresponding estimate from the replications as well as the average squared difference between the estimated values and the actual value across all replications. These two quantities are also known as the bias and mean squared error (MSE), respectively. That is, for any parameter estimate computed (e.g. for any $\hat{\theta} = \{\hat{\beta}_1, \hat{\sigma}_\nu^2, \hat{\rho}\}$) as

1. Bias($\hat{\theta}$) = $E(\hat{\theta}) - \theta$

2. MSE($\hat{\theta}$) = $(\text{Bias}(\hat{\theta}))^2 + \text{Var}(\hat{\theta})$

Additionally, we compute the average mean squared error for each level of the factor smooth, which is computed as the average distance error over the grand total of replications, R ,

$$\text{MSE}_f = \frac{\sum_{i=1}^N \left(f_k(x_2)\check{x}_{1i} - \hat{f}_k(x_2)\check{x}_{1i} \right)^2}{\sum_{i=1}^N n_i} \quad \text{AMSE}_f = \frac{\text{MSE}_f}{R}$$

where $x_2 \equiv \check{x}_{2i}$,³¹ with $k = \{1, 2\}$. We simulate datasets under varying parameters, and fit each dataset with a generalized additive mixed model with a random intercept (assumed Gaussian) and first-order, temporal autocorrelation AR(1) structure in all cases. For the datasets we simulate, we also test and observe effects of model “misspecification” on the zero values of the random effects’ variance and the autoregressive coefficient (either $\sigma_\nu^2 = 0$ or $\rho = 0$). From a practical standpoint, $\sigma_\nu^2 = 0$ signifies that

³¹When $x_2 = x_{2ij}$, $\text{MSE}_f = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} \left(f_k(x_2)\check{x}_{1i} - \hat{f}_k(x_2)\check{x}_{1i} \right)^2}{\sum_{i=1}^N n_i}$.

subject-level clustering does not aid in explaining any variance of the counts/response (counts from subjects are completely independent regardless if they come from if observed from the same subject or not), and $\rho = 0$ signifies (conditional) temporal independence of within-subject observations. All models' components and fixed effect coefficients will be estimated using penalized quasi likelihood from the **mgcv** package utilizing the **gamm()** function. For estimating the true curve (either polynomial or linear) in each GAMM, we will utilize only thin plate regression splines fitted as **s(...bs='tp')**. Also, we use the default basis in all scenarios for the thin plate regression spline with $d = 1$ since we have one covariate in the smooth function, and we penalize the second degree derivative $m = 2$.

To simulate the datasets with varying parameters for modeling the Poisson repeated responses with a stationary and non-stationary covariance structures, I write a function in R, that takes, as inputs: $(N, n_i, \beta, \sigma_v^2, \rho)$. For the GAMM fitting process, the underlying fitting engine for **gamm()** utilizes **lme**. Given the randomness associated with simulating the covariates, the Poisson distributed correlated responses with stationary/non-stationary structures, and the attempt at modeling the explicit correlation via an AR(1) models with these fairly large datasets, **lme** can be a bit unstable in terms of convergence particularly when modeling non-independence in the data for the generalized case due to the inability to separate heterogeneity from trend ([9]). In addition to increasing the number of iterations for the GAMM fitted by PQL to **niterPQL=250**, an additional error catching process is formulated. For simulating and fitting R datasets of n subjects repeatedly observed n_i times, the fitted GAMM calls are first wrapped in an error handling process that is flagged if any convergence error is obtained. Results are appended until 500 instances of model fits are obtained. Additionally, each GAMM model was fitted with the default basis of size **k=10** for the thin plate regression spline in both designs for all cases.

3.5.2 Simulation Results: Design I.

For this design, we find that the results are relatively consistent regardless of whether the subjects were observed equally $n_i = 40$ versus unequally (e.g. a random sample with replacement assigned to each cluster) $n_i = 25 - 40$. The unequal or unbalanced case will be reported in the appendix at the end of the chapter. For this design, we report the results for a cluster size of $N = 300$ and a balanced case with $n_i = 40$.

Table 3.2: Simulation Results I: Balanced Stationary Panel

Case 1:					Case 2:					Case 3:				
Parameter Values	Polynomial: $f_1(\bar{x}_{2i})\bar{x}_{1i}$		Linear: $f_2(\bar{x}_{2i})\bar{x}_{1i}$		Parameter Values	Polynomial: $f_1(\bar{x}_{2i})\bar{x}_{1i}$		Linear: $f_2(\bar{x}_{2i})\bar{x}_{1i}$		Parameter Values	Polynomial: $f_1(\bar{x}_{2i})\bar{x}_{1i}$		Linear: $f_2(\bar{x}_{2i})\bar{x}_{1i}$	
	Bias	MSE	Bias	MSE		Bias	MSE	Bias	MSE		Bias	MSE	Bias	MSE
$\beta_1 = 1$	0.01515	0.006632	0.020869	0.014604	$\beta_1 = 1$	0.014560	0.00084	0.0813	0.00025	$\beta_1 = 1$	0.04099	0.035653	-0.030052	0.012213
$\sigma_v^2 = 0.50$	0.010131	0.00080	-0.07225	0.008919	$\sigma_v^2 = 0$	0.01340	0.000185	0.00002	0.00001	$\sigma_v^2 = 0.50$	0.11289	0.099306	-0.03042	0.004542
$\rho = 0.50$	0.018121	0.000367	0.036476	0.001444	$\rho = 0.50$	0.021909	0.000507	-0.04387	0.0020	$\rho = 0$	-0.025952	0.000680	0.003096	0.000541

Table 3.3: Stationary Estimated Smooth Functions - Performance (Case 1 - Stationary)

	Factor Smooth: Level (0), (x2)x1:(0)	Factor Smooth - Level (1), (x2)x1:(1)
	Average Mean Squared Error (AMSE _f)	Average Mean Squared Error (AMSE _f)
Polynomial	0.029144	0.02348
Linear	0.01548	0.01230

The following results for the three primary cases for the first design are shown in Table 3.2. In Case 1 (the correctly specified case), the results are shown for conditionally serially correlated Poisson longitudinal counts with temporal correlation. The model was fit to 500 datasets with moderate random effect's variance ($\sigma_\nu^2 = 0.50$), moderate temporal autocorrelation ($\rho = 0.50$), and the parametric coefficient on the non-baseline covariate $\beta_1 = 1$. For the case where we fit both the polynomial and linear functions for $\check{x}_2$ factor smooths, we note that both the bias and MSEs for the coefficient, β_1 , for $\check{x}_{i1}$ (parametric GAMM component) are relatively negligible. Also, for the polynomial case, both the random effect's variance estimate and the autoregressive coefficient show negligible bias. Similarly, the MSEs for both parameter estimates are slight as well. In estimating the linear function with a factor smooth in Case 1, the bias for the estimated random effect's variance and the temporal AR(1) coefficient are clearly more biased (negatively and positively, respectively). The MSEs for the linear case are negligible for the estimated variance and the autoregressive coefficient.

In the misspecified cases, we highlight Case 2 ($\sigma_\nu^2 = 0$) and Case 3 ($\rho = 0$). For Case 2, in which we estimated the smooth function factor smooth $f()$ with a polynomial, the biases for the covariance parameters and dummy parametric parameter are relatively small, while the autoregressive parameter estimate in the linear function factor smooth is moderately downwardly biased. Also, Case 2, the parametric coefficient for β_1 is higher as well. For Case 3 in which the 500 datasets were simulated with $\rho = 0$, the random effect's variance in the polynomial is moderately biased and has a relatively large MSE for the polynomial. In the case of the linear factor smooth for this misspecified case, the bias and MSEs are manageable.

Table 3.3 highlights average mean squared errors for the $R = 500$ replications of the factor smooth of the binary factor ($\check{x}_1$) by the time-independent "trend" covariate $\check{x}_2$. In the case of both levels of the polynomial fit of $f_1(\check{x}_{2i})\check{x}_{1i}$, the AMSEs are relatively low, yet are slightly lower in the linear ($f_2(\check{x}_{2i})\check{x}_{1i}$) case. This seems to suggest, in the correctly specified balanced case (Case 1), the thin plate regression spline fits the linear true curve slightly better, but impressively, still fits the polynomial quite well for both smooths across the two levels of the factor. Lastly, we show the average factor smooth fits superimposed upon the true curve $f()$ for the 500 replications for the correctly specified Case 1. The x-axis represents $\check{x}_{2i}$, the time-independent covariate, and the black points represents the (mean) values of $f(x_{2i})$. As can be seen in Figure 3.1, the mean estimates corresponding to each factor level both the polynomial and linear in the 500 replications approximate the true curve quite well

3.5.3 Simulation Results: Design II

We now highlight the simulation results from including a time-dependent covariate in the nonstationary case. We report the correctly specified case and two misspecified cases for $N = 200$ (a decrease in the cluster size relative to the first design) and $n_i = 40$. As in the previous simulation design, these results are consistent with unbalanced scenario with $n_i = 25 - 40$ (shown in the Appendix). For this design, we reduce both the values of the parameters for $\{\sigma_\nu^2, \rho\}$ to 0.25. As can be seen in Table 3.4, we see similar results to the first design.

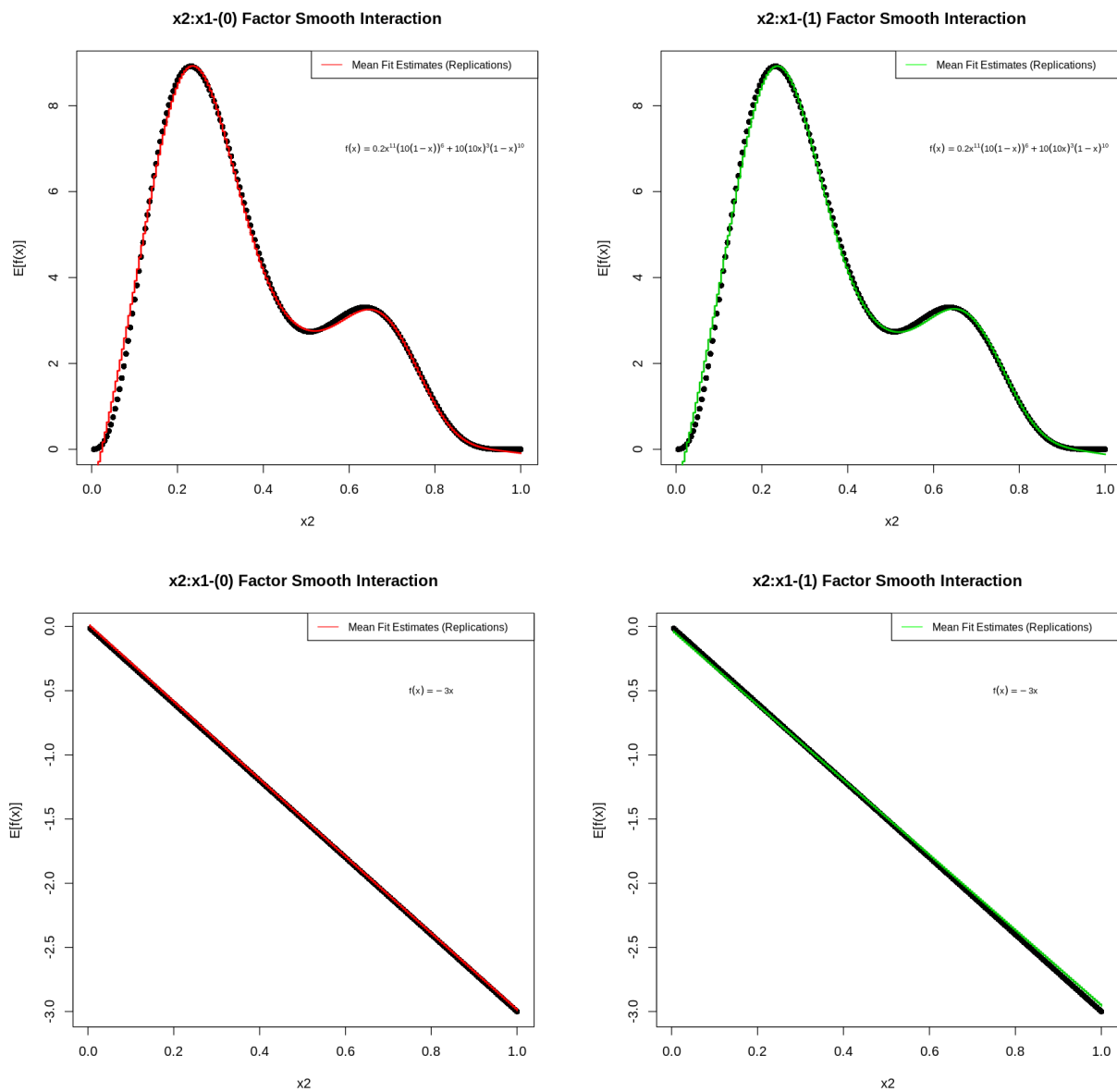


Figure 3.1: Factor Smooth Plots of Mean Estimates (Case 1)

In the correctly specified case, the biases and MSEs for both the polynomial and linear cases with the time-varying trend factor smooth are relatively negligible although a little high for the bias regarding ρ and β_1 in the linear case. In Case 2, the bias for the parametric component is a little high as well is the case for ρ in the polynomial case. Lastly, in Case 3, we note that random effect's variance is quite underestimated via the bias in the polynomial case as in the previous simulation design (although not as biased). The average mean squared errors reported for each level of the factor smooths is also comparable to the previous simulation design as seen in Table 3.5. In this case, however, the polynomial factor smooths outperforms the linear cases.

The goal of these simulation designs was to empirically investigate parameter performance of variance components, a fixed coefficient, and associated factor smooths associated with unknown functions for conditionally serially correlated counts using generalized additive mixed models. The extent to which we could simultaneously estimate the variance components of interest (the temporal AR(1) correlation coefficient conditional upon a simple random effect) as well as estimating unknown smooth functions with consideration of factor smooths are promising given our fixed effect under balanced and unbalanced longitudinal designs. Particularly in the correctly specified cases, given consideration for both stationary and non-stationary Poisson counts in both balanced and unbalanced panel structures, the results seem to mitigate the biases and MSEs for the polynomial cases quite well for the parametric parameter and the variance components.

Additionally, the average MSEs associated with the polynomial vs. linear functions associated with the factor smooths in both the stationary and non-stationary correlation structure for each level of the parametric factor are relatively similar, imbuing us with confidence in that the low-rank, thin plate regression spline utilized in the GAMM setup can properly estimate linearity vs. nonlinearity relatively equally (in terms of error) in balanced scenarios for Case 1. Similar results are seen in the unbalanced case for a non-stationary correlation structure in the appendix for Case 1. In the misspecified balanced cases, specifically when we generate datasets with the autoregressive coefficient set to 0, we note a higher relative biases in the random effect of the variance specifically in the polynomial case. For the unbalanced case seen when the data sets are generated with the autoregressive coefficient set to zero seen in the appendix of this chapter in the stationary case, there seems to be a higher bias

Table 3.4: Simulation Results II: Balanced Non-Stationary Panel

Case 1:					Case 2:					Case 3:				
Parameter	Polynomial: $f_1(x_{2ij})\tilde{x}_{1i}$		Linear: $f_2(x_{2ij})\tilde{x}_{1i}$		Parameter	Polynomial: $f_1(x_{2ij})\tilde{x}_{1i}$		Linear: $f_2(x_{2ij})\tilde{x}_{1i}$		Parameter	Polynomial: $f_1(x_{2ij})\tilde{x}_{1i}$		Linear: $f_2(x_{2ij})\tilde{x}_{1i}$	
Values	Bias	MSE	Bias	MSE	Values	Bias	MSE	Bias	MSE	Values	Bias	MSE	Bias	MSE
$\beta_1 = 1$	0.01327	0.00180	0.0608	0.0152	$\beta_1 = 1$	0.03764	0.00628	0.04569	0.0063	$\beta_1 = 1$	-0.01971	0.0086	-0.01591	0.0036
$\sigma_v^2 = 0.25$	0.00758	0.00108	0.0178	0.0007	$\sigma_v^2 = 0$	2.7e-10	0.00001	3.1e-10	1e-19	$\sigma_v^2 = 0.25$	-0.1039	0.0007	0.0044	0.0001
$\rho = 0.25$	0.0429	0.0020	0.0363	0.0014	$\rho = 0.25$	0.0430	0.0019	0.0316	0.0013	$\rho = 0$	-0.0072	0.0001	-0.0046	0.0008

Table 3.5: Non-stationary Estimated Smooth Functions - Performance (Case 1)

	Factor Smooth: Level (0), (x2)x1:(0)	Factor Smooth - Level (1), (x2)x1:(1)
	Average Mean Squared Error (AMSE)	Average Mean Squared Error (AMSE)
Polynomial	0.02734607	0.0313961
Linear	0.03587293	0.0370131

related to the fixed effect overall. Nevertheless, the results seem to be quite manageable and signifies that an assumed conditionally serially correlated count process can be well fit by a generalized additive mixed model with an autoregressive coefficient via the penalized quasi-likelihood estimation in the cases for which we have time-invariant and time-varying coefficients with factor smooths as biases never exceed $|0.08|$. In the cases in which either the autoregressive AR(1) coefficient or random effect is not present in the dataset (or misspecified cases, e.g. $\sigma_\nu^2 = 0$ or $\rho = 0$) with factor smooths, these semi-parametric models seems relatively capable of its detection where biases never exceed $|0.12|$ in any case. This seems to align with the main result from [39], which suggests that, generally, although the penalized quasi-likelihood estimators are biased in finite samples, the asymptotic bias of the PQL coefficients associated with the parametric portion of our GAMM (re-parameterized as a GLMM) as well as our variance components (of order $o(||\theta||)$), no bias correction is needed (at least relative to the case of binary outcomes). Given these promising simulation results, we now turn to an application continuing from Chapter 2.

3.6 Application: Household weekly visits during non-promotional period 2

In Chapter 2, we developed a IV-GMM approach to estimate the impact of coupon promotions on the rate of customer visits in the presence of endogeneity for Type A campaigns. We found that all campaign exposures in general had positive and significant impact of exposure effects relative to non-promotional periods in the promotional cycle (or washout periods). Here, we investigate the carryover of the campaign exposure effects from the prior promotional period preceding this non-promotional period. We aim to model the magnitude and duration of the increased visit rate relative to households' baseline in a specified period of non-promotion immediately following the first promotional period via a potentially nonlinear function of time. Additionally, as flexible as GAMMs are for capturing potentially nonlinear effects, there exists an equal need to properly capture the non-independence attributed to individual level heterogeneity and conditional temporal dependence associated with repeated visits of households after the first promotional period expire. We focus on the period immediately following the first promotional period. Since this is the second period of non-promotion (the first being the baseline period), we refer to it as non-promotional period 2 or NP2. We intuit that the carryover effect of prior promotions is to attenuate or decay over the time points in NP2. In addition to this suspected attenuation, we place no specific functional form restriction on the suspected attenuating carryover effect for each of the different exposures, signifying that the effect can vary linearly or non-linearly across time, requiring the use of a generalized additive mixed model with factor smooths.

In analyzing any type of carryover effect of the first promotional period's promotion on the visit rates in NP2, we continue to use our longitudinal or panel data format; however, we adjust the time dimension. Since our aim in this chapter involves the analyzing the first non-promotional period beyond our baseline period 1 during the test cycle, our scale for our cross-section of interest is on a weekly basis where we observe

the visit counts of household i in week j . We utilize a covariate, NP2Week, that highlights the weekly time increment of interest spanning this intermittent period of campaign exposure (weekly visits). For each level, we condition on the exposure effect in the previous period and estimate a separate, non-linear functional model without a strict presupposition of the preceding promotional campaign treatments' linear relationship to the responses in NP2. We include an offset, t_i^* , which now represents each individual household's baseline visit rate in period 1. This offset contributes to a ratio of rates that will allow us to measure the visit rates in our adjusted response variable y_{ij}^* (visits of household i in week j) relative to its baseline rate of visitation in the households' pre-promotional period. This is of paramount importance in this analysis because the visiting behavior varies substantially more in NP2 relative to the panel structure in Chapter 2. The ratio of rates formulation will allow us to interpret a strictly parametric covariate's impact on weekly visitation rates relative/over each consumer's baseline period.

Lastly, we decide to concatenate the three main prior exposure effects of interest ("Type A", "Type B", and "Type C") into two categories for our models which represent the prior periods' campaign exposure:

- Type A $\rightarrow$ Personalized Rewards Campaign (P)
- Type B and Type C $\rightarrow$ Targeted Rewards Campaign (T)

The decision to combine the exposure effects "Type B" and "Type C" is made in order to circumvent the low frequency or instances of "Type C" campaign exposure relative to "Type A" and "Type B" campaigns. In the next subsection, we provide aggregate descriptive statistics that reveal average rate differences. It is important to note multiple campaign exposure periods never precede NP2; there was only one promotional exposure before NP2. Only single effects of a prior promotional period are valid as there are no joint campaign exposure effects that ever precedes a subsequent period of non-promotion as campaigns "expire" at different and non-overlapping times as discussed in Chapter 2.

3.6.1 Aggregate Descriptive Statistics in Periods of Non-Promotion

We descriptively capture the first five non-promotional periods (NP1 represents the baseline cycle, whereas NP2-NP5 represents non-promotional periods during the test cycle immediately after promotional periods. The initial baseline rate is defined to be the visits per week of households during number of non-promotional days in the baseline. For the remaining households that received promotional campaigns later, the rate of visits per week has been calculated (test phase). The variability of visiting behavior for the aggregate households' cohort can be seen in the proceeding analysis by observing the reversion to states of non-promotion beyond the initial control phase of NP1. Some households received only one campaign; this explicitly signifies that such households will have solely two non-promotional periods (the days that spans the pre-promotional phase corresponding to a household's respective baseline as well as the reversion to an epoch of non-promotion after being exposed to either a personalized {Type A} or targeted {Type B, Type C} campaign immediately after the loyalty promotional program begins).

Our statistical models in the next section for non-promotional periods will examine, for households, the difference in the rate of visits per week for the first baseline period versus the immediate subsequent non-promotional periods during the test phase. The average and standard error of these differences have been captured in Table 3.6.

Table 3.6 indicates the average rate difference of visits per week relative to NP1 (baseline) shows an increase or significant lift relative to the baseline period. This suggests an increased visitation rate, at least descriptively, after households were exposed to a campaign from prior promotional periods for the four non-promotional periods after the baseline. It is notable that, although we observe an increase, the magnitude of increase for the rate of visits per week decreases substantially across non-promotional periods. It is not clear at this point why we observe such a decrease in the later non-promotional periods, but it could be representative of an overall fatigue effect in campaign the latter cycle of the study. Additionally, the aggregated descriptive summary in table 3.6 above gives insight into visiting behavior through these periods of non-promotion, but a more in-depth analysis of the data must be addressed.

For modeling considerations, another important omission not displayed descriptively in Table 3.6 is accounting for how active a household is likely to be based off receipt of promotional exposures in the test cycle. In terms of fitting our non-promotional model GAMM, we will consider an additional integer-valued covariate, “Activity Level”³², that can serve as a proxy for a household’s utility to the entire loyalty promotional program. The larger values of “Activity Level” represent the most highly active customers receiving the most campaigns who remain active until the end of the promotional phase. Active customers in baseline phase are likely targeted more, which means the individuals contained in such a cohort have shorter periods in the test phase and are likely to have truncated non-promotional periods as compared to less active customers. Controlling for activity level gives us greater control for accurately estimating the carryover effect of the promotional campaigns’ prior exposure effects. Given the repeated weekly visits for our households, we are likely to have temporal autocorrelation which will be considered in our model fitting process. This leads us to the formulation of our primary hypothesis for our models.

H1: The personalized and/or targeting campaign’s exposure in a prior promotional period has a significant marginal impact upon the weekly visitation rate in NP2 relative to the baseline visitation rate in the baseline period (NP1).

Table 3.6: Average Rate Difference and Standard Error of Rate Difference in Visits per Week

Non-promotional Periods	N	Average Rate Difference	SE(Rate difference)	% Decrease (N)
Non-Promotional Period 1 (NP1)	1584	-	-	-
Non-Promotional Period 2 (NP2)	1584	0.066	0.0083	0%
Non-Promotional Period 3 (NP3)	815	0.035	0.0064	48.85%
Non-Promotional Period 4 (NP4)	648	0.011	0.0084	20.50%
Non-Promotional Period 5 (NP5)	435	0.001	0.0160	32.87%

³²This variable can also be interpreted as the number of “periods” a household has in Chapter 2.

3.6.2 R Package: `mgcv`

We will utilize R’s `mgcv` package to aid in our analyses ([9]). This package offers a variety of smoothing bases used for the smoothing function estimation such as cubic regression splines, p-splines, and thin plate regression splines. The distributional families for modeling the response include those from the `stats` package for the `glm()` function. A caveat, however, is that all non-normally distributed responses must be estimated by penalized quasi-likelihood. This function uses the `nlme` which essentially transforms smooth terms into “random effects” for estimation ([50]).

This function is advantageous as it allows all semi-parametric GAMM models to be re-parameterized into a generalized linear mixed model (GLMM) based on a “surrogate” response vector and it is, generally, computationally faster than its analog `gam`.

3.6.3 Non-Promotional Period 2 (NP2) GAMMs

There are 66 weeks that span the first non-promotional period immediately following the first promotional period with 1,584 households and $n = 18821$. Throughout the entire 66 weeks, there were approximately 21.32% of the observations that received a personalized campaign in the previous promotional period, and 71.68% that received a targeted campaign. We observe the aggregated average weekly visits (weekly visits divided by number of households) across all the weeks in the non-promotional period 2 (NP2) given the campaign type received in the first promotional period (Previous Campaign Type). Recall, the personalized campaign types (P) refer to campaigns that distributes customized coupon sets, whereas targeted campaign types (T) refer to campaigns that distributed either themed-event coupons or solo mailers. As seen in Figure 3.2, the average visits per week initially follow an attenuating trend; yet, for those who previously received a targeted campaign is higher in the initial weeks of NP2. At approximately the 16th week in NP2, the average weekly visits for both a prior personalized and prior targeted campaign are relatively similar, followed by four weeks in which the personalized campaign average weekly visits are higher. Beyond the 20th week in NP2, the trends for average visits per week seems to be relatively inconsistent and quite variable.

Overall, the trajectories of these average weekly visits given a prior promotional campaigns does seem to show a relatively decreasing trend overall across the weeks in the NP2. However, the degree to which these averages attenuate is of interest particularly once accounting for households’ individual baselines prior to receiving the first promotional campaigns is of principal interest. Does mere exposures to prior promotional campaigns positively induce a higher weekly visit rate over the baseline in this non-promotional epoch? If so, is the exposure of the time purely linear? We explore these questions in the subsequent models. For the model formulation, we utilize a subject-specific random effect for the cross-section of households which provide utility in our modeling scheme as we have repeated weekly visits for households in this non-promotional period. Due to the repeated weekly visits within each household i in NP2, there is likely non-independence or correlation within each cluster which necessitates the need for incorporating a simple random effect

into the full predictor in order to capture the clustering across households. Visiting behavior from within a household i across the time points in NP2 are likely to be similar relative to visiting behaviors between households, the utility of its inclusion in this longitudinal setting is obvious. These random effects are used to correctly capture these correlations between observations and represent the influence each household on the repeated, in-store visits that is unaccounted for by the observed covariates [51]. The degree of heterogeneity of households is captured by the variance. Additionally, there is a likely chance that in-store grocery visits has a temporal association or are autocorrelated within each household. It is believed that the baseline average weekly visits can offers useful interpretation regarding the visits per week after the first promotional period (NP2) in terms of covariate interpretation, necessitating the utility of a baseline constructed offset. Thus, we consider the rate of visits in NP2Week over the baseline average weekly visits for individual i by including $\text{offset}=\log(\text{Baseline Average Weekly Visits})$ in the mean function of the GAMM. This enables the assessment of the weekly rate of visits in NP2 over the individual-level baseline visit per week. Lastly, the variable Previous Campaign Type's utility in our model formulation is two-fold: we are able to assess the parametric/linear partial effect on weekly visit rate in NP2 relative to each household-specific baseline visit rate and the potentially nonlinear smoother for the interactions between NP2Week and Previous Campaign Type. For the interaction³³ between $s(\text{NP2Week})$ and Previous Campaign Type, the smooth for each level are centered. This signifies that if the means of the values at each level of the Previous Campaign Type are different, then the difference in means of the model will not degrade or be taken into account in any of the fitted GAMMs ([9]).

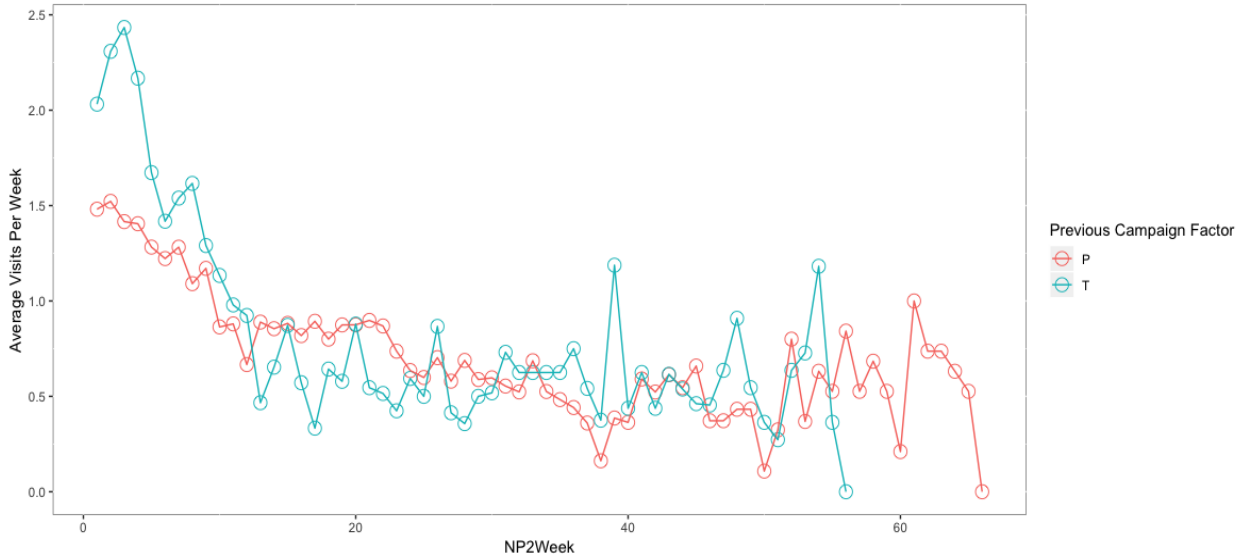


Figure 3.2: Average Visits Per Week given Previous Campaign Type

³³The estimated smooth effect of the prior promotional campaign's type varying over the temporal covariate NP2Week.

The model takes the form

$$\log\left(\frac{E[y_{ij}|\mathbf{x}_i, \boldsymbol{\nu}_i]}{t_i^*}\right) = \boldsymbol{\eta}_{ij} = \mathbf{x}_i' \boldsymbol{\beta} + f(w_{ij})x_{2i} + \mathbf{z}_{ijR}' \boldsymbol{\nu}_i$$

where $E[y_{ij}|\mathbf{x}_i, \boldsymbol{\nu}_i] = \mu_{ij}$ is the conditional expected count of visits for individual i in week j , $\mathbf{x}_i = (1, \text{Activity Level}, \text{Previous Campaign Type})'$, w_{ij} is the covariate representing is the household's time (in weeks) from the start of NP2, x_{2i} corresponds to the Previous Campaign Type (represented as a factor covariate) in the first promotional period, $\boldsymbol{\nu}_i$ is the household-specific random effect assumed $\boldsymbol{\nu}_i \sim N(0, \boldsymbol{\Sigma}_\nu)$, $\mathbf{z}_{ijR}$ is the design matrix corresponding to the random effect³⁴, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ are the coefficients corresponding to the parametric terms in the model, and t_i^* is the offset variable defined to be the baseline average weekly visits for each household i for $i = 1, \dots, 1584$, $j = 1, \dots, n_i$ ($1 \leq j \leq 66$). Also, the log link function is used that relates the expected visits per week in NP2 to the mean function. The Previous Campaign Type factor is specified in the parametric portion of the GAMM in addition to its interaction with NP2Week. Its inclusion in the parametric portion of this semi-parametric model is necessary to ensure the identifiability, “sum-to-zero”, constraint is met as discussed in a prior subsection³⁵ ([9],[10]). Thus, Previous Campaign Type is centered at zero and to ensure that if there were any significant difference in mean visits per week for level P vs. T, the estimated fit from the smooth from, say, the reference level (T) is shifted up or down by some constant relative to the other level (P) requiring this covariate to be specified in the parametric portion of the model³⁶. In sum, the parametric portion of the predictor reflects shifts in the log counts at the mean of NP2Week while the factor smooth interaction captures nonlinear variation away from that point similar to interactions between linear and factor terms in regression.

As shown in Table 3.7, we first fit two Poisson GAMMs to our non-promotional data and again utilize the default basis the thin plate regression spline with $d = 1$ since we have one time-varying covariate in the smooth function (NP2Week), penalizing the second degree derivative $m = 2$. As frequently present in longitudinal data, monotonically decreasing correlations among residuals are commonly modeled with an AR(1) process ([52]). This motivates the inclusion of an AR(1) covariance structure for the GAMM model comparison. The first model is fit without an AR(1) coefficient whereas the second is fit with a first-order autocorrelation component given the Gaussian random effects considered in each case. For the longitudinal response, visits per week in NP2, we assume the responses follow a Poisson distribution. However, as we are interested in the lingering effect of the prior promotional campaign's effects on weekly store visits in NP2, we

³⁴Since we only consider random intercepts in the GAMM formulation, $\mathbf{z}_{ijR}$ is a sparse “matrix” of 0s and 1s only that indicates which week a household visited in. Thus, the element will be 1 if household i visited in NP2Week j , and 0 otherwise.

³⁵Particularly computing GAMMs in the **mgcv** package, the default basis for $\mathbf{s}()$, two bases are automatically computed in the null space and unpenalized: the linear function in the parametric portion and a horizontal function (congruously, the model's intercept). Any value could be added to the intercept term and subtracted from the horizontal function, resulting in the same fit with a new model, resulting in an infinite number of model fits. Therefore, over the covariate values' range, it is necessary to guarantee the smooth is constrained to sum to zero to ensure the horizontal function is removed from the unpenalized basis of the smooth.

³⁶The parametric terms represent the deviation of the indicated group from the mean of the reference group. Without its inclusion in the linear predictor, the resulting smooths from the GAMM fits may become misspecified as they'd become “wiggly” by virtue of trying to account for the mean difference/shifts within the levels of Previous Campaign Type.

Table 3.7: Poisson GAMM Results

Response: Visits Per Week	Poisson GAMM	Poisson GAMM AR(1)
Intercept	-0.4600319*** (0.052285)	-0.451617*** (0.052186)
Previous Campaign Type (P)	-0.170710*** (0.04778)	-0.171377*** (0.048029)
Activity Level	0.077849*** (0.003327)	0.077355*** (0.003308)
s(NP2Week):Previous Campaign Type (P) [EDF]	2.946*	2.638**
s(NP2Week):Previous Campaign Type (T) [EDF]	5.900***	5.567***
R^2	.166	.167
$\hat{\sigma}_{ID}^2$ (95% C.I.)	0.4568186 (Incomputable)	0.4401753 (Incomputable)
$\hat{\rho}$ (95% C.I.)	-	0.06772976 (Incomputable)
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$		

employ factor smooth interactions with Previous Campaign Type over the weeks in NP2. For completeness, this application directly ties into our Simulation Design II. described in Section 3.5 where the trend covariate is clearly time-varying when modeling correlated Poisson counts with a non-stationary temporal structure.

From the results, we first observe the parameter estimates from our fixed, parametric component of the GAMM. For example, in assessing the Poisson GAMM coefficient for the parametric Previous Campaign Type (P), we note that holding all covariates constant, we’d expect the weekly relative visit rate over the baseline period would be $1 - \exp(-0.170710) \approx 15.69\%$ less compared to those in the sample receiving a targeted promotion (T). Thus, the targeted promotions received from the first promotional campaign have a stronger relative effect on visit rates compared to the personalized ones. All the strictly parametric coefficients in Table 3.7 can be interpreted similarly. For the factor smooth terms, the values represent the EDFs, with both being statistically significant. The further values are away from 1 indicates the more “wiggly” or nonlinear one would expect the factor smooths to be. We now turn to the parameter estimates for the variance of the random effects along with the autoregressive coefficient for the AR(1) temporal specification in each model fit. For example, the resulting random effects variance for the Poisson GAMM and the Poisson GAMM AR(1) models are 0.4568186 and 0.4401753, respectively. These values represent the estimated variances attributed to the clustering or dependencies of visits per week between households after controlling for the included covariates. While we are able to retrieve the point estimates easily from our GAMM fits, we are unable to compute the “approximate 95% confidence intervals” for these variance components³⁷. Approximate confidence intervals are computed using the `intervals.lme()` function from the `nlme` package in R. In both model cases, an error is returned: `cannot get confidence intervals on var-cov components: Approximate variance-covariance matrix not available`. As errors are returned in both cases, this seems to suggest that there is an error with the log (quasi) likelihood function to the point in which no other optimization can occur, and that the Hessian is non-positive definite.

In other words, the covariance structure is misspecified in the model fitting procedure as the Hessian calculated from the models’ fitted values is not positive definite. With a non-positive definite Hessian, the

³⁷These are “approximate confidence intervals” assuming a Gaussian approximation to the distribution of the REML estimators. These estimators are assumed to have a normal distribution centered at the true values given a covariance matrix equal to the negative inverse Hessian matrix of the log-likelihood evaluated at the estimated parameters ([29]).

(estimated) covariance matrix is incomputable (as the Hessian is the inverse of the covariance matrix) ([46]). As this is the case, both of these Poisson GAMM models seem to be misspecified. There is a likely issue that the response of the distribution is misspecified due to potential *overdispersion*.

A closer look at the mean/variance comparisons of visits per week in NP2, we turn our attention to the visiting behavior in NP2 for our sample. As can be seen in the time plot highlighting both the mean (left axis) and variance (right axis) of visits over the weeks in NP2 for Figure 3.3, there is a clear lack of matching measures of center and dispersion across weeks (represented by the lines). Essentially Figure 3.3 shows that the variance of counts is consistently larger than mean visits across the weeks of NP2 indicating overdispersion. This motivates the replacement of the Poisson assumption with a negative binomial assumption for the responses. The negative binomial is capable of handling overdispersed counts. Failure to account for this extra variability would not effect the parametric parameter estimates, but would distort inference by inflated standard errors which affects hypothesis tests and confidence intervals. As the variance of a negative binomial model is a quadratic function to the mean, employing its distribution has a particular effect on the model fitting process in terms of estimating regression coefficients (relative to, for example, the quasipoisson).

In a generalized linear (mixed) model's iterative process for estimating $\hat{\beta}$, utilizing a negative binomial imbues the terms (or elements) inside the matrix $\mathbf{W}$ with an incurvated or concave weight relative to the size of the mean values. In essence, smaller mean values receive lesser weights, whereas larger means has weights that approach κ ([53]). This model form can be easily accommodated in a GAMM (GLLM) framework for the model fitting process. As it relates to estimation using the `gamm()` function from the `mgcv` package, and in order for the negative binomial distribution to be a member of the exponential family, there is a requirement for the shape parameter κ to be fixed or known. That is, for any fixed shape parameter κ , the negative binomial density function is

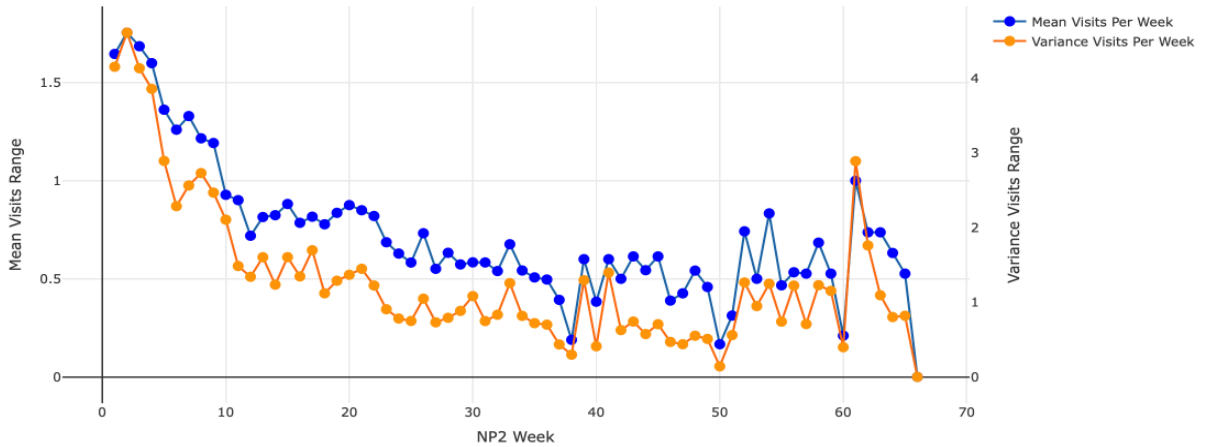


Figure 3.3: NP2: Mean and Variance Visits Per Week

$$NB(y_{ij}|\mu_{ij}, \kappa) = \exp\left(\frac{y_{ij}\xi_{ij} - \zeta(\xi_{ij})}{\phi} + c(y_{ij}, \phi)\right)$$

where $\xi_{ij} = \log(\frac{\mu_{ij}}{\mu_{ij} + \kappa})$, $\phi = 1$, $\zeta(\xi_{ij}) = -\kappa \log(1 - \exp(\xi_{ij}))$, and $c(y_{ij}, \phi) = \log\left(\frac{\Gamma(y_{ij} + \kappa)\kappa^\kappa}{\Gamma(\kappa)y_{ij}!}\right)$.

For utilizing the negative binomial in the `gamm()` function, the shape parameter is required to be known in order for it to coincide with the exponential family condition for estimation. To estimate κ (the shape parameter), we utilize the analog `gam()` with `family=nb()` and use this estimated value of κ as the shape required in the `family` argument of the `gamm()` function as the estimation, generally, is equivalent, or at least provides a reasonable estimate for use in the GAMM formulation ([28]).

The NB GAMM model is fitted with the random effect only and the NB GAMM - AR(1) is fitted with an autoregressive component of order 1 and a random effect and using the default basis for the thin-plate low rank smoother. As seen, the parametric coefficients are all of similar magnitudes in terms of both the estimates and the standard errors. All are highly statistically significant at an $\alpha = 0.05$. For the parametric factor smooths, or the terms corresponding for allowing the linear effect of Previous Campaign Type to vary smoothly with NP2Week, or the week number spanning the weeks of the intermittent promotional campaigns, the values displayed are the effective degrees of freedom (EDF). The higher the values of the EDF determines the amount of nonlinearity associated with the factor smooths. Note, the value of 1.000 indicates a strictly linear effect for NP2Week by prior campaign type interaction.

Additionally, as previously seen in the Poisson GAMM case, the NB GAMM model (without any consideration of serial correlation) is perhaps a bit misspecified as it relates to the estimation of the variance-covariance components (hence the approximate 95% confidence interval for the estimate of the within-subject variability cannot be computed). The approximate 95% confidence intervals for the GAMMs were computed using the `intervals.lme()` function as before.

As this is the case with our previous Poisson fits, this motivates the inclusion of additional complexity that is likely caused by temporal autocorrelation for households' visiting behavior across weeks within each cluster. Therefore, an autoregressive component of order 1 is again considered. Specifically, we utilize the covariate NP2Week as the trend covariate with household key as the grouping variable to specify that the residuals themselves are only correlated within the household keys and that visiting behavior correlation relative to baseline visiting behavior decreases exponentially further out in the weeks spanning NP2Week, conditional on the random effect. Additionally, from Table 3.8, there is no issue with convergence as the (approximate) intervals for both the household keys' standard deviation and the estimated parameter $\hat{\phi}$ are computable. Although not displayed, the variance of the working residual from the NB GAMM - AR(1) model is 0.9686372, which, combined with the estimated variance attributed to the household keys' variance (0.5259956²), indicates that the proportion of variation attributed to households' visit rates due to clustering over the baseline visit rate in NP1 is approximately 22.2% (as $\frac{0.5259956^2}{0.5259956^2 + 0.9686372} = 0.2221709$). As the entire confidence interval for the random effects' variance does not contain zero, this suggests that

the random effect is indispensable for modeling as there is a notable of variability, after controlling for the fixed terms in our model, that is attributed to the clustering mechanism. Additionally, the smoothing parameters $\hat{\omega} = (\lambda_P^{-1}, \lambda_T^{-1})$ which were jointly estimated along with random effects via penalized quasi likelihood for both the unknown smooth functions that vary by levels of previous campaign types' exposure effects. In our case, we have $\hat{\omega} = (0.000108887, 2.226861e^{-17})$ which signifies that $\hat{\omega}^{-1} = (\lambda_P, \lambda_T) = (9183.83, 4.490627e^{16})$. As $\lambda \rightarrow \infty$, a straight line is expected for the estimated smooth function. Thus, it is quite clear that the factor smooth associated with the targeted campaign is completely linear. We will observe a plot of the factor smooths momentarily. In Figure 3.4, we compare the autocorrelation of the residuals from the NB GAMM model no additional temporal correlation structure with the normalized residuals from the AR(1) consideration³⁸. The normalized residuals take into account the specified AR(1) of the temporal correlation structure specified in the model. The serial correlation presented in the autocorrelation function for the NB GAMM with random effects only could also explain the non-convergence in the variance-covariance parameters in Table 3.8. Nevertheless, the lack of significant residual autocorrelation from the normalized AR(1) specification indicates that the standard errors associated with the parametric coefficients are appropriate. We consider this the best GAMM model and proceed with interpretations. We turn our attention to the parametric coefficients associated from Table 3.8. They are all statistically significant at the level $\alpha = 0.01$, which verifies **H1**. The coefficient for Previous Campaign Type P is -0.31560. Exponentiating for ease of interpretation, we have $\exp(-0.31560) = 0.7293511$.

An interpretation for the Previous Campaign Type P would be that if we were to hold all other covariates constant, then the weekly visit rate relative to the baseline period is 27.06% less compared to receiving a targeted promotion (T) at the mean of NP2Week. Reasonably, this could perhaps be attributed to the lack of a specific degree of true or adequate personalization associated with the personalized campaigns for the cohort as this models visiting behavior based solely on the at least 232 days of devising these campaigns in the control period. In other words, the mechanism of personalized campaign exposure based on customer loyalty and personal buying profiles resulting from buying behavior in the baseline period may not be quite tailored personally enough to induce in-store visiting relative to the themed coupons

Table 3.8: Negative Binomial GAMM Results

Response: Visits Per Week	NB GAMM	NB GAMM - AR(1)
Intercept	0.20754*** (0.04684)	0.21927*** (0.04654)
Previous Campaign Type (P)	-0.30847*** (0.04091)	-0.31560*** (0.04065)
Activity Level	0.01806*** (0.00280)	0.01776*** (0.00278)
s(NP2Week):Previous Campaign Type (T) [EDF]	1.814	1.000**
s(NP2Week):Previous Campaign Type (P) [EDF]	5.820***	5.589***
R^2	0.357	0.355
$\hat{\sigma}_{ID}^2$ (95% C.I.)	0.2909524 (Incomputable)	0.2765708 (0.2487416, 0.3072817)
$\hat{\rho}$ (95% C.I.)	-	0.05802 (0.0404, 0.07886)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

³⁸The normalized residuals are the standardized residuals that are pre-multiplied by the inverse of the square root of the error correlation matrix estimated from the GAMM fit.

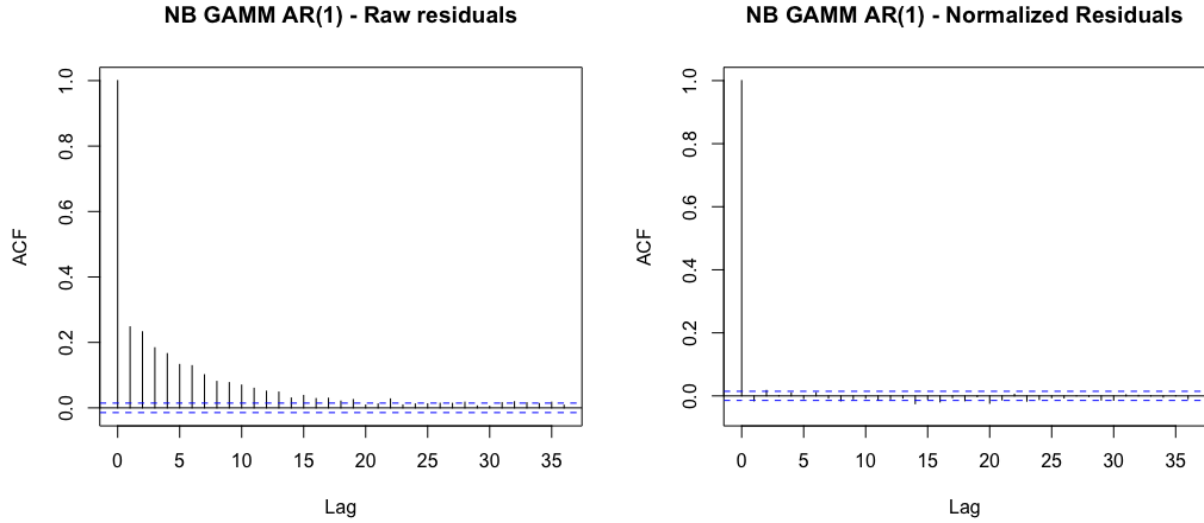


Figure 3.4: Raw Residuals vs. Normalized Residuals

(which offers discounts based on groups of related products) as it relates to transition into the first non-promotional period. Also, it could be the combined effect of both Type B and Type C, as they are aggregated here, could have a stronger parametric effect on the response.

In observing the interaction of $s(\text{NP2Week}, 5.59) : \text{Previous Campaign Type}(P)$ in Figure 3.5, we observe the partial response plot of this isolated effect on the response over the weeks in the first intermittent promotional cycle beyond the baseline. Although the models are estimated with the entire span of weeks in the second non-promotional period, we choose to plot the first 15 weeks (instead of the entire 66 weeks that span NP2) in Figure 3.5. Our reasoning for doing this is to graphically highlight the realistic prior promotional period's lasting campaign effect on the weekly visits (as any realistic effect would be most believable in the beginning weeks). Note, in both plots for 3.5, we add the intercept, 0.21927, to all values of the fitted response as to ensure that the y-axis represents the estimate of the response (adjusted for the offset) given the partial effect on the log scale. Thus, in both plots, the y-scale represents the expected log weekly visits in NP2 over the baseline visits for the effect of either NP2Week:Previous Campaign Type (P) or NP2Week:Previous Campaign Type (T). Note, these partial response plots do not include any other parametric coefficients in its interpretation (such as Activity Level). The solid line represents the estimated value of the response as a function of NP2Week. The dotted blue lines on either side of the solid line represents ± 1 standard error of the estimates of the nonlinear smoothers. Overall, we observe that across the weeks of NP2Week, the weekly visit rate of those who received a personalized campaign descends below the intercept at week 5 and remains below the intercept throughout the remaining 10 weeks (with a slightly more positive effect beginning at week 11-13). In the beginning weeks (weeks 1-4), the estimated response via this nonlinear function of the interaction is indeed above the intercept (mean), which implies that the personalized promotion received during the first promotional period does contribute to a lingering,

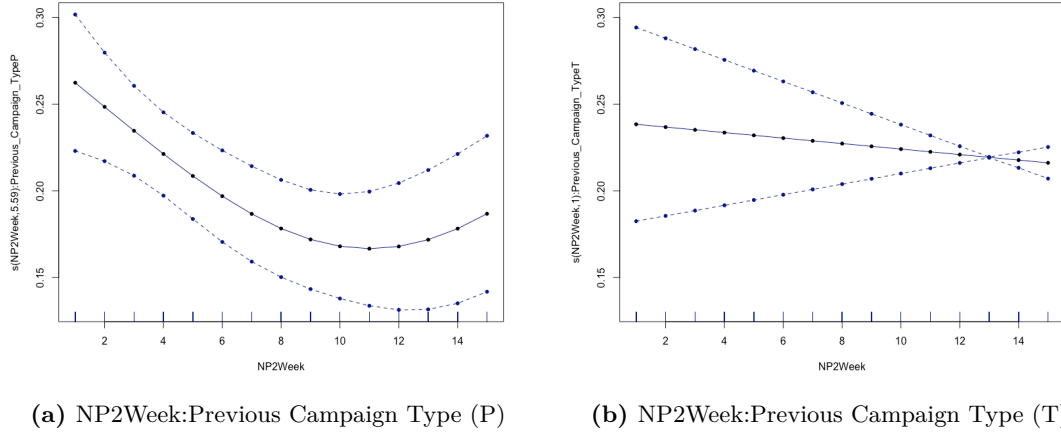


Figure 3.5: Partial Dependence Factor Smooth Plots - NP2

nonlinear exposure effect for increased visit rate in the first few weeks in this non-promotional cycle relative to the baseline period. This suggests that not only does the promotional campaigns' exposure effects clearly induces a higher visit rate in all promotional periods relative to no campaigns (as seen in Chapter 2), but also that the campaign effect from the first promotional period maintains a positive effect or carryover effect on visits into the first intermittent cycle of non-promotion. Most importantly, however, is that the GAMM formulation is able to capture the nonlinearity of the interactive effect particularly around weeks 7-15. Slightly differing behavior can be assessed from the partial plot representing the $s(\text{NP2Week}, 1.00)$:Previous Campaign Type(T) smoothing interaction as well. Firstly, we note that the factor smooth interaction of the prior targeted campaign effect given the weekly trend is completely linear (as indicated by the EDF value of 1.00). This could be attributed due to the fact that the targeted and solo campaign mailers in the prior period are not customized to the active households' needs, thus having a linear, or "expected" effect for the immediate weeks of non-promotion on visits. For the first 13 weeks, the estimated response exceeds the intercept, which implies that a lingering effect holds for increased weekly visitation over the baseline given the households received a targeted campaign in the first promotional period. This large carryover campaign exposure effect indicates that the targeted campaigns are quite effective in increasing visits over the baseline, and substantially more relative to the personalized campaigns. This suggests households respond much more strongly from prior exposures alone to targeted campaigns at least in the beginning of the promotional cycle. At week 13, however, there does seem to be very tight bound factor smooth interaction's estimate with the added intercept for the first 15 weeks of NP2. The corresponding lower and upper limits are provided in Table 3.9.

Table 3.9: NP2 Factor Smooth Interactions - Estimates and Limits

NP2 Week Factor Smooth Interactions	s(NP2Week):Personalized (P)			s(NP2Week):Targeted (T)		
	Point Estimate	Lower Limit	Upper Limit	Point Estimate	Lower Limit	Upper Limit
Week 1	0.262408805	0.223029905	0.301787704	0.23838785	0.182459407	0.294316297
Week 2	0.248469312	0.217125947	0.279812677	0.23679981	0.185516968	0.288082647
Week 3	0.234680002	0.208738905	0.260621098	0.23521176	0.188574515	0.28184901
Week 4	0.221274575	0.19722711	0.245322039	0.23362372	0.191632045	0.275615392
Week 5	0.208563618	0.183722835	0.2334044	0.23203567	0.194689554	0.269381796
Week 6	0.196904696	0.170478018	0.223331374	0.23044763	0.197747041	0.263148225
Week 7	0.186672564	0.159118997	0.214226131	0.228859593	0.200804499	0.256914687
Week 8	0.178230698	0.150156554	0.206304843	0.227271557	0.20386192	0.250681193
Week 9	0.171905116	0.143255236	0.200554995	0.225683524	0.206919282	0.244447767
Week 10	0.167961075	0.137801196	0.198120953	0.224095498	0.209976533	0.238214462
Week 11	0.166582985	0.133622887	0.199543082	0.222507478	0.213033514	0.231981442
Week 12	0.167857623	0.131260864	0.204454382	0.220919466	0.216089442	0.22574949
Week 13	0.171760763	0.131540298	0.211981227	0.219331463	0.219077749	0.219585178
Week 14	0.178147468	0.13502365	0.221271285	0.217743471	0.213275639	0.222211304
Week 15	0.186746632	0.141729763	0.231763501	0.21615549	0.207043936	0.225267044

3.7 Discussion

Summary: In Chapter 3, we exploit the utility of penalized spline regression as a direct connection to the mixed model framework in order to explore the value of generalized additive mixed models (GAMMs). A relatively overlooked component of GAMMs in literature is the ability to use nonlinear interactions via smoothing covariates with any levels of factors.

Via broad simulations designs, we investigated the simultaneous estimation of the variance parameters, fixed effects, and functions of smoothed covariates' interactions via factor smooths for conditionally serially correlated AR(1) Poisson count models with random effects for both unbalanced and balanced longitudinal structures via standard performance metrics. We relied exclusively on penalized quasi-likelihood estimation (PQL) for parameter estimations in all cases. Additionally, we further parsed parameter performances given situations of stationary vs. non-stationary temporal, AR(1) correlation structures with the inclusion of time-independent vs. time-dependent (time-varying) covariates. Regarding the smooths functions, we simulated the true curve to take on a complex polynomial structure as well as a strictly linear one. We varied parameters ranging from the subject/cluster size, the number of observations per subject, and realistic magnitudes of the variance associated with a random effect (Gaussian distributed random intercept) and the parameter corresponding to a temporal, autocorrelation AR(1) model. In all simulation designs, we decided to keep constant the fixed effect associated with one level of the parametric model portion constant in all simulation designs. In all cases of the simulation designs of conditionally serially correlated AR(1) Poisson counts for longitudinal structures, the generalized additive mixed models seems to recover the variance parameters very well in addition to the fixed effect. Regarding the thin plate regression splines' approximation to the true polynomial and linear functions we simulated, we note that these optimal, low rank, isotropic smoothers performs well in terms of approximation and shape recovery. Even in the misspecified cases in which we manually set either the variance of the random intercept or the AR(1) parameter to zero, our GAMMs

were basically able to appropriately identify and estimate those cases for both unbalanced and balanced considerations.

In a continuing application, we fit Poisson GAMMs with AR(1) temporal correlation structure given the random effect in order to assess any lingering effect of the prior promotional period campaign type on visits relative to the cross-section baseline’s visit in second non-promotional period (NP2) with arguably tenuous results. While the parametric parameter estimates seem reasonable as well as the partial dependence factor smooth fits, there was a bit of an issue in terms of computing approximately normal 95% confidence intervals for the variance parameters, indicating model misspecification. After taking into account the phenomenon of overdispersion using an extra parameter relative to the Poisson utilizing the negative binomial, we re-compute our GAMMs with confidence in these results. The NB GAMM AR(1) model captures the random effects’ variance as well as the estimate for the AR(1) parameter and their respective approximately normal confidence intervals. This model seems to completely remove the autocorrelation in the normalized residuals. Overall, we were able to utilize the factor smooths to quantify the lingering exposure effects from a prior promotional period’s campaign on visits relative to the individual baseline. The results convincingly suggest that the personalized campaigns stimulate households’ visits relative to the baseline for four weeks after immediately outside the first promotional period with a nonlinear attenuation, whereas the targeted campaigns attenuate strictly linear with an increased stimulation of visits of thirteen weeks relative to the baseline.

References

- [1] S. Hamel, N. G. Yoccoz, and J.-M. Gaillard, “Statistical evaluation of parameters estimating autocorrelation and individual heterogeneity in longitudinal studies,” *Methods in Ecology and Evolution*, vol. 3, no. 4, pp. 731–742, 2012. [73](#)
- [2] M. P. Kain, B. M. Bolker, and M. W. McCoy, “A practical guide and power analysis for glms: detecting among treatment variation in random effects,” *PeerJ*, vol. 3, p. e1226, 2015. [73](#)
- [3] J. Ferron, “Reconsidering the use of the general linear model with single-case data,” *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 3, pp. 324–331, 2002. [73](#)
- [4] M. Sciandra, “Flexible modelling of serial correlation in glmm,” in *Riunione scientifica della Società italiana di Statistica*, CLEUP, 2008. [73](#)
- [5] G. J. McKeown and I. Sneddon, “Modeling continuous self-report measures of perceived emotion using generalized additive mixed models,” *Psychological methods*, vol. 19, no. 1, p. 155, 2014. [73](#), [81](#)
- [6] S. N. Wood, N. Pya, and B. Säfken, “Smoothing parameter and model selection for general smooth models,” *Journal of the American Statistical Association*, vol. 111, no. 516, pp. 1548–1563, 2016. [74](#)
- [7] S. N. Wood, “Thin plate regression splines,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, 2003. [74](#), [79](#), [80](#), [83](#), [89](#)
- [8] D. McNeish and K. Kelley, “Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations,” *Psychological methods*, 2018. [74](#)
- [9] S. N. Wood, *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017. [76](#), [77](#), [79](#), [80](#), [81](#), [83](#), [84](#), [85](#), [89](#), [95](#), [102](#), [103](#), [104](#)
- [10] S. N. Wood, “Low-rank scale-invariant tensor product smooths for generalized additive mixed models,” *Biometrics*, vol. 62, no. 4, pp. 1025–1036, 2006. [76](#), [104](#)
- [11] D. Ruppert, M. P. Wand, and R. J. Carroll, “Semiparametric regression during 2003–2007,” *Electronic journal of statistics*, vol. 3, p. 1193, 2009. [77](#)
- [12] M. Abrahamowicz and A. Ciampi, “Information theoretic criteria in non-parametric density estimation: Bias and variance in the infinite dimensional case,” *Computational statistics & data analysis*, vol. 12, no. 2, pp. 239–247, 1991. [77](#)
- [13] S. Wood, *Generalized Additive Models: An Introduction with R*. CRC Press, 2006. [78](#)

- [14] J. Duchon, “Splines minimizing rotation-invariant semi-norms in sobolev spaces,” in *Constructive theory of functions of several variables*, pp. 85–100, Springer, 1977. [78](#)
- [15] A. Brezger, *Bayesian P-splines in structured additive regression models*. PhD thesis, lmu, 2005. [78](#)
- [16] I. Basinas, V. Schlünssen, D. Heederik, T. Sigsgaard, L. A. Smit, S. Samadi, Ø. Omland, C. Hjort, A. M. Madsen, S. Skov, *et al.*, “Sensitisation to common allergens and respiratory symptoms in endotoxin exposed workers: a pooled analysis,” *Occup Environ Med*, vol. 69, no. 2, pp. 99–106, 2012. [80](#)
- [17] J. C. Gutiérrez-Estrada, I. Pulido-Calvo, I. de la Rosa, and B. Marchini, “Modeling inflow rates for the water exchange management in semi-intensive aquaculture ponds,” *Aquacultural engineering*, vol. 48, pp. 19–30, 2012. [80](#)
- [18] K. Millidine, I. Malcolm, C. Gibbins, R. Fryer, and A. Youngson, “The influence of canalisation on juvenile salmonid habitat,” *Ecological indicators*, vol. 23, pp. 262–273, 2012. [80](#)
- [19] C. M. Davey, D. E. Chamberlain, S. E. Newson, D. G. Noble, and A. Johnston, “Rise of the generalists: evidence for climate driven homogenization in avian communities,” *Global Ecology and Biogeography*, vol. 21, no. 5, pp. 568–578, 2012. [80](#)
- [20] R. O. Martínez-Rincón, S. Ortega-García, and J. G. Vaca-Rodríguez, “Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (*acanthocybium solandri*) in the mexican tuna purse-seine fishery,” *Ecological Modelling*, vol. 233, pp. 20–25, 2012. [80](#)
- [21] J. Feekings, V. Bartolino, N. Madsen, and T. Catchpole, “Fishery discards: factors affecting their variability within a demersal trawl fishery,” *PloS one*, vol. 7, no. 4, p. e36409, 2012. [80](#)
- [22] P. Guo and K. Li, “Trends in esophageal cancer mortality in china during 1987–2009: age, period and birth cohort analyzes,” *Cancer epidemiology*, vol. 36, no. 2, pp. 99–105, 2012. [80](#)
- [23] F. Donat and G. Marra, “Discrete responses in bivariate generalized additive models,” *arXiv preprint arXiv:1508.01302*, 2015. [80](#)
- [24] F. Donat, *Discrete Responses in Penalized Generalized Linear Models*. PhD dissertation, McGill University, 2016. [80](#)
- [25] K. Larsen, “Gam: The predictive modeling silver bullet,” 2015. [80](#)
- [26] A. Groll and G. Tutz, “Regularization for generalized additive mixed models by likelihood-based boosting,” *Methods of Information in Medicine*, vol. 51, no. 02, pp. 168–177, 2012. [81](#)
- [27] X. Lin and D. Zhang, “Inference in generalized additive mixed modelsby using smoothing splines,” *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 61, no. 2, pp. 381–400, 1999. [81](#), [86](#), [88](#)

- [28] G. L. Simpson, “Modelling palaeoecological time series using generalized additive models,” *BioRxiv*, p. 322248, 2018. [81](#), [107](#)
- [29] J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media, 2006. [83](#), [105](#)
- [30] N. M. Laird and J. H. Ware, “Random-effects models for longitudinal data,” *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982. [83](#)
- [31] E. M. Chi and G. C. Reinsel, “Models for longitudinal data with random effects and ar (1) errors,” *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 452–459, 1989. [83](#)
- [32] A. F. Sterrantino *et al.*, “Varying coefficient models as mixed models: reparametrization methods and bayesian estimation,” tech. rep., Department of Statistics, University of Bologna, 2013. [83](#), [85](#)
- [33] S. N. Wood, Z. Li, G. Shaddick, and N. H. Augustin, “Generalized additive models for gigadata: modeling the uk black smoke network daily data,” *Journal of the American Statistical Association*, vol. 112, no. 519, pp. 1199–1210, 2017. [83](#), [89](#)
- [34] N. E. Breslow and D. G. Clayton, “Approximate inference in generalized linear mixed models,” *Journal of the American statistical Association*, vol. 88, no. 421, pp. 9–25, 1993. [86](#), [88](#)
- [35] R. Wolfinger and M. O’connell, “Generalized linear mixed models a pseudo-likelihood approach,” *Journal of statistical Computation and Simulation*, vol. 48, no. 3-4, pp. 233–243, 1993. [86](#), [89](#)
- [36] W. Jang and J. Lim, “Pql estimation biases in generalized linear mixed models,” 2006. [86](#)
- [37] J. M. Ver Hoef, J. M. London, and P. L. Boveng, “Fast computing of some generalized linear mixed pseudo-models with temporal autocorrelation,” *Computational Statistics*, vol. 25, no. 1, pp. 39–55, 2010. [86](#)
- [38] X. Lin and N. E. Breslow, “Bias correction in generalized linear mixed models with multiple components of dispersion,” *Journal of the American Statistical Association*, vol. 91, no. 435, pp. 1007–1016, 1996. [86](#), [87](#)
- [39] X. Lin, “Estimation using penalized quaslikelihood and quasi-pseudo-likelihood in poisson mixed models,” *Lifetime data analysis*, vol. 13, no. 4, pp. 533–544, 2007. [86](#), [87](#), [91](#), [99](#)
- [40] D. Handayani, K. A. Notodiputro, K. Sadik, and A. Kurnia, “A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (glmm),” in *AIP Conference Proceedings*, vol. 1827, p. 020033, AIP Publishing, 2017. [87](#)
- [41] C. E. McCulloch and J. M. Neuhaus, “Generalized linear mixed models,” *Wiley StatsRef: Statistics Reference Online*, 2014. [87](#)

- [42] J. Jiang, *Asymptotic analysis of mixed effects models: theory, applications, and open problems*. Chapman and Hall/CRC, 2017. [87](#)
- [43] R. R. Corbeil and S. R. Searle, “Restricted maximum likelihood (reml) estimation of variance components in the mixed model,” *Technometrics*, vol. 18, no. 1, pp. 31–38, 1976. [89](#)
- [44] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, “Package ‘mass’,” [89](#)
- [45] S. N. Wood, “Stable and efficient multiple smoothing parameter estimation for generalized additive models,” *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 673–686, 2004. [89](#)
- [46] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC, 2014. [90](#), [106](#)
- [47] B. C. Sutradhar, *Dynamic mixed models for familial longitudinal data*. Springer Science & Business Media, 2011. [92](#), [93](#), [117](#)
- [48] “Explanatory variables in the ar (1) count data model,” [93](#)
- [49] B. C. Sutradhar and W. Bari, “On generalized quasilielihood inference in longitudinal mixed model for count data,” *Sankhyā: The Indian Journal of Statistics*, pp. 671–699, 2007. [93](#)
- [50] E. J. Pedersen, D. L. Miller, G. L. Simpson, and N. Ross, “Hierarchical generalized additive models: an introduction with mgcv,” *PeerJ Preprints*, vol. 6, p. e27320v1. [102](#)
- [51] D. Hedeker, “Generalized linear mixed models,” *Encyclopedia of statistics in behavioral science*, 2005. [103](#)
- [52] R. D. Wolfinger, “Heterogeneous variance: covariance structures for repeated measures,” *Journal of agricultural, biological, and environmental statistics*, pp. 205–230, 1996. [104](#)
- [53] J. M. Ver Hoef and P. L. Boveng, “Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data?,” *Ecology*, vol. 88, no. 11, pp. 2766–2772, 2007. [106](#)
- [54] B. C. Sutradhar *et al.*, “An overview on regression models for discrete longitudinal responses,” *Statistical Science*, vol. 18, no. 3, pp. 377–393, 2003. [117](#)
- [55] E. McKenzie, “Some arma models for dependent sequences of poisson counts,” *Advances in Applied Probability*, vol. 20, no. 4, pp. 822–835, 1988. [117](#)
- [56] R. C. Jung and A. Tremayne, “Binomial thinning models for integer time series,” *Statistical Modelling*, vol. 6, no. 2, pp. 81–96, 2006. [117](#)

Appendix

Simulation Designs I. and II. - means, variance, and correlations for conditionally serially correlated AR(1) DGPs for stationary and non-stationary correlation structures

For a stationary correlation structure, the properties, following [54], [47], are:

Mean: $E[Y_{ij}] = \check{\mu}_i$

Variance: $\text{Var}[Y_{ij}] = \check{\mu}_i$.

(Conditional) Correlation: $\text{Corr}[Y_{i(j-1)}, Y_{ij} | \nu_i] = \rho^\ell$, where $\ell = 1, \dots, j-1$ corresponds to a unit time-period lag.

For a non-stationary correlation structure, the following properties hold:

Mean: $E[Y_{ij}] = \mu_{ij}$

Variance: $\text{Var}[Y_{ij}] = \mu_{ij}$.

(Conditional) Correlation: $\text{Corr}[Y_{ih}, Y_{ij} | \nu_i] = \rho^{j-h} \left(\frac{\mu_{ih}}{\mu_{ij}} \right)^{\frac{1}{2}}$ (where $h < j$, and $j > 2, \dots$)

As obviated by the forms, note that the statistical parameters for a stationary, AR(1) correlation structure are completely time independent and its correlation simply depends upon a prior lag ℓ . In the case of a non-stationary AR(1) structure, clearly there is time dependence for the mean, variance, and correlation.

Binomial Thinning Process

Following [55], [56], we briefly highlight the binomial thinning process. This is a method employed to ensure the integer discreteness of the process of first-ordered serially autocorrelated Poisson counts. The thinning methods employed are a stochastic operations performed independently of each other with a given probability distribution. Note, also, that $y_{i(j-1)}$ and is independent of q_{ij} .

The term $\rho \circ (y_{i(j-1)})$ denotes a “binomial-thinning” parameter. In essence, $\rho \circ (y_{i(j-1)}) = \sum_{l=1}^{y_{i(j-1)}} u_l(\rho)$ where $u_l(\rho)$ is a ρ -dependent, iid sequence of $\{0, 1\}$ Bernoulli random variables with $P(u_l(\rho) = 1) = \rho$. Thus, $u_l(\rho) \sim \text{Bern}(\rho)$ with $\rho \in [0, 1)$.

The sum of iid Bernoulli random variables is the Binomial distribution which easily follows from employing the probability generation function of a Bernoulli random variable.

Let $X \sim \text{Bern}(\rho)$. Thus, the probability generating function of X is

$$\Pi_X(s) = (1 - \rho) + \rho s \quad 0 \leq s \leq 1$$

where s is a dummy placeholder and $X = 0$ with probability $1 - \rho$ and $X = 1$ with probability ρ . The probability generating function for the sum of n trials is

$$(\Pi_X(s))^n = ((1 - \rho) + \rho s)^n = \sum_{x=0}^n \binom{n}{x} (\rho s)^x (1 - \rho)^{n-x}$$

resulting in the probability generating function for the Binomial distribution (thus, the sum is Binomial).

Simulation Designs I. and II. (Unbalanced Cases)

Here are our cases that corresponds to the unbalanced designs where $25 \leq n_i \leq 40$ with $R = 500$. The results seem to mimic the balanced cases in all instances, particularly for the fixed effects and variance parameter biases and MSEs. However, it is noticeable that the polynomial AMSEs are highest for the stationary estimated smooth functions for the polynomial for the correctly specified case for Simulation Design I.

Table 3.10: Simulation Results I - Unbalanced Stationary Panel

Case 1:					Case 2:					Case 3:				
Parameter Values	Polynomial		Linear		Parameter Values	Polynomial		Linear		Parameter Values	Polynomial		Linear	
	Bias	MSE	Bias	MSE		Bias	MSE	Bias	MSE		Bias	MSE	Bias	MSE
$\beta_1 = 1$	0.01581	0.01161	0.04939	0.00944	$\beta_1 = 1$	-0.03504	0.02844	-0.00489	0.00221	$\beta_1 = 1$	0.07226	0.01143	0.05148	0.00646
$\sigma_v^2 = 0.50$	0.04328	0.01580	0.02214	0.01781	$\sigma_v^2 = 0$	.084439	.067378	-0.00517	0.00056	$\sigma_v^2 = 0.50$	0.05213	0.02218	0.00106	0.00002
$\rho = 0.50$	0.00116	0.00002	-0.05216	0.00302	$\rho = 0.50$	-0.03146	0.00104	0.006630	0.00041	$\rho = 0$	-0.01042	0.00211	-0.02801	0.00623

Table 3.11: Unbalanced Stationary Estimated Smooth Functions - Performance (Case 1)

Factor Smooth: Level (0), (x2)x1:(0)		Factor Smooth - Level (1), (x2)x1:(1)	
Average Mean Squared Error (AMSE _f)		Average Mean Squared Error (AMSE _f)	
Polynomial	0.07163	0.09663	
Linear	0.040867	0.039652	

Table 3.12: Simulation Results II - Unbalanced Non-Stationary Panel

Case 1:					Case 2:					Case 3:				
Parameter Values	Polynomial		Linear		Parameter Values	Polynomial		Linear		Parameter Values	Polynomial		Linear	
	Bias	MSE	Bias	MSE		Bias	MSE	Bias	MSE		Bias	MSE	Bias	MSE
$\beta_1 = 1$	-0.00198	0.00555	0.03525	0.01629	$\beta_1 = 1$	0.0424790	0.0094501	0.09960	0.01122	$\beta_1 = 1$	0.00019	0.00275	0.01329	0.00322
$\sigma_v^2 = 0.25$	0.00505	0.00032	0.03691	0.00201	$\sigma_v^2 = 0$	.00017e-10	0.0030e-20	0.00037	0.00201	$\sigma_v^2 = 0.25$	-0.02107	0.00144	-0.01005	0.00130
$\rho = 0.25$	0.04020	0.00168	0.02952	0.000989274	$\rho = 0.25$	0.02523171	0.00075018	0.02994	0.00102	$\rho = 0$	-0.01345	0.00021	-0.01143	0.00031

Table 3.13: Unbalanced Non-Stationary Estimated Smooth Functions - Performance (Case 1)

Factor Smooth: Level (0), (x2)x1:(0)		Factor Smooth - Level (1), (x2)x1:(1)	
Average Mean Squared Error (AMSE _f)		Average Mean Squared Error (AMSE _f)	
Polynomial	0.034214	0.043898	
Linear	0.02675854	0.0303986	

CHAPTER 4

Forecasting Out-of-Sample Inflation Gaps: A Penalized Quantile Combination Approach

Abstract

The idea is to utilize an existing penalization procedure to minimize the effects of weak forecasting models on the out-of-sample prediction capability of monthly annualized inflation gaps across multiple horizons framed in a forecasting combination problem. As the optimal combination forecast solution of the Granger-Ramanathan regression form under quadratic loss is the conditional mean, we seek to extend the literature using an approximation to the conditional mean by additively combining of conditional quantiles that discards underlying, useless forecasters via an L1-penalization quantile forecast combination procedure at various, equally-spaced deciles of our monthly annualized inflation gap. The proposed L1PQFC forecast combination solution based on this soundly outperforms an equal weight benchmark model as well as other existing combination models on a variety of performance metrics.

4.1 Introduction

Over the last half-century, the methodology of combination forecasting has been shown to be an indispensable method for enhancing forecasting accuracy. Combination forecasting can be defined as the resulting combination of multiple, individual forecasts from a selection of forecasts, which are then unified to produce a single, pooled forecast ([1]). The combination methodology has been utilized in various research domains including, but not limited to logistics ([2], [3]), power systems ([4]), statistics, management science, and psychology ([5]). Unanimously considered in the literature as the genesis of the theory development and the practical utility of combination forecasting, [6] first posited that a linear combination of multiple forecasts can yield a greater degree of forecast accuracy when independent and relevant information is taken into account. The principal method presented in the seminal work suggests computing multiplicative linear constants, or linear “weights”, such that the error variance of the final combination is minimized under the assumption of unbiased, individual forecasts. The vital conclusion is that a composite set of independent forecasts can yield lower mean square error (MSE) than a single forecast. Another invaluable benefit from a linear forecasting combination is that the combined solution utilizing different underlying forecasting models is more robust or can be relatively unaffected by the misspecification biases that may plague the data generating process from which the individual forecasts were constructed ([7]).

As the practical utility of combination forecasting has been illuminated, the theory related to this technique is both well-explored and ever-evolving in the literature with the linear formulation being a recurring focal point. Established methods inspired by the linear weight solution proposed by [6] include the simple average method, the outperformance method ([8]), the discounted mean squared error (MSE) method ([9]), the shrinkage weight method ([10]), the time-varying parameter weighted method ([11]), and various adaptive extensions¹.

Additionally, beyond utilizing a forecasting combination to predict a traditional univariate time-series response or forecasting target, a variety of forecasting combination schemes have been studied and outlined for conditional variances, conditional densities, interval forecasts, and discrete random variables ([13]). A recent work presents a thorough overview of theoretical results of the linear forecasting combination problem under two closely related goals of the forecasting combination methodology: combination for adaptation (given a set of individual forecasts, the final combination is chosen such that it equivalently performs as well as the singular best forecasting individual model) and combination for improvement (given a set of individual forecasts, the constrained linear combination of all forecasts is chosen) under a variety of theoretical scenarios ([13]). This literature shows their algorithm, AFTER, is an effective tool when the primary objective is combination for adaptation. Additionally, [14] provides a brief, yet comprehensive overview of well-founded combination forecasting techniques derived from a standard, linear forecasting form including equal weight

¹One proposed extension is an “independent” assumption of the associated covariance matrix associated with the constructed linear weight such that the covariance matrix is a pure diagonal matrix of individual forecast error variances. Another proposed extension bounds the linear weights to lie within the interval $[0,1]$ ([12]).

forecasting combination, optimal minimum variance forecasting combination, outperformance probabilities, as well as forecasting combinations from both unconstrained and constrained ordinary least squares (OLS).

Certain avenues have explored, with positive combination results for inflation, forecasting combinations from univariate, dynamic factor, and an amalgam of VAR and Bayesian VAR models², which mostly outperform their univariate counterparts based on lower forecast error ([17]). Another work uses performance-based forecast comparative combination methods for European inflation to hedge against univariate forecasting models that perform poorly in times of prevalent volatility ([18]). Another unique approach explores a model selection method for individual forecasting models used in a forecasting combination where the five “best” models are chosen through a sequential cointegration/encompassing procedure where forecast accuracy is increased ([19]). This last work leads into the criticisms surrounding the forecasting combination problem. Despite the various methods proposed for estimating weights are vast and numerous, there is no agreeable, single best technique that is guaranteed to consistently produce uniformly optimal accuracy. More severely, although a combination solution can be considered generally robust, the inclusion of estimation errors that pervade combination weights diminish forecasting performance. The likely cause of such performance distortion can be attributed to the inclusion of underlying “weak” forecasts that have virtually no relevance in forecasting any response of interest.

Arguably the most beneficial and lauded extensions of [6] beyond those cited previously is [20]³. These authors explored an ordinary least squares solution for the estimation of weights. Under the quadratic (MSE) loss function and utilizing individual forecasting models as regressors in a standard ordinary least square regression in addition to a constant, leads to an optimal forecast solution, which does not require that the unbiasedness of the forecasts⁴. Thus, under the MSE loss function, the optimal forecast solution is the conditional mean. This popularized “linear-in-weights” combination form in conjunction with the notion that the conditional mean under the MSE loss function can be approximated by the combination of conditional quantiles ([21],[22], [23],[24]) leads to an uninvestigated avenue of potential increased out-of-sample (OOS) forecasting performance involving a penalized approach for discarding models that can contaminate linear weights. The focus of this chapter is to employ a technique for accurate model discrimination in a linear combination solution that both discards all underlying forecasting models that do not contribute positively to out-of-sample (OOS) forecasting gains while retaining such models that only partially contributes to OOS forecasting gains at certain “segments” of a forecasting target. To our knowledge, this is a completely unexplored approach to extricate weak forecasting models’ ability to contaminate accuracy. Our central focus is the development of such a criterion, or penalty, that is dually capable of nullifying underlying forecasting models in a linear, composite combination model that contributes null to OOS forecasting gains while retaining underlying forecasting models that, at certain quantiles of our target, contributes positively

²A VAR (vector autoregressive) model is simply a direct, straightforward extension of standard univariate autoregressive models which demonstrates how values of variables at time t depends linearly upon previous time points such as $t - 1$. The Bayesian VAR simply refers to estimation of standard VAR parameters where prior probabilities are assigned to the parameters to be estimated wherein the parameters themselves are considered random variables ([15], [16]).

³As the Granger-Ramanathan form is germane to our analysis, we provide a brief, yet explicit, overview in Chapter 1.

⁴Unbiasedness of the individual forecasts are not required as any bias can be adjusted through the intercept.

to OOS forecasting gains. This precisely corresponds to the novelty to the forecasting literature reflected in this chapter. As the optimal forecasting combination solution under MSE loss can be approximated by the aggregation of conditional quantiles from [20], we utilize a penalization procedure via a LASSO⁵ quantile regression method that discriminates useless or partially useless forecasting models that aid in marginal or no utility in forecasting our target of interest. We deem this solution to be the “L1-penalized” quantile forecast combination L1PQFC.

This proposed method utilizes a penalty to classify the presence of “weak”, “partially weak”, and “strong” underlying forecasting models. Weak forecasting models are defined to be the underlying forecasters in the combination solution that predict no quantiles of our response at the forecast horizon $t + h$. Partially weak forecasters represent underlying forecasters that predict some, but not all quantiles of the forecasting target at the forecasting horizon $t + h$; strong forecasters are those such forecasters that help forecast all quantiles of the target. As we take particular interest in conditional quantiles in our combination solution due to its protection against estimation errors relative to the competitive standard ordinary least squares (OLS) estimator’s solution, we first select statistically significant forecasting models at the 5% level corresponding to the quantile functions, and then employ penalized quantile regressions on the selected forecasting models which results in L1-penalized quantile estimators. We combine these L1-penalized quantiles to obtain a point forecast of the target of interest resulting in the L1PQFC solution.

Thus, the resulting forecasting classification of weak, partially weak, and strong forecasters in the L1PQFC solution are chosen via the selection procedure for the using the penalized quantile regression technique proposed by [26] for quick and efficient dimensionality reduction in a high-dimensional, sparse modeling context. This has been applied successfully towards forecasting a one-period ahead forecast of equity premium by [24] as a variable selection tool. Building from this framework, we emphasize our novel, primary contribution highlighting that while the combination of quantile forecasts will lead to a prediction equation that has a conditional mean interpretation as in [20], the coefficients of such prediction equation now reflects the presence of strong, partially weak and weak forecasting models.

Following [27], we apply our proposed forecasting approach to monthly annualized inflation gaps. The choice of inflation gaps is due to the well-known, omnipresent trend component that can be observed by measuring inflation over time (particularly spiking during The Great Inflation period). This trend is “slow-varying” and controls for a low-frequency component central to inflation measurement in recent past, and prior literature recommends accounting for this phenomenon, particularly for forecasting inflation in the long run. Therefore, our target of interest represents the deviation of the actual monthly annualized inflation rate and its corresponding trend. The breadth of this application essentially involves a two-part, composite forecasting analysis that expands on previous literature by contributing to the forecasting combination problem with model selection via the LASSO (L1) penalization procedure. We use this penalty as a powerful, intuitive model selection procedure for our combination solution. First, we utilize ten linear, univariate

⁵LASSO is an acronym representing “Least Absolute Shrinkage and Selection operator” re-popularized by [25] in a traditional regression context.

forecasting models to construct out-of-sample forecasts of the monthly annualized inflation gap across six forecasting horizons: $h = \{1, 2, 3, 4, 8, 12\}$. Next, we treat these forecasts, per each horizon, as covariates in the linear Granger-Ramanathan form for a forecasting combination model.

We show that the proposed L1PQFC method results in superior out-of-sample (OOS) forecast accuracy over an equal weight combination benchmark as well as other competing combination considerations. Additionally, we test and verify the robustness and success of our proposed L1PQFC models over our six horizons $h = \{1, 2, 3, 4, 8, 12\}$. The empirical results show that, at varying horizons h , the model selections ultimately utilized in the L1PQFC considerations vary considerably across three notable levels: 1.) the underlying chosen models via the penalty at a specified quantile, 2.) the frequency of models selected via the penalty at a specified quantile, and 3.) the time at which a model is chosen for the span of 48 months chosen in our out-of-sample data per each considered horizon h . We measure the success of OOS forecast accuracy by the R_{OS}^2 statistic via statistical significance of the Clark and West and Diebold-Mariano tests for all our horizons; additionally, we also show a decomposition of the mean squared prediction error (MSPE) into two additive terms in order to assess the contribution of each step of the L1PQFC approach for our combination solution following [24]. Our results indicate that the additional loss is overwhelmingly attributable to the efficiency loss from the presence of partially weak models in the population.

The remainder of this paper is structured as follows: Section 4.2 briefly reviews the methodology of forecast combination problem as well as prior works' results regarding the optimal forecast weights under both a general and MSE loss functions as well as the data generating process (DGP) coinciding with our forecasting combination problem. Section 4.3 presents the penalization procedure, the methodology regarding the competing forecast combinations models, and our weighting scheme for LASSO-chosen quantile forecasts. Section 4.4 presents the univariate models used to compute the forecasts used in quantile combination considerations in addition to our principal evaluative metric. Section 4.5 presents all combination results. Section 4.6 concludes our analyses.

4.2 The Forecast Combination Problem

Following previous literature ([28], [29]), we review the forecasting combination problem. We begin by outlining [30] who uses a traditional diversification problem as a preamble. Suppose, for instance, we want to forecast a response variable y with two available and separate point forecasts, $\hat{y}_1$ and $\hat{y}_2$. Let g_1 and g_2 be two arbitrary functions that takes inputs to create point forecasts of interest. The first forecast is estimated from two independent explanatory covariates, x_1 and x_2 , as $\hat{y}_1 = g_1(x_1, x_2)$, whereas the second forecast is produced with three other independent covariates as $\hat{y}_2 = g_2(x_3, x_4, x_5)$. Assuming all covariates are not perfectly collinear and are all observed, a natural step would be to produce a forecast of all available covariates $\hat{y}_3 = g_3(x_1, x_2, x_3, x_4, x_5)$. For a combination consideration, let g_c indicate a function that combines two forecasts. If only the point forecasts themselves are available ($\{\hat{y}_1, \hat{y}_2\}$), and the underlying

covariates are obfuscated from a practitioner (perhaps due to input privatization), then two, non-exhaustive options for a forecasting combination model would be a combination solution of form $\hat{y} = g_c(\hat{y}_1, \hat{y}_2)$ or even the estimation the quantile function $\hat{y} = q_y(\tau) = g_c(\hat{y}_1, \hat{y}_2)$. The forecaster user's information set may comprise k individual forecasts, $\Omega = (\hat{y}_1, \dots, \hat{y}_k)$, where Ω is often a much smaller subset of the union of the information sets underlying the each of the individual forecasts, $\bigcup_{i=1}^k \Omega_i$ ([7]).

Further expounding on the overall perspectives of forecasting combinations methods stated in the introduction, we highlight advantages and disadvantages of defined and outlined from previous literature. Point forecasts may be very differently affected by “structural breaks” caused, for example, by technological developments. Some forecasting models can quickly adapt may and will only temporarily be affected by breaks in the structure of a time series, while others have parameters that only adjust marginally to new data proceeding a structural break ([31], [11], [32]). Therefore, a combination of forecasts from models of varying adaptability from structural breaks can yield the best forecasting performance relative to any individual models. Perhaps the most important utility offered by the forecast combination problem is the relative “protection” offered by a combination's solution comprised from point forecasts that are produced by models that are inherently misspecified due to unknown reasons ([1], [7], [33]). Traditionally, as data generating processes are likely much more complex for a univariate, stationary series, it is likely that this phenomenon would be beyond the scope of econometric models available in a forecaster's arsenal. Thus, the forecast combination procedure, as a method, produces a solution such that the final combination is essentially impervious to misspecification biases and measurement errors. Lastly, the point forecasts that comprise a vector of forecasts used for a final combination may employ varying loss functions given the same information set, which can allow for the simultaneity of greater forecast accuracy and less significant forecast errors that can contaminate the optimal weights for the final combination. Regarding forecasting combination disadvantages, we emphasize the potential for overfitting when estimating forecasts in a regression scenario where the sample size is small relative to the number of coefficients to be estimated given our training set ([10], [28]). Additionally, if a forecast practitioner does actually observe the union of information sets underlying each individual forecasts, $\bigcup_{i=1}^k \Omega_i$, then prior literature ([34], [10]) have suggested the notion of “combining information” as opposed to “combining forecasts” (e.g. a “super” model construction) which nests all of this information into one forecasting model rather than the combination strategy in hopes that this aggregation tactic could lead to decreased forecast errors at the expense of increased forecast bias. Lastly, the forecasting combination procedure is disadvantaged relative to any single forecast model as it can introduce parameter estimation errors in instances in which combination weights must be estimated. This is such a foundational issue that simple combination schemes, such as the simple averaging scheme (equal weights), have been notably superior relative to “optimal” combination schemes in the absence of parameter estimation ([32]).

For our application, we are interested in forecasting the future value of a stationary, univariate time series $\{y_{t+h}\}_{h=1}^{\infty}$ based on a k -vector of point forecast of this variable $\hat{\mathbf{y}}_{t+h,t}$ which possibly includes a constant

determined ex-ante. This is also adapted into the information set Ω_t (whereas y_{t+h} is not, prohibiting cases in which there exists a $\hat{y}_{t+h,t}^i$ such that $\hat{y}_{t+h,t}^i = y_{t+h}$ [perfect prediction] and where i indexes an individual forecaster). The goal is to seek an aggregator that transforms the information in $\hat{\mathbf{y}}_{t+h,t} \in \mathbb{R}^k$ to a summary measure $C(\hat{\mathbf{y}}_{t+h,t}, \boldsymbol{\gamma}_i) \in \mathbb{R}^i \subset \mathbb{R}^k$. The term $\boldsymbol{\gamma}_i$ is a parameter vector corresponding to the combination and depends on both a loss function and an unknown forecast error distribution. Additionally, we define $F_{t+h,t}$ to be the conditional distribution of y_{t+h} , given the information set Ω_t . As ubiquitous as the combination forecasting problem has been in the literature, it is imperative to note that the point forecasts themselves may generally provide insufficient information for a decision maker who may be interested in the degree of uncertainty surrounding the forecast. For our application, however, point forecasts will suffice.

We have the conditional model used in our forecasting combination problem following the general form presented as

$$y_{t+h} = \hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\beta} + (\hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\phi}) \eta_{t+h} \quad (4.1)$$

$$\eta_{t+h} | \Omega_t \sim F_{\eta,h}(0, 1)$$

where $F_{\eta,h}(0, 1)$ is a defined distribution function with zero mean and a variance of unity, which does not depend on Ω_t , but depends on h , $\hat{\mathbf{y}}_{t+h,t} \in \Omega_t$ is a $k \times 1$ vector of point forecasts/models for an arbitrary forecast horizon h observable at time t ⁶, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are $k \times 1$ parameter vectors including the intercepts: β_0 and ϕ_0 . The location-scale model form implies the conditional mean and conditional quantile functions will be affected by the vector of point forecasts $\hat{\mathbf{y}}_{t+h,t}$ (as the point forecasts affect the location and scale of the conditional distribution of y_{t+h}). Specifically, this conditional model form is very convenient as it can naturally be investigated through the lens of a quantile regression framework with linear conditional quantiles, which is precisely a benefit we will exploit in our application ([29]). Notably, no parametric structured is assumed for $F_{\eta,h}$ which is indicative of a general loss function for the error process. This model form subsumes an extensive class of models including ARCH as well as stochastic volatility models such as SABR, GARCH, and Heston models.

We briefly review optimal weight estimation results as it relates to combination forecast methodology. Following seminal literature ([20], [35] [36], [29]), we begin by outlining a necessary assumption regarding the loss function. An optimal forecast solution $\hat{y}_{t+h}^i$ is the minimizer of the expected value of a general loss function L^i . Thus, we begin with our first primary assumption.

⁶Note, as our vector of predictors in our application will correspond to point forecasts from univariate models across specified forecast horizons, we use the terms “point forecasts” and “models” synonymously.

Assumption 1 (Loss Function): The loss function, L^i , $i \in (0, 1)$ is a homogeneous function solely of the forecast error $e_{t+h} \equiv y_{t+h} - \hat{y}_{t+h,t}$. That is, $L^i = L^i(e_{t+h,t})$ and $L(ae) = g(a)L(e)$, $a \neq 0$, for some positive function g ⁷.

The superscript i associated with the loss function L is indicative of a stream of loss functions that satisfies Assumption 1.⁸

Proposition 1. Under the DGP in (4.1) and Assumption 1, the optimal forecast combination will be

$$\hat{y}_{t+h,t}^i = \hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\beta} + (\hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\phi}) \phi_h^i = \gamma_{i,0} + \gamma_{i,1} \hat{y}_{t+h,t}^1 + \dots + \gamma_{i,k-1} \hat{y}_{t+h,t}^{k-1} \quad (4.2)$$

where $\gamma_{i,0} = (\beta_0 + \phi_0 \phi_h^i)$, $\gamma_{i,1} = (\beta_1 + \phi_1 \phi_h^i)$, $\gamma_{i,k-1} = (\beta_{k-1} + \phi_{k-1} \phi_h^i)$, where ϕ_h^i is a constant that is dependent upon the forecast horizon h , distribution function $F_{n,h}(0, 1)$ and the loss function L^i . The proof for Proposition 1 can be found in the appendix.

From Proposition 1, an essential corollary is immediate:

Corollary 1. Under (4.1), the optimal forecast combination under the quadratic loss function (MSE) is

$$\hat{y}_{t+h,t}^i = E[y_{t+h} | \Omega_t] = \hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\beta} = \beta_0 + \beta_1 \hat{y}_{t+h,t}^1 + \dots + \beta_{k-1} \hat{y}_{t+h,t}^{k-1} \quad (4.3)$$

where $E[y_{t+h} | \Omega_t]$ is the conditional mean of y_{t+h} . Here, $\phi_h^i = 0$ as $E[\eta_{t+h} | \Omega_t] = 0$. Thus, the optimal weights are fixed at $\gamma_{i,j} = (\beta_j + E[\phi_j * \phi_h^i]) = \beta_j$, $j = 0, \dots, k-1$.⁹

This solution is the least squares estimator for the conditional mean of y_{t+h} from the Granger-Ramanathan combination method [20].

4.3 Estimation

Under MSE Loss, Corollary 1 shows that the optimal combination solution corresponds directly to the conditional mean of the response y_{t+h} . The sample analogs of the weights presented under Corollary 1 is the least squares estimator on a constant and a vector of forecasts ([20]). From Proposition 1 and utilizing the conditional model in (4.1), utilizing a general loss function, we clearly see that the optimal forecast solution differs from the solution proposed by [20] under MSE loss. Utilizing the conditional model in (4.1), with dynamics for both the mean and variance, and invoking the fact that the conditional distribution of a random variable given a k -vector of covariates/forecasts (with available data from their joint distribution) can be equivalently characterized by either the conditional distribution function or the conditional quantile function ([37],[38]), we can utilize a method following [29] for estimating optimal combination forecasting

⁷Homogeneity of a loss function is a crucial property for an optimal forecast solution since the optimal forecast will itself be invariant for different values of $g(a)$.

⁸There are many loss functions that satisfies Assumption 1 such as MSE, EKT, MAE, lin-lin, and a host of other asymmetric quadratic losses.

⁹Note, we omit the solution of the optimal weights derived under the complete lack of scale effects $\phi_1 = \dots = \phi_{k-1} = 0$ under the given conditional model. The results can be found in ([29], [14]).

weights that is appropriate under a homogeneous loss function. Given the optimal forecast solution in (4.2) and, then the conditional distribution of y_{t+h} can be employed as

$$\begin{aligned}
F_{t+h,t}(\hat{y}_{t+h,t}^i) &= P(y_{t+h} < \hat{y}_{t+h,t}^i | \Omega_t) \\
&= P\left(\hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\beta} + (\hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\phi})\eta_{t+h} < \hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\beta} + (\hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\phi})\phi_h^i | \Omega_t\right) \\
&= P\left((\hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\phi})\eta_{t+h} < (\hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\phi})\phi_h^i | \Omega_t\right) \\
&= P(\eta_{t+h} < \phi_h^i | \Omega_t) = F_{\eta,h}(\phi_h^i) \\
&= \tau_i \in (0, 1)
\end{aligned}$$

Clearly, $\hat{y}_{t+h,t}^i = F_{t+h,t}^{-1}(\tau_i)$, which means that the optimal forecast combination $\hat{y}_{t+h,t}^i$ corresponds to the conditional quantile function of y_{t+h} given level τ_i , as

$$\hat{y}_{t+h,t}^i = Q_{y_{t+h}}(\tau_i | \Omega_t) = \gamma_{i,0} + \gamma_{i,1}\hat{y}_{t+h,t}^1 + \dots + \gamma_{i,k-1}\hat{y}_{t+h,t}^{k-1} \quad (4.4)$$

for some $\tau_i \in (0, 1)$

where $Q_{y_{t+h}}(\tau_i | \Omega_t)$ is the conditional quantile of y_{t+h} at level τ_i and weights are estimated as in (4.2), explicitly represented as $\gamma_{i,j} = (\beta_j + \phi_j F_{\eta,h}^{-1}(\tau_i))$, $j = 0, \dots, k-1$. Note, $\phi_h^i = F_{\eta,h}^{-1}(\tau_i)$ is the unconditional quantile corresponding to a future horizon $t+h$ of the error process, η_{t+h} , for the loss function L^i .

We can employ a quantile regression of y_{t+h} on a constant and a vector of forecasts to find the optimal weights under the general loss function L^i and $F_{\eta,h}(0, 1)$.¹⁰

We invoke Assumption 1 under a singular, asymmetric absolute loss function, ρ_τ ¹¹, corresponding to conditional quantile function (necessarily dropping the previous subscript i). For convenience, given our vector of point forecasts representing our known information set, $\hat{\mathbf{y}}_{t+h,t}$, we re-write equation (4.4) as

$$Q_{y_{t+h}}(\tau | \hat{\mathbf{y}}_{t+h,t}) = \beta_0(\tau) + \hat{\mathbf{y}}_{t+h,t}'\boldsymbol{\beta}_1(\tau) \quad (4.5)$$

with $\beta_0(\tau) = \beta_0 + \phi_0 F_{\eta,h}^{-1}(\tau)$, $\boldsymbol{\beta}_1(\tau) = \boldsymbol{\beta} + \boldsymbol{\phi} F_{\eta,h}^{-1}(\tau)$ ¹², with an associated vector of size $(k-1) \times 1$ point forecasts $\hat{\mathbf{y}}_{t+h,t} = (\hat{y}_{t+h,t}^1, \dots, \hat{y}_{t+h,t}^{k-1})'$ utilized as explanatory covariates in the linear combination for the proposed combination solution. Given this form, we highlight the result of combination forecast optimality

¹⁰If $F_{\eta,h}(\phi_h^i) = \tau_i$, then it's clear that $\phi_h^i = F_{\eta,h}^{-1}(\tau_i)$.

¹¹This asymmetric absolute loss function, ρ_τ , is explicitly defined as $\rho_\tau = [\tau - \mathbf{1}(u < 0)]u$. This is a continuous, piecewise linear function that is non-differentiable at $u = 0$. In this case, u is tantamount to a "regression error", and ρ_τ can be thought of as the loss associated with the error.

¹²Clearly, $\boldsymbol{\beta} \equiv (\beta_1, \dots, \beta_{k-1})$ and $\boldsymbol{\phi} \equiv (\phi_1, \dots, \phi_{k-1})$.

of a forecasting target partitioned into two individual, additive components: the conditional mean and a measure of bias.

Proposition 2. Under Equation (4.5) utilizing the singular, asymmetric loss function, the optimal forecast combination solution is:

$$\begin{aligned}\hat{y}_{t+h} &= Q_{y_{t+h}}(\tau|\hat{\mathbf{y}}_{t+h,t}) \\ &= E(y_{t+h}|\hat{\mathbf{y}}_{t+h,t}) + \theta_\tau\end{aligned}$$

where $\theta_\tau = \hat{\mathbf{y}}'_{t+h,t} \phi F_{\eta,h}^{-1}(\tau)$ is an additional bias relative to the conditional mean function that is dependent upon the associated vector of point forecasts associated with some arbitrary forecast horizon, $\hat{\mathbf{y}}_{t+h,t}$, the distribution function $F_{\eta,h}$ and the loss function L .

We use the approach for obtaining robust mean squared prediction error (MSPE) forecasts of monthly annualized inflation gaps following [24]. That is, in the presence of outlying observations, an approach to obtaining MSPE forecasts of our response is through combining quantile forecasts. Let $\boldsymbol{\tau} = (\tau_1, \dots, \tau_m)$ be a set of strictly increasing quantiles. Therefore, we have:

$$\begin{aligned}\sum_{\tau=\tau_1}^{\tau_m} w_\tau Q_{y_{t+h}}(\tau|\hat{\mathbf{y}}_{t+h,t}) &= E(y_{t+h}|\hat{\mathbf{y}}_{t+h,t}) + \sum_{\tau=\tau_1}^{\tau_m} w_\tau \theta_\tau \\ &= E(y_{t+h}|\hat{\mathbf{y}}_{t+h,t}) + \hat{\mathbf{y}}'_{t+h,t} \phi \sum_{\tau=\tau_1}^{\tau_m} w_\tau F_{\eta,h}^{-1}(\tau)\end{aligned}$$

with w_τ being the weight attributed to the conditional quantile $Q_{y_{t+h}}(\tau|\hat{\mathbf{y}}_{t+h,t})$. The weights themselves are quantile dependent as their purpose is directed at estimating/approximating the mean of η_{t+h} which, by assumption, is 0. A notable result for integrating the quantile function over its entire domain in a univariate setting results in the sample mean ([39],[24]). Given that the error η_{t+h} is independently and identically distributed, with $F_{\eta,h}^{-1}(\tau) = Q_\tau(\eta_{t+h})$, immediately implies $E(\eta_{t+h}) = \int_0^1 F_{\eta,h}^{-1}(t)dt = 0$. We can utilize this result and apply it to a sample for conditional mean approximation utilizing a set of arbitrary quantiles. With a quantile grid (e.g. $\boldsymbol{\tau} = \tau_1, \dots, \tau_m$) in a sample set of data, we can approximate $E(\eta_{t+h}) = \int_0^1 F_{\eta,h}^{-1}(t)dt$ by $\sum_{\tau=\tau_1}^{\tau_m} w_\tau F_{\eta,h}^{-1}(\tau)$. The assigned weight w_τ reflects the heavy tails/kurtosis or potential asymmetry of the conditional distribution function of η_{t+h} , which is $F_{\eta,h}$. When $F_{\eta,h}$ is symmetric, equal weight assignment to the grid quantiles in the vicinity of the median ($\tau = 0.5$) guarantees $\sum_{\tau=\tau_1}^{\tau_m} w_\tau Q_{y_{t+h}}(\tau|\hat{\mathbf{y}}_{t+h,t}) = E(y_{t+h}|\hat{\mathbf{y}}_{t+h,t})$ ¹³. By utilizing the quantile regression estimator, this approach is essentially impervious to extreme observations of a response. As the lower and higher ends of the quantiles

¹³The brief, theoretical justification of the approximation of the combination of quantiles to the conditional mean (which is our optimal forecast combination solution) (e.g. [21],[40]) follows directly from [24] and can be applied indiscriminately with no additional justification for the extension to a forecasting combination solution.

specified in our quantile grid produce downwardly and upwardly, respectively, biases of the conditional mean of a response, the combination approach results in the point forecast $\sum_{\tau=\tau_1}^{\tau_m} w_{\tau} Q_{y_{t+h}}(\tau | \hat{\mathbf{y}}_{t+h,t})$ and combines opposing biased predictions which nullify each other and remove the aggregation bias ([41]).

As the conditional model is unknown (4.1), the forecasting model based on combining conditional quantiles with fixed (standard) point forecasts, may potentially be misspecified, particularly when ineffective or weak. The penalization procedure that follows will address the problem of eliminating its adverse effects in our combination of conditional quantiles to yield an optimal forecasting combination solution.

4.3.1 The L1-Penalized Quantile Forecast Combination Method

Following [26], identifiability of weak forecasting models by using a lasso-based penalty to the model weights is of principle interest. Utilizing equation (4.5), the resulting LASSO quantile regression estimator solves the following objective function

$$\min_{\beta_0, \beta_1} \sum_t \rho_{\tau}(y_{t+h} - \beta_0(\tau) - \hat{\mathbf{y}}'_{t+h,t} \beta_1(\tau)) + \lambda \frac{\sqrt{\tau(1-\tau)}}{N} \|\beta_1(\tau)\|_{L1} \quad (4.6)$$

where ρ_{τ} is the asymmetric absolute deviation function (or “check” function) defined for any scalar a as $\rho_{\tau}(a) \equiv (\tau - \mathbb{1}(a \leq 0))a$ with $\mathbb{1}(\cdot)$ as the indicator function, N is the estimation (sample) size, $\|\cdot\|_{L1}$ denotes the L1-norm, $\|\beta_1\|_{L1} = \sum_{i=1}^{k-1} |\beta_{1i}|$, and $\mathbf{y}_{t+h,t} = (\hat{y}_{t+h|t}^1, \dots, \hat{y}_{t+h|t}^{k-1})'$ ¹⁴. Therefore, the full objective function in (4.6) is the sum of a quantile regression estimator and a scaled penalty of $\beta_1(\tau)$ based on the L1-norm; The level of penalization, $\lambda \frac{\sqrt{\tau(1-\tau)}}{N}$, depends on each quantile τ where λ on the set of specified quantile indices of interest $\boldsymbol{\tau}$. Details can be found in the appendix. The L1-norm is a well-studied regularization technique that performs “feature selection” that shrinks the weights associated with the least important or weak individual forecasts ([42]). Following [26] and [43], this LASSO method will first select the individual forecasting models from the information in $\hat{\mathbf{y}}_{t+h,t}$ based on the conditional models for each specified quantile at each time period t .

Next, we employ a quantile regression post-penalty generation for the selected point forecasts in $\mathbf{y}'_{t+h,t}$ for the L1-penalized quantile forecast of the response at $t+h$ defined as $q_{t+h,t}^{\tau} = \beta_0(\tau) + \mathbf{y}_{t+h|t}'^* \beta(\tau)$. The matrix $\mathbf{y}_{t+h,t}'^*$ now corresponds to the L1-penalized quantile forecasts (L1PQF) forecast at time t for various horizons of h by the LASSO selection at the $\alpha = 5\%$ level of significance. This procedure is repeated to obtain L1PQFs for the discrete quantiles of $\boldsymbol{\tau} = (\tau_1, \dots, \tau_m)$. Finally, we add these quantile forecasts after LASSO selection to obtain the L1 penalized quantile forecast combination (L1PQFC) $\hat{y}_{t+h} = \sum_{j=1}^m w_{\tau_j} q_{t+h,t}^{\tau_j}$, resulting in the mean squared prediction error (MSPE) point forecast of the target at future time $t+h$.

¹⁴Note, the subscripts $t+h, t$ and $t+h|t$ are synonymous.

4.3.2 L1PQFC Forecast: Example

The L1PQFC forecast can be represented simply as an equation. Let $\hat{\mathbf{y}}_{t+h,t} = (1, \hat{y}_{t+h,t}^1, \hat{y}_{t+h,t}^2, \hat{y}_{t+h,t}^3)'$ be a vector of point forecasts available at time t (forecasting an arbitrary horizon h) with a discrete set of quantiles $\tau \in (\tau_1, \dots, \tau_5)$. Given $\hat{\mathbf{y}}_{t+h,t}$ and τ , we can obtain the L1-penalized quantile forecasts (L1PQF) a forecasting target for $t+h$ as $q_{t+h,t}^{\tau_j}$, $j = 1, \dots, 5$. Thus, we have

$$\begin{pmatrix} q_{t+h,t}^{\tau_1} \\ q_{t+h,t}^{\tau_2} \\ q_{t+h,t}^{\tau_3} \\ q_{t+h,t}^{\tau_4} \\ q_{t+h,t}^{\tau_5} \end{pmatrix} = \begin{pmatrix} \beta_0(\tau_1) & \beta_1(\tau_1) & 0 & \beta_3(\tau_1) \\ \beta_0(\tau_2) & \beta_1(\tau_2) & 0 & \beta_3(\tau_2) \\ \beta_0(\tau_3) & \beta_1(\tau_3) & 0 & \beta_3(\tau_3) \\ \beta_0(\tau_4) & 0 & 0 & \beta_3(\tau_4) \\ \beta_0(\tau_5) & 0 & 0 & \beta_3(\tau_5) \end{pmatrix} \times \begin{pmatrix} 1 \\ \hat{y}_{t+h,t}^1 \\ \hat{y}_{t+h,t}^2 \\ \hat{y}_{t+h,t}^3 \end{pmatrix}$$

Here, the model $\hat{y}_{t+h,t}^3$ is “strong” in the population since it predicts all quantiles, yet $\hat{y}_{t+h,t}^2$ is considered “fully weak” as it does not predict any quantile. The point forecasts represented by $\hat{y}_{t+h,t}^1$ is deemed “partially weak” as it aids in predicting some (τ_1, τ_2 , and τ_3) but not all (τ_4 and τ_5) quantiles. Thus, the L1PQFC forecast for estimating the target is produced as

$$\begin{aligned} \hat{y}_{t+h} &= \sum_{j=1}^5 w_{\tau_j} \beta_0(\tau_j) + \sum_{j=1}^3 w_{\tau_j} \beta_1(\tau_j) \hat{y}_{t+h,t}^1 + \sum_{j=1}^5 w_{\tau_j} \beta_3(\tau_j) \hat{y}_{t+h,t}^3 \\ &= \beta_0 + \beta_1 \hat{y}_{t+h,t}^1 + \beta_3 \hat{y}_{t+h,t}^3 \end{aligned} \quad (4.7)$$

with $\beta_0 = \sum_{j=1}^5 w_{\tau_j} \beta_0(\tau_j)$, $\beta_1 = \sum_{j=1}^3 w_{\tau_j} \beta_1(\tau_j)$, and $\beta_3 = \sum_{j=1}^5 w_{\tau_j} \beta_3(\tau_j)$.

The purpose of using the L1-penalized quantile regression is to determine the individual models that has the most utility for predicting the response variable y_{t+h} amongst all candidate, underlying forecasting models. In the example given above, the L1PQFC procedure completely excludes the fully weak model $\hat{y}_{t+h,t}^2$ from the linear equation, yet $\hat{y}_{t+h,t}^1$ and $\hat{y}_{t+h,t}^3$ are included for their partial and complete predictive utility, respectively, to $\hat{y}_{t+h}$. Additionally, as observed from the example above, since the utility of $\hat{y}_{t+h,t}^1$ in predicting $(q_{t+h,t}^{\tau_4}, q_{t+h,t}^{\tau_5})$ is partially weak (as the respective coefficients $\beta_1(\tau_4)$ and $\beta_1(\tau_5)$ are zero), the forecasting combination penalization procedure completely omits $\beta_1(\tau_4)$ and $\beta_1(\tau_5)$ in the prediction equation (4.7), while still retaining.

In terms of the weight selection for the L1PQFC models, we consider two simple, time-invariant weight schemes following [24]. The notion of time-invariance here arises from the fact that we consider the importance of quantiles in the determination of the optimal combination forecasts to be constant throughout the entire out-of-sample forecasting period for each forecast horizon h . Thus, the proposed L1-penalized quantile forecast combination method seeks to approximate the point (MSPE) forecast by assigning equal weights to a discrete set of conditional quantiles. Considering a discrete list of quantiles in the vector

$\tau = (\tau_1, \dots, \tau_m)$ with a specified equal weight, $w_\tau \equiv w$, we have our proposed penalized combination, L1PQFC, under two distinct sets of quantiles and equal weight considerations:

$$\text{L1PQFC1} = \frac{1}{3}q_{t+h,t}^{0.3} + \frac{1}{3}q_{t+h,t}^{0.5} + \frac{1}{3}q_{t+h,t}^{0.7}, \quad w = \frac{1}{3}$$

$$\text{L1PQFC2} = \frac{1}{5}q_{t+h,t}^{0.3} + \frac{1}{5}q_{t+h,t}^{0.4} + \frac{1}{5}q_{t+h,t}^{0.5} + \frac{1}{5}q_{t+h,t}^{0.6} + \frac{1}{5}q_{t+h,t}^{0.7}, \quad w = \frac{1}{5}$$

where L1PQFC1 and L1PQFC2 are simple mean aggregates.

4.3.3 Combination Model Comparisons

For comparison, we will separately include standard nested cases that can arise which likely results in inefficiency given the omitted impact of the inclusion of partially weak forecasting models.

Case 1: Standard Quantile Regression (SQR): Firstly, we can estimate linear quantile regression models with the selected forecasters (serving as predictors) given $\tau = (\tau_1, \dots, \tau_m)$ to obtain a standard quantile regression (SQR) forecast as:

$$\hat{y}_{t+h} = b_0 + b_1\hat{y}_{t+h,t}^1 + b_3\hat{y}_{t+h,t}^3 \quad (4.8)$$

$$\text{with } b_0 = \sum_{j=1}^5 w_{\tau_j} \beta_0(\tau_j), b_1 = \sum_{j=1}^5 w_{\tau_j} \beta_1(\tau_j), \text{ and } b_3 = \sum_{j=1}^5 w_{\tau_j} \beta_3(\tau_j).$$

The SQR forecasts are based on the linear quantile estimator with a set τ on the selected models $\hat{y}_{t+h,t}^1, \hat{y}_{t+h,t}^3$ and unity (intercept). In this prediction equation above, we have 3 discrete quantiles $\tau = (\tau_1, \dots, \tau_3)$. We also, for comparison, consider the 5 quantile specification as a separate model given $\tau = (\tau_1, \dots, \tau_5)$, resulting in forecast combination models: SFQR₁ and SFQR₂, respectively.

Case 2: Standard Ordinary Least Squares (SOLS). Here, we simply employ ordinary least squares to obtain a fixed linear regression forecast on equation (4.8) with y_{t+h} on the selected underlying forecasters given identical sets of discrete quantiles as in Case 1 resulting in models SOLS₁ and SOLS₂.

Note, for Case 2, in addition to it sharing the primary disadvantage of not utilizing the predictive utility of partially weak predictors in a given population as in Case 1, there is also efficiency loss resulting from the fact that the linear estimator will not be impervious to outliers' influence on estimation.

Following the idea of [44], we also further partition the difference between SOLS and L1PQFC forecasts into two elements. To show the relative importance of for including the presence of partially weak forecasting models, we have the following equation:

$$\text{MSPE}_{\text{SOLS}} - \text{MSPE}_{\text{L1PQFC}} = [\text{MSPE}_{\text{SOLS}} - \text{MSPE}_{\text{SQR}}] + [\text{MSPE}_{\text{SQR}} - \text{MSPE}_{\text{L1PQFC}}] \quad (4.9)$$

The first term on the right-hand side of (4.9) yields the additional loss of the SOLS forecast given the linear estimator's lack of robustness to the errors in estimation, while the second represents extra loss or inefficiency due to partially weak forecasting models' presence.

Case 3: Egalitarian LASSO (EL). Here, for comparison, we augment the standard LASSO¹⁵ penalized estimation problem to the Egalitarian LASSO procedure where the weights we shrink are the deviations from equal weights toward zero. In terms of the implementation into software, the Egalitarian LASSO construction is tantamount to performing the standard LASSO selection on a transformed response where y_{t+h} is centered by subtracting the average forecasts across all time periods t . Let $\hat{y}_{t+h,t}^i$ represent the vector of point forecasts for model i available at time t . Thus, the objective function for the method is

$$\text{EL} = \arg \min_{\beta} \left(\sum_t \left(y_{t+h} - \sum_{i=1}^K \beta_i \hat{y}_{t+h,t}^i \right)^2 + \lambda \sum_{i=1}^K \left| \beta_i - \frac{1}{K} \right| \right)$$

with β_i representing the coefficients or weights associated with each of the forecasts from the conditional models, and λ is the penalty term which is chosen such that the minimum mean cross-validated error is obtained¹⁶. Note, in our estimation, we do utilize the cross-validation procedure where the number of folds are arbitrarily set to 10 folds. Also, we will show the relationship of this method's objective function to the standard LASSO for software usage in the Appendix.

Case 4: Robust Forecast Combination (RFC). As our final comparison to our L1PQFC proposed method, we will utilize an adjusted version of [40] which is just an average point forecasts obtained from the aforementioned considerations. The idea is that this model can potentially be less sensitive to the sample split for the recursive estimation. Let $\hat{r}_{t+h,t}^j$ represent the point forecasts from our combination models L1PQFC1, L1PQFC2, SQR1, SQR2, SOLS1, SOLS2, and EL, $j = 1, \dots, 7$, across a specified forecast horizon h . The Robust Forecast Combination (RFC) is defined as:

$$\text{RFC} = \frac{1}{M} \sum_{j=1}^M \hat{r}_{t+h,t}^j$$

with $M = 7$, and $h = \{1, 2, 3, 4, 8, 12\}$.

4.4 The Forecasting Procedures: Forecasting the Monthly Inflation Gap

The Consumer Price Index, or CPI, represents goods and services purchased for consumption by a reference population and is often employed as a gauge of effectiveness of certain economic policies. It is the most commonly used method to measure inflation after seasonal adjustments. Thus, a seasonally adjusted CPI serves as the basis to compute our principle target of interest: inflation. Inflation rate is the rate in which goods and/or services increases over time for a given nation. Intuitively, this increased rate of inflation

¹⁵Note, the standard LASSO's objective function involves the selection to and shrinkage of weights toward 0. That is, the objective function given for standard LASSO lacks the " K^{-1} " in the second (additive) term in absolute value.

¹⁶If there exists such a situation as to the minimum penalty for λ selects exactly 0 models, then and only then will we adjust the penalization parameter to be $\lambda.1\text{SE}$ instead of $\min(\lambda_i)$

directly corresponds to decreased purchasing power of currency. A voluminous amount of previous research has been proposed in the econometrics literature regarding forecasting inflation in a variety of contexts. An important, perhaps seminal survey of forecasting inflation reviews state-of-the-art procedures in forecasting ([27]). In this work, the authors utilize a well-established, robust roundup of forecasting models¹⁷ based on four principles with differences in forecasting evaluated by the root mean squared prediction error (RMSPE). Another piece of literature that solely focuses on the effectiveness of the Phillips-Curve (PC) model’s utility in forecasting inflation. Primary interest surrounds how the covariates composing the PC model that uses alternative, aggregate macroeconomic indicators, aside from unemployment, ultimately contributes to smaller mean-squared prediction errors (MSPE) with additional statistically significant forecasting gains ([45]). Another significant, recent analysis ([46]) narrows its focus to extending the literature for inflation forecasting in the context of emerging forecasting economies (EMEs) by focusing on the utility gained from using a traditional variant of the random walk model (RW-AO) with incorporation of subjective intuition for mitigating structural time series shifts. Though the focus of the previously cited literature presents efficacious results regarding success of forecasting inflation through the use of comparison of various model performance metrics, the utility from the theory of combination models to forecast inflation is clear and will be demonstrated in the proceeding analyses. Below, we will apply the L1PQFC model and other combination model comparisons described in subsections 4.3.1-4.3.3 to the problem of forecasting a function of inflation rate: the inflation gap.

4.4.1 Part I. Production of the out-of-sample (OOS) forecasts

We assume that there are economic institutions that use different econometric models to make forecasts and pretend where only the point forecasts from each institution are observed. These point forecasts are represented by the conditional mean of each model. We first estimate these 9 conditional models for the forecasting target: annualized monthly inflation gap. We principally follow the idea and modeling considerations of ([27], [29]) with a key difference in terms of the response variable. All models that will be considered here will be focused on forecasting monthly annualized inflation “gaps”. Contrasting from previous literature, we employ the annualization transformation on the inflation rate response, which is simply a process of adjusting the inflation rate to reflect the amount inflation would have changed over a year’s time given that it had constantly grown at the given rate. Thus, the time series of the annual inflation rate is the growth rate with respect to the corresponding month of the previous year. We denote monthly annualized inflation as π_t , which is $\pi_t = \log(P_t/P_{t-12}) * 100\%$ where P_t is defined as the consumer price index (CPI) at time t . We define τ_t as the corresponding inflation trend. The inflation trend is computed as a moving average process based on the previous 12 months of the observed real time monthly annualized

¹⁷The roundup of forecasting models presented in this literature are quite varied which includes (but not limited to) recursive autoregression (RAR), a Phillips-Curve (PC), unobserved component stochastic volatility (UCSV), and Bayesian Model Averaging (BMA) on quarterly inflation “gaps” (computed as the difference between quarterly inflation rates and a quarterly trend measure)

inflation ($\tau_t = \frac{1}{12} \sum_{i=1}^{12} \pi_{t-i}$). Thus, we can compute our response of monthly annualized inflation gap as $g_t = \pi_t - \tau_t$. We seek to compute the out-of-sample (OOS) forecasts over the six forecasting horizons: $h = \{1, 2, 3, 4, 8, 12\}$.

We estimate these conditional models to the annualized inflation gap which are then used to compute forecasts for the given forecast horizons of interest. A general setup for monthly annualized inflation gap is presented for the conditional model of interest as follows

$$y_{t+h} \equiv g_{t+h} = \pi_{t+h} - \tau_{t+h} = \mathbf{\Pi}'_{t+h,t} \boldsymbol{\alpha} + \varepsilon_{t+h} \quad (4.10)$$

$$\varepsilon_{t+h} \sim N(0, \sigma_{t+h}^2)$$

where g_{t+h} represents the monthly annualized inflation gap at some future time $t+h$, π_{t+h} represents the annualized inflation rate at some future time period $t+h$, τ_{t+h} represents the annualized 12-month inflation trend, and $\mathbf{\Pi}_{t+h,t}$ is a vector of covariates that the linear out-of-sample forecast production whose values are known at time t . We present the format of the 10 econometric models (8 of which follow [29]), with an additional consideration for the historical average as point forecasts of our response. Finally, we consider the equal weight forecast combination¹⁸.

- *Model 1.* First Order Autoregressive Model (AR(1)): Equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = g_t$ with $\boldsymbol{\alpha} = 0.46$
- *Model 2.* Second-Order Autoregressive Model (AR(2)): Equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = (1, g_t, g_{t-1})$ with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$.
- *Model 3.* RW-AO: Variation of a pure random walk with equation (4.10) as $\mathbf{\Pi}'_{t+h,t} = \frac{1}{4} \sum_{j=1}^4 g_{t-j+1}$ and $\boldsymbol{\alpha} = 1$ [47].
- *Model 4.* PC (Backward Looking): This model is the Phillips Curve, which has a very established history of inflation forecasting as it relates the unemployment rate or, some other measure of aggregate macroeconomic activity, to inflation rate via an equation. Equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = (1, g_t, u_t)$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$ with u_t as the monthly annualized unemployment rate.
- *Model 5.* PC (Backward Looking, Extended): This is the same as Model 4, but with $\mathbf{\Pi}'_{t+h,t} = (1, g_t, g_{t-1}, u_t, u_{t-1})$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)'$.
- *Model 6.* PC-hybrid (Backward and Forward Looking): This is equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = (1, g_{t+h,t}, g_t, u_t)$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$, with $g_{t+h,t}$ is the market expectation of annualized inflation

¹⁸Although we do have the covariates underlying our forecasts that will ultimately be used for our combination model considerations in the subsequent section, we will assume immediately, that these “inputs” are indeed privatized or unknown to avoid the computation of a likely inefficient “super model” that constructs a (quantile) forecast that utilizes all available inputs/covariates.

rate. The data here comes from Aruoba Term Structure of Inflation Expectations (ATSIX) published by the Philadelphia FED.

- *Model 7.* PC-hybrid 2 (Backward and Forward Looking, Extending): This is the equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = (1, g_{t+h,t}, g_t, u_t, \Delta u_t)$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)'$.
- *Model 8* Survey-Based Forecasts: This is the equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = g_{t+h,t}$ and $\boldsymbol{\alpha} = 1$.
- *Model 9.* Historical Average: This is simply the equation (4.10) with $\mathbf{\Pi}'_{t+h,t} = g_{t+h,t} = \frac{1}{N} \sum_{j=1}^N g_{t-j+1}$ and $\boldsymbol{\alpha} = 1$.
- *Model 10.* Equal Weight Combination Model. Let $\hat{y}_{t+h,t}^i$ represent the out of sample (OOS) forecast from each of Models $i = 1, \dots, 9$. The Equal Weight Combination Model given as: $\bar{y}_{t+h,t} = \frac{\sum_{l=1}^M \hat{y}_{t+h,t}^l}{M}$ with $M = 9$ ¹⁹

Figure 4.1 shows the Monthly annualized inflation gap from 1998:M1-2018:M8. The monthly annualized inflation gap demonstrably exhibits volatility between 2008 until 2012 perhaps attributable to the mortgage crisis in the United States. Via the Augmented Dickey-Fuller Test, the annualized inflation gap is a confirmed stationary process as the null hypothesis of a unit root is rejected with a p-value less than 0.01 given $\alpha = 0.05$.

The starting point for the estimation sample is always 1998:M1. The actual forecasts themselves are generated strictly with a recursive scheme (e.g. sample size expansion). Choosing the recursive scheme as opposed to another (such as the moving window scheme) is advantageous as all the data for each scenario available at a point in time are utilized and the precision of our forecasts sharpens further over time. Here, the individual models are estimated using a sample that starts at 1998:M1 and ends at (2002:M4)- h

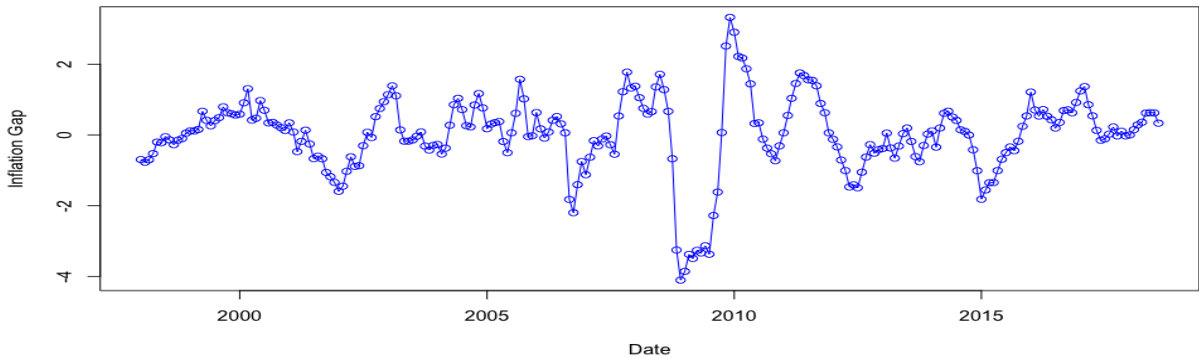


Figure 4.1: U.S. Annualized Monthly Inflation Gap (1998:M1 - 2018:M8)

¹⁹From previous literature ([32]), this equal weight combination model is a natural benchmark as it has a degree of optimality when forecast errors from individual models have identical correlations (pairwise) and homogeneous variances.

(conditional on the forecast horizon h) and continuing into the last observation in the out-of-sample period. The full forecast evaluation runs from 2002:M4 until 2018:M8. The entire out-of-sample period results in 197 observations of monthly annualized inflation gap. Thus, the resulting point forecasts for these models are represented as $\hat{y}_{t+h,t}^i$ with $j = 1, \dots, 10$ for all forecast horizons $h = \{1, 2, 3, 4, 8, 12\}$. In terms of assessing the out-of-sample performance of models 1-10 from this procedure, we superimpose the empirical point forecasts for all 10 models against the OOS monthly annualized inflation gap across these forecast horizons shown in Figure 4.2.

In addition to the graphical display for assessing our point forecasts performance across all horizons, one important OOS accuracy metric that is a trustworthy measure of decent forecast performance is the Theil's U statistic. This statistic is an accuracy measure that emphasizes the importance of large errors (as in MSE) as well as providing a relative basis for comparison with naïve forecasting methods [48]. In addition to the visualizations of forecasting provided previously, we utilize this statistic to quantify forecast quality. The Theil's U statistic is computed (model i and forecast horizon h dependent) for all 10 models for all 6 horizons as

$$U_i = \sqrt{\frac{\sum_{t=1}^{n-1} \left(\frac{\hat{y}_{t+h,t}^i - y_{t+h}}{y_t} \right)^2}{\sum_{t=1}^{n-1} \left(\frac{y_{t+h} - y_t}{y_t} \right)^2}} \quad i = 1, \dots, 10 \quad h = 1, 2, 3, 4, 8, 12$$

The Theil's U -statistic can be interpreted as dividing the RMSE of the proposed model by the RMSE of a naïve model. The naïve model is simply defined to be the most recently observed forecast value of y . If $U > 1$, the proposed forecasting model is worse than the predictive performance of a naïve model. If $U < 1$, the proposed forecasting model (the smaller, the better), then our forecasting models produces more accurate point forecasts for the conditional mean of our response variable.

In Table 4.1, we report the results for this metric across all models. As can be seen, for each horizon, the majority of the models yield a value less than 1, which indicates decent forecast qualities for these models' predictions. Eight out of the 10 models for $h = \{1, 3, 12\}$ outperform a naïve model. The models PC-EXT, PC-BF, SBF, and Hist Avg yield more accurate forecasts relative to the naïve model which may be indicative of their utility in our subsequent combination considerations. The RW-AO model consistently performs the worst across all horizons. In the appendix, we have a table that also shows various OOS accuracy metrics as well for the holdout sample of monthly annualized inflation gap.

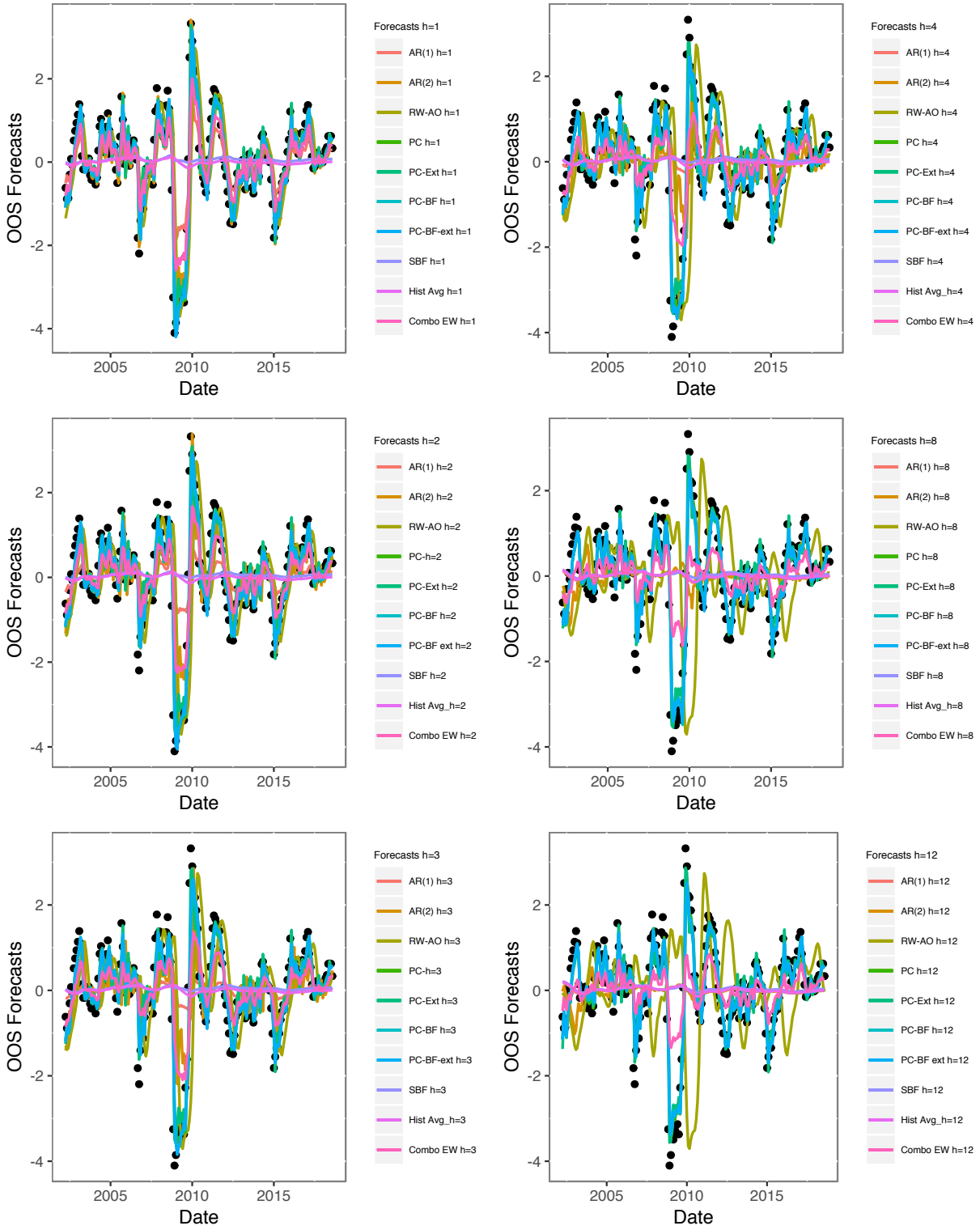


Figure 4.2: OOS Forecasts, $h=1,2,3,4,8,12$

Table 4.1: Out-of-Sample Accuracy Measure: Theil's U

	Theil's U					
Models	h=1	h=2	h=3	h=4	h=8	h=12
AR(1)	0.99563163	1.04446822	0.99362701	1.02355381	0.98132999	0.97681607
AR(2)	0.894149079	1.3365389	0.69367738	1.5152365	1.19190884	1.0136447
RW-AO	1.259741076	1.42021203	1.66193752	1.91229286	1.26715396	3.40767187
PC	1.011746001	1.01563066	1.0148535	1.02199768	1.01848879	0.99605634
PC-Ext	0.957415681	0.96458131	0.96244906	0.96792295	0.97188511	0.982221
PC-BF	0.97724463	0.98342343	0.98422837	0.99503159	0.99897814	0.98386463
PC-BF ext	0.981050529	0.98904299	0.99080679	0.99947757	1.00236475	0.99213601
SBF	0.967864618	0.96720252	0.9658693	0.96547471	0.95350275	0.96750843
Hist Avg	0.991624443	0.99160316	0.99037785	0.99022862	0.97515881	0.98781436
Combo EW	0.992521848	1.05412882	0.9973427	1.12342914	0.93639027	0.74131405

4.4.2 Part II: Forecasting Combination Procedures

We produce 197 point forecasts using an out-of-sample recursive scheme for all horizons $h = \{1, 2, 3, 4, 8, 12\}$ for all 9 conditional models and an equal weight benchmark. This results in six datasets of dimension 197×10 with rows representing the values of the point forecasts of 10 models (columns) conditional on the six forecast horizons. We denote the point forecasts from our conditional models as $\hat{\mathbf{y}}_{t+h,t} = (1, \hat{y}_{t+h,t}^1, \dots, \hat{y}_{t+h,t}^9)'$, which also includes unity for the intercept. The last model, which is the equal weight combination model, denoted as $\bar{y}_{t+h,t}$, will be used as a benchmark for our next step. The equal weight combination model is a reasonable combination candidate as forecasts constructed from simple averages typically achieve a substantial reduction in both biases and variances through averaging out the biases from underlying forecasters ([49]). Note, also, as the last 8 observations do not span a full year, we decide to remove the last 8 months of our datasets corresponding to 2018 such that, for all h .

In this portion of our analysis highlighting the principle application of the forecasting combination problem, we use our point forecasts $\hat{\mathbf{y}}_{t+h,t}$, corresponding to six horizons, $h = \{1, 2, 3, 4, 8, 12\}$, to generate out-of-sample forecasts of monthly annualized inflation gap h -step ahead, y_{t+h} , in the form of 8 models²⁰:

- 1.) The L1-penalized quantile forecast combination (L1PQFC) method using the two weighting schemes presented earlier,
- 2.) The Standard Quantile Regression (SQR) method,
- 3.) The Standard Ordinary Least Squares (SOLS) method,
- 4.) Egalitarian LASSO method, and
- 5.) Robust Forecast Combination Method.

If the models forecasting monthly annualized inflation gaps are viable in the covariate vector $\hat{\mathbf{y}}_{t+h,t} = (1, \hat{y}_{t+h,t}^1, \dots, \hat{y}_{t+h,t}^9)'$ for forecasting the out-of-sample portion of monthly annualized inflation gap, y_{t+h} , then the point forecasts in $\hat{\mathbf{y}}_{t+h,t}$ should outperform the equal weight combination benchmark, $\bar{y}_{t+h,t}$, computed in our previous step, across all forecast horizons h . Our aim here is to assess whether this penalization procedure in our proposed L1PQFC methods (which accounts for partially weak models/point forecasts) outperforms: 1.) the methods that do not account for the partial nature of our underlying

²⁰For robustness, we assess whether the proposed L1PQFC method is superior to the other combination procedures when predicting the next h step ahead forecast of monthly annualized inflation gap, noting that y_{t+h} for $h = \{1, 2, 3, 4, 8, 12\}$ are the same OOS observations in each scenario.

forecasting models (SQR, SOLS), II.) the Egalitarian Lasso (EL) model which does entail selection of covariates as well as weight shrinkage ([50]) and III.) a mean aggregate combination of all of these combination considerations (RFC).

We use a recursive estimation scheme here as well, dividing the sample into an in-sample and out-of-sample period. The in-sample period consists of 141 observations that ranges from 2002:M4 to 2013:M12 that expands one month forward periodically for all horizons. The out-of-sample period consists of 48 observations that ranges from 2014:M1 to 2017:M12. One evaluation statistic that we use is R_{OS}^2 , which compares the combination forecast models' predictive ability from our conditional models, denoted as $\hat{r}_{t+h,t}^i$, to our equal weight combination benchmark's, $\bar{y}_{t+h,t}$, predictive ability for all horizons. We can compute this evaluative metric for all of our proposed combination models in our six horizons as follows

$$R_{OS_{i,h}}^2 = 1 - \frac{MSPE_{\hat{r}_{t+h,t}^i}}{Var(y_{t+h})} \quad (4.11)$$

where $MSPE_{\hat{r}_{t+h,t}^i} = \frac{1}{T^\circ} \sum_t (y_{t+h} - \hat{r}_{t+h,t}^i)^2$, $i = 1, \dots, 8$ and $Var(y_{t+h}) = \frac{1}{T^\circ} \sum_t (y_{t+h} - \bar{y}_{t+h,t})^2$, $h = \{1, 2, 3, 4, 8, 12\}$, and T° is the number of forecasts in our out-of-sample forecast epoch. The competing forecast combinations superiority to the benchmark model is clear if $R_{OS_{i,h}}^2 \geq 0$. In addition to the $R_{OS_{i,h}}^2$ metric, we employ both the Diebold-Mariano (DM) test which tests the null of equal forecast predictive accuracy, on average, relative to our benchmark, and the (adjusted) Clark and West test under the null $R_{OS_{i,h}}^2 \leq 0$. This test can be viewed as a test for equal MSE of non-nested forecasts. Details can be found in the appendix.

4.5 Forecasting Combination Results

Figures 4.3-4.8 show the bivariate relationship between the cumulative squared prediction error and the competing combination forecast models compared to the equal weight linear combination benchmark. The cumulative squared prediction error (cpe) can be defined as the cumulative sum of the difference between the squared prediction error of the OOS responses and the benchmark forecast and the squared prediction error between the OOS response and the combination model considerations. Recalling $\bar{y}_{t+h,t}$ represents the equal weight benchmark for a given horizon and that $\hat{r}_{t+h,t}^i$ represents the point forecasts from our forecasting models $i = 1, \dots, 8$. These panels show the cumulative performance of a given combination forecasting model compared to the equal weight combination over time. When the curve for each panel displays an upward trend, we note that the combination model of interest outperforms the benchmark, whereas if the line trend is downward, the benchmark is the outperforming model. Additionally, a downward trend in these panels for Figures 4.3-4.8 indicate indicates that the MSPE for each competing combination model is higher than the benchmark across the 48 month OOS period. Therefore, for every panel representing the cumulative squared prediction error, these combination models have a lower MSPE across the out-of-sample period relative to the benchmark

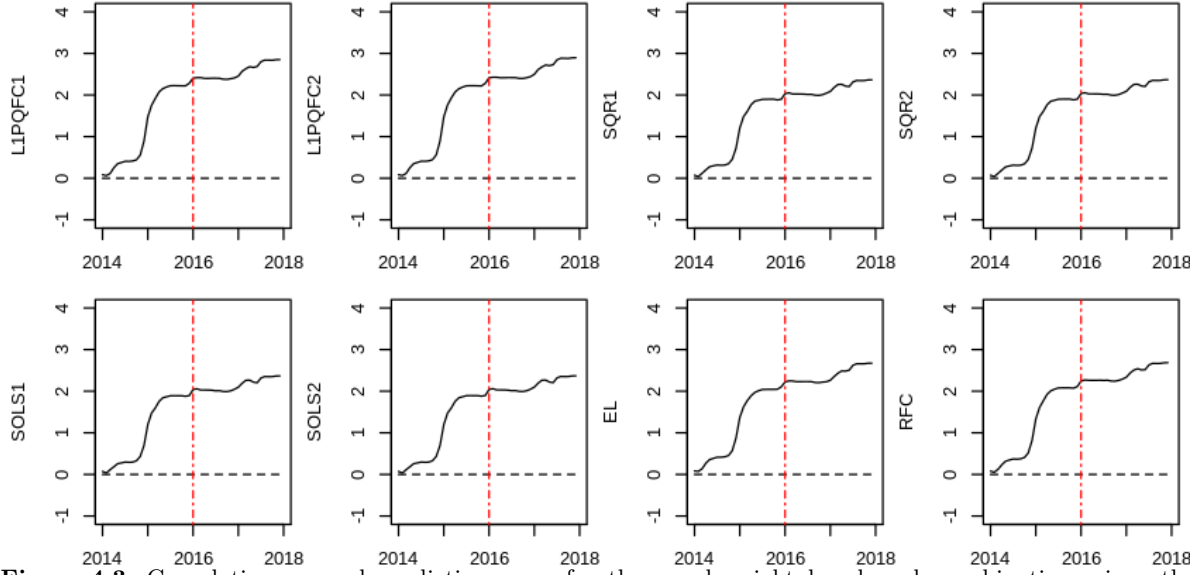


Figure 4.3: Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=1$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.

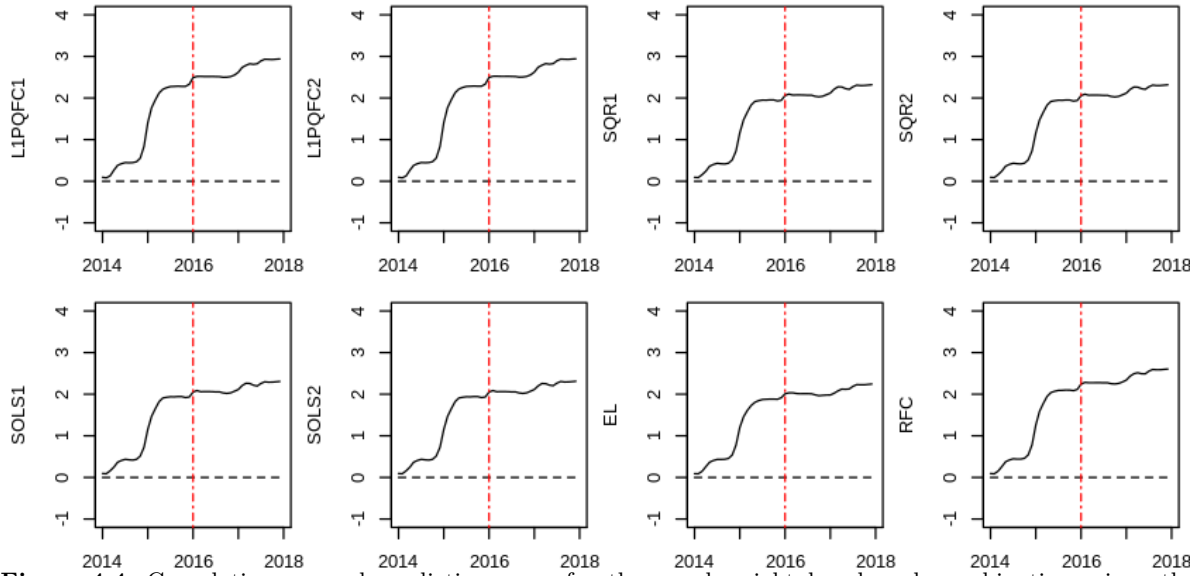


Figure 4.4: Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=2$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.

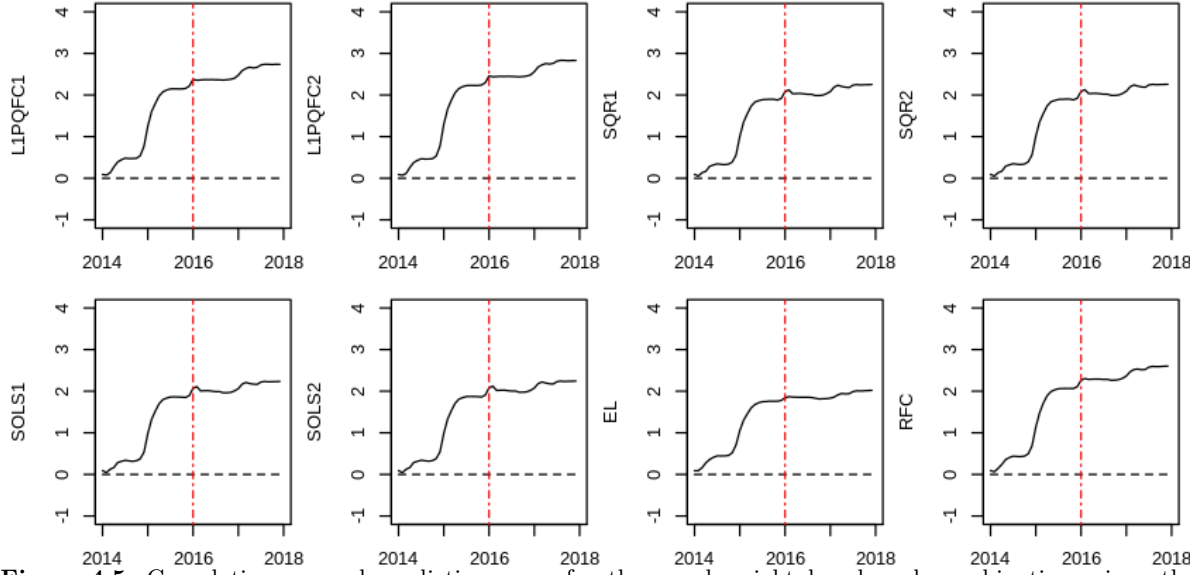


Figure 4.5: Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=3$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.

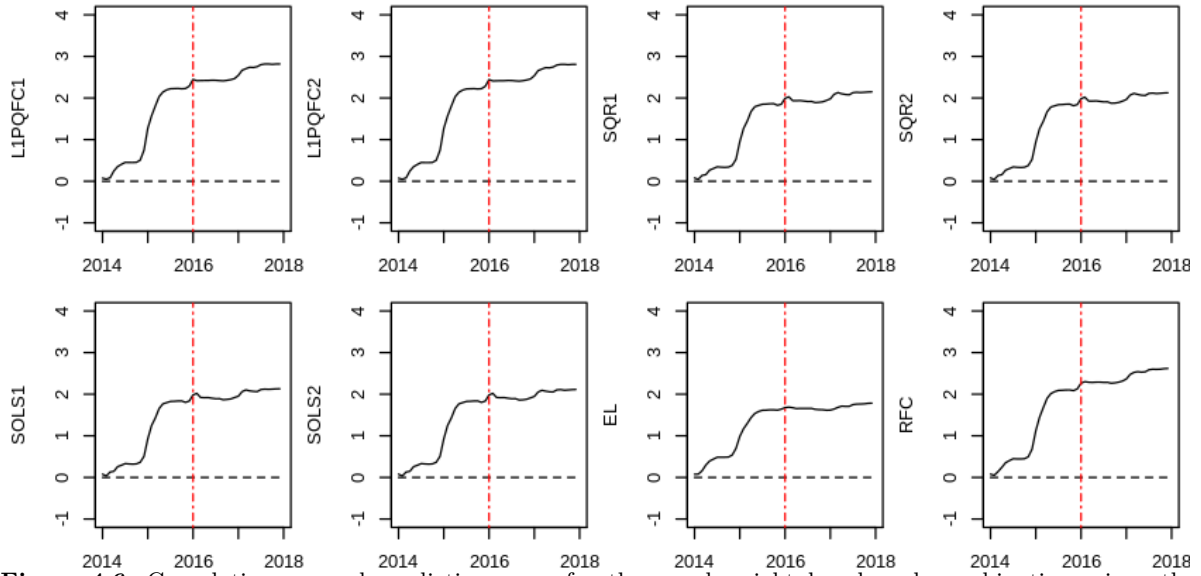


Figure 4.6: Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=4$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.

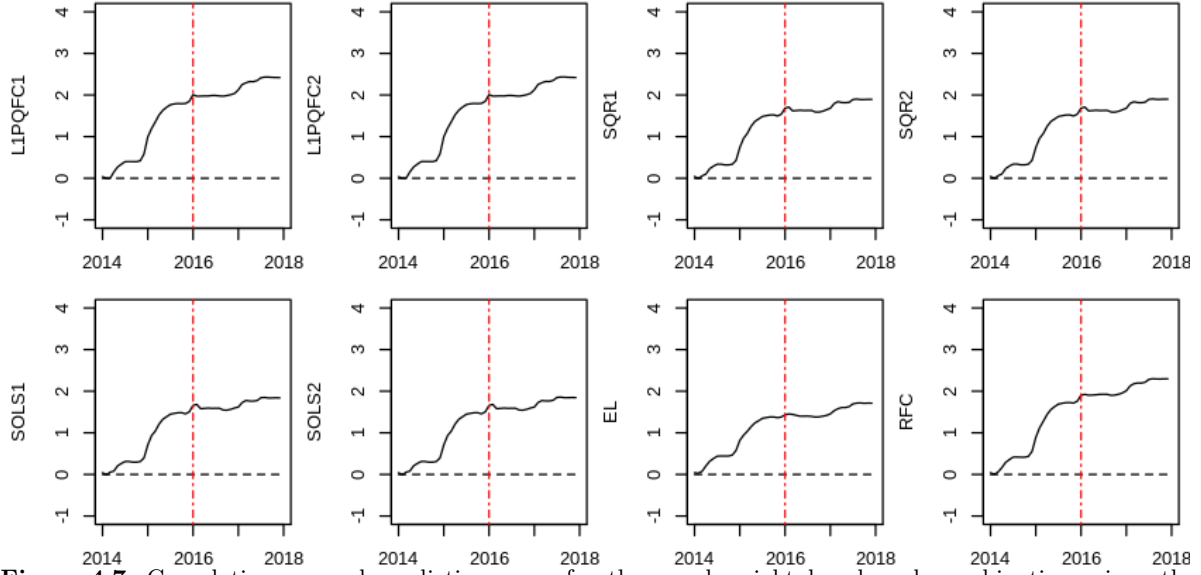


Figure 4.7: Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=8$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time..

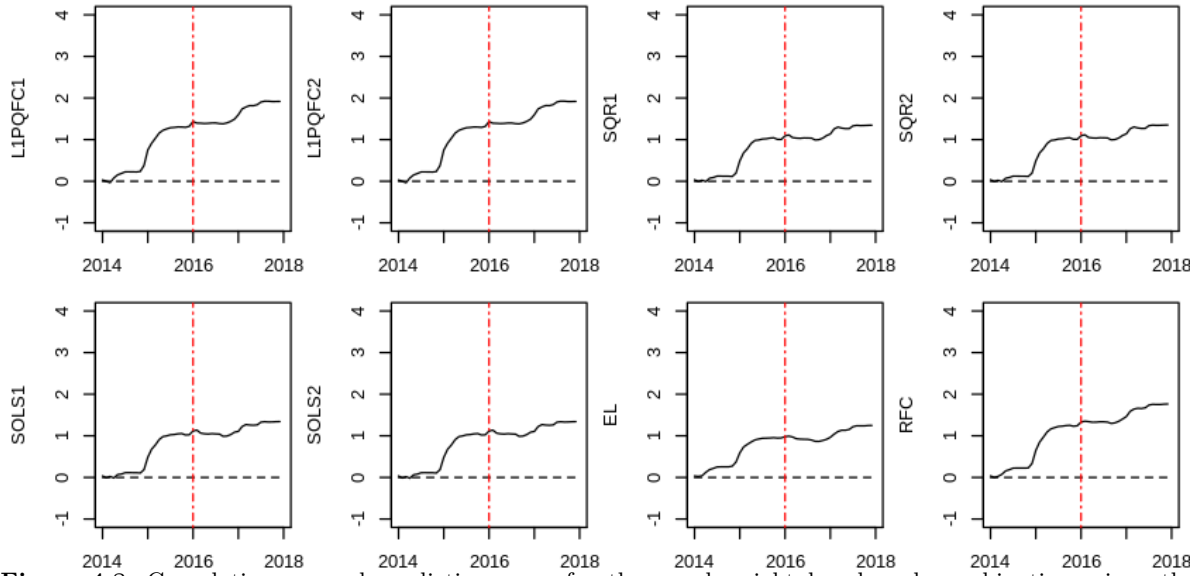


Figure 4.8: Cumulative squared prediction error for the equal weight benchmark combination minus the cumulative squared prediction errors for the competing combination forecasting models $h=12$, 2014:M1 – 2017:M12. The non-negative, increasing line reveals that our proposed forecasting combination models prevails in OOS predictive ability relative to the equal weight combination model over time.

displaying quite significant out-of-sample gains in comparison. All Figures 4.3-4.8 consistently and clearly show the superior nature of our proposed L1PQFC1 and L1PQFC2 models in terms of predictive performance reaching its apex in the latter months (after 2016) of our out-of-sample period which indicates not only clear superiority (performance-wise) relative to the equal weight benchmark, but also between the other combination considerations.²¹

Before 2016, all combination models in Figures 4.3-4.8 seem to show relatively the same upward trajectory for out-of-sample performance. The sloping line from 2015 until 2016 increases sharply particularly in all cases. After 2016, the out-of-sample cumulative squared prediction error seems to relatively level out without increasing slopes across all models for each horizon. We note that for Figures 4.7 and 4.8, unsurprisingly, for these combination considerations, the predictive gains relative to the shorter forecast horizons are significantly lower. Therefore, we can conclude that using point forecasts as covariates with a further horizon seemingly in terms of predictive performance of the OOS response variable relative to our combination benchmark computed from a shorter horizon. However, in each panel, the superiority of the cumulative squared prediction errors in terms of performance for our proposed L1PQFC over the equal weight benchmark is clear across all horizons.

For direct model combination comparisons and recalling that our proposed combination considerations L1PQFC1 and L1PQFC2 compared to SQR1, SQR2, SOLS1, and SOLS2 primarily differ through the latter four models' inability to include the presence of partially weak predictors given τ , we observe the clear utility for including the partially weak models in Figures 4.3-4.8. Before 2016, both sets of L1PQFC and SQR models perform equally well; however, after 2016, L1PQFC models clearly outperform the SQR models which seems to indicate the utility of the inclusion of partially weak predictors for out-of-sample predictive performance using the point forecasts (predictors) across all horizons.

Also, noting the comparison between the SQR models and SOLS models, we observe similar trajectories throughout the entire out-of-sample periods across all horizons. As the regression equation is the same for SQR1, SQR2, SOLS1, and SOLS2, where the only difference is how estimators are constructed via the difference in the QR estimator and the OLS estimator, then a clear conclusion is that there are hardly any outliers in the response variable, y_{t+h} (for the chosen underlying models given from the L1 penalty), indicating the similarity in estimation errors for these methods for the two sets of quantiles considered in τ .

This is also reaffirmed by assessing the out-of-sample forecasting performance in Table 4.2. We display the out-of-sample diagnostics for all combination models considered in this analysis. For our models where our point forecasts are based on horizons $h = \{1, 2, 3, 8, 12\}$, we note that the L1PQFC2 models is the superior choice based off the models' predictive performance in our OOS response indicated by $R_{OS}^2(\%)$, the Clark and West (CW), and Diebold-Mariano (DM) Test. For both the DM and CW tests, we note that the alternative hypothesis we specifically test is one-sided; thus, the clear statistical significance of both tests ($\alpha = 0.10$) indicates that the models presented have far superior OOS predictive ability relative

²¹The dashed (red) lines of demarcation in Figures 4.3-4.8 have no significance other than to highlight the midway point of out-of-sample periods M1:2014-M12:2015 and M1:2016-M12:2017.

Table 4.2: OOS Combination Forecasting Diagnostics, 2014:M1-2017:M12, $h = \{1, 2, 3, 4, 8, 12\}$

	h=1			h=2			h=3			h=4			h=8			h=12		
Models	$R_{OS}^2(\%)$	DM	CW	$R_{OS}^2(\%)$	DM	CW	$R_{OS}^2(\%)$	DM	CW	$R_{OS}^2(\%)$	DM	CW	$R_{OS}^2(\%)$	DM	CW	$R_{OS}^2(\%)$	DM	CW
L1PQFC1	82.93%	0.00	0.00	86.44%	0.00	0.00	82.09%	0.00	0.00	85.47%	0.00	0.00	83.28%	0.00	0.00	79.31%	0.00	0.00
L1PQFC2	84.19%	0.00	0.00	86.54%	0.00	0.00	84.89%	0.00	0.00	85.21%	0.00	0.00	83.32%	0.00	0.00	79.32%	0.00	0.00
SQR1	68.81%	0.00	0.00	68.15%	0.00	0.00	67.58%	0.00	0.00	65.14%	0.00	0.00	65.23%	0.00	0.00	55.72%	0.00	0.00
SQR2	68.89%	0.00	0.00	68.13%	0.00	0.00	67.77%	0.00	0.00	64.47%	0.00	0.00	65.39%	0.00	0.00	55.93%	0.00	0.00
SOLS1	68.88%	0.00	0.00	67.83%	0.00	0.00	67.06%	0.00	0.00	64.63%	0.00	0.00	63.25%	0.00	0.00	55.48%	0.00	0.00
SOLS2	68.94%	0.00	0.00	67.96%	0.00	0.00	67.31%	0.00	0.00	64.10%	0.00	0.00	63.51%	0.00	0.00	55.44%	0.00	0.00
EL	77.55%	0.00	0.00	65.99%	0.00	0.00	60.62%	0.00	0.00	54.14%	0.00	0.00	58.94%	0.00	0.00	31.97%	0.00	0.00
RFC	78.11%	0.00	0.00	76.61%	0.00	0.00	78.13%	0.00	0.00	79.42%	0.00	0.00	79.11%	0.00	0.00	72.00%	0.00	0.00

to the equal weight combination benchmark. Interestingly, for $h = 4$, we note that the model based on 3 distinct quantiles $\tau = (0.3, 0.5, 0.7)$ (as opposed to 5) has the superior performance for all models. Also, noticeable is the stunningly poor performance of the Egalitarian LASSO (EL) model which employs a form of penalty selection as our proposed L1PQFC models across all horizons h . Specifically, this is evident in the higher forecast horizons which highlights the stunning degradation of out-of-sample R_{OS}^2 metric. A few reasons of the poor performance seem obvious. Firstly, as stated in [50], this is likely due to the Egalitarian LASSO's inability to properly select the underlying forecasting models in the correct direction. As evident in its objective function, the Egalitarian LASSO selects and shrinks towards equal weights rather than the standard LASSO's shrinkage towards zero weights, which induces the inevitability of underlying forecasting models to not actually be nullified when they contribute little towards accurately predicting the response at our arbitrary levels of τ ²². Lastly, the EL model estimation does not have the built-in protection against traditional estimation errors resulting from lack of outlier robustness in comparison to our quantile estimator, which also contributes to its overall inferiority particularly in relation to the L1PQFC and SQR combination model considerations.

Table 4.3 highlights the utility of the L1PQFC forecasts via the decomposition in the mean squared prediction error (MSPE). The decomposition measures the additional MSPE loss of the FOLS forecasts relative to the L1PQFC forecasts. The term $(MSPE_{SQR} - MSPE_{L1PQFC})$ represents the extra loss caused by the presence of partially weak forecasting models in the population, while the term $(MSPE_{SOLS} - MSPE_{SQR})$ represents the quantile regression estimators' impervious nature in being resistant to outlying observations. For the point forecasts forecasting all horizons $h = \{1, 2, 3, 4, 8, 12\}$, we note the dominating impact of the additional loss resulting from not including partially weak models/covariates (models that help predict some, but not all quantiles), whereas the OLS lack of robustness has relatively little consequence regarding the additional loss. Also, for $h = 4$ only, we measure the additional MSPE loss of SOLS forecasts relative to the L1PQFC forecast using only 3 quantiles ($\tau = (0.3, 0.5, 0.7)$) as opposed to 5 ($\tau = (0.3, 0.4, 0.5, 0.6, 0.7)$) as it is the only case in which L1PQFC1 outperforms L1PQFC2 in terms of out-of-sample predictive performance.

²²The selection of underlying forecasters towards equality of weights rather than zero weights presumes that the underlying forecasting models essentially belongs to the set to be combined [50].

Table 4.3: Mean squared prediction error (MSPE) decomposition

$\text{MSPE}_{\text{SOLS}} - \text{MSPE}_{\text{L1PQFC}} =$	$(\text{MSPE}_{\text{SOLS}} - \text{MSPE}_{\text{SQR}})$	$+$	$(\text{MSPE}_{\text{SQR}} - \text{MSPE}_{\text{L1PQFC}})$
h	% of total		% of total
1	0%		100.00%
2	0.92%		99.08%
3	2.62%		97.38%
4	1.75%		98.25%
8	9.48%		90.52%
12	2.02%		97.98%

Note: The decomposition measures the additional MSPE loss of FOLS forecasts relative to the L1PQFC forecasts. The first element $(\text{MSPE}_{\text{SOLS}} - \text{MSPE}_{\text{SQR}})$ measures the additional loss from the OLS estimator's lack of robustness to estimation errors, while the second element $(\text{MSPE}_{\text{SQR}} - \text{MSPE}_{\text{L1PQFC}})$ represents the extra loss caused by the presence of partially weak predictors in the population. Note, that the L1PQFC, SQR and SOLS forecasts correspond to models L1PQFC2, SQR2 and SOLS2 for $h = \{1, 2, 3, 8, 12\}$ in the paper. However, L1PQFC, SQR and SOLS forecasts corresponds to models L1PQFC1, SQR1, and SOLS1 for $h = 4$ in the paper as L1PQFC1 is superior to L1PQFC2 based off R_{OS}^2 .

4.5.1 Additional benefits of L1PQFC models

The MSPE of a forecasting model can be split into two parts: the squared forecast bias as well as the unconditional forecast variance. We recall that in this application, the MSPE of the forecast of monthly annualized inflation gap compared to our combination models, is $\text{MSPE}_{\hat{r}_{t+h,t}^i} = \frac{1}{T^\circ} \sum_t (y_{t+h} - \hat{r}_{t+h,t}^i)^2$, $i = 1, \dots, 8$, $h = \{1, 2, 3, 4, 8, 12\}$ where T° represents the total number of forecasts that spans our out-of-sample size (which are the 48 months between M1:2014 until M12:2017). The unconditional variance for the forecast for each horizon, thus, is $\frac{1}{T^\circ} \sum_t (\hat{r}_{t+h,t}^i - \frac{1}{T^\circ} \sum_t \hat{r}_{t+h,t}^i)^2$, $i = 1, \dots, 8$, $h = \{1, 2, 3, 4, 8, 12\}$, and the squared bias is calculated as the difference between the MSPE and the unconditional forecast variance ([51]). The relative forecast variance (squared bias) is computed as the difference between the forecast variance (squared bias) of the i th model and the forecast variance (squared bias) of the combination benchmark. This implies that the value of the relative forecast variance (squared bias) of the EW benchmark is zero. Each point on the diagonal represents a forecast with the same MSPE as the benchmark. Figures 4.9 - 4.11 represent bivariate graphical displays for the difference in the squared bias forecast difference (x-axis) and the forecast variance difference (y-axis) for each of the combination models (measured on the same scale) over the entire OOS period (2014:M1-2017:M12). If a point exists to the right side of the line, then the corresponding combination forecasting models are outperformed by the benchmark, whereas points to the left signify the opposite. As we can see in all Figures 4.9 - 4.11, all combination forecasting models outperform the benchmark. For example, we note, for $h = 3$ in Figure 4.10, we see that not only do all of our forecast considerations outperform the benchmark model, but also that the relationship between the squared bias relationship as it relates to the forecast variance that highlights the superiority. For the L1PQFC model, a relatively significant decrease in squared bias of approximately 8% is gained at the expense of a 2.5% increase in forecast variance. Therefore, the extent to which squared bias can be decreased relative to a marginal increase in unconditional forecast variance contributes to improved forecast accuracy relative to our equal weight combination benchmark.

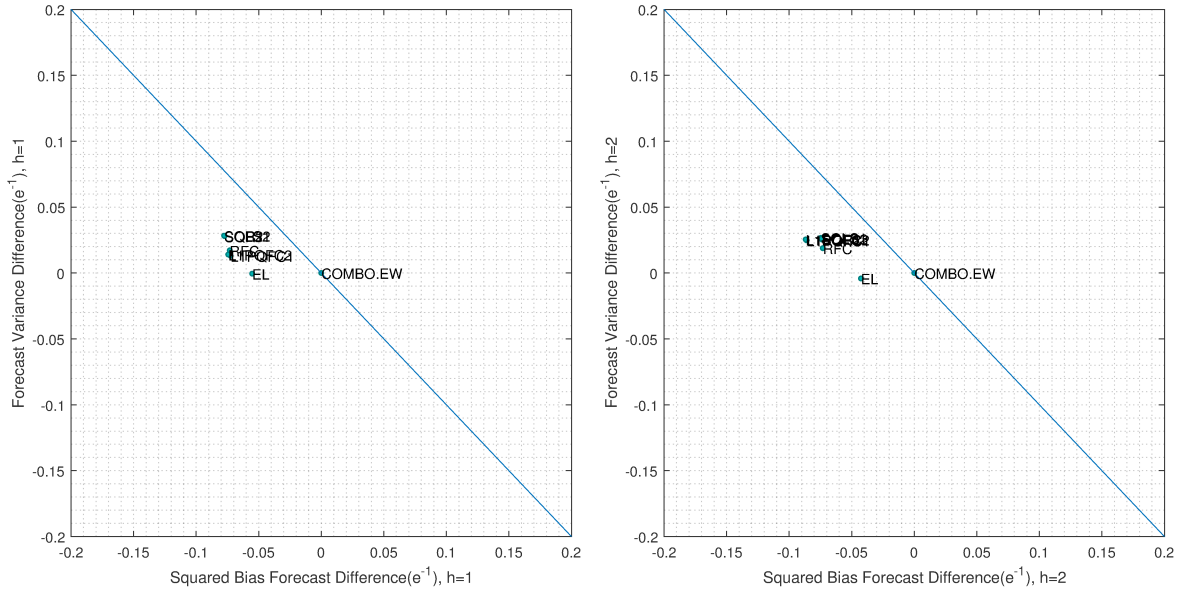


Figure 4.9: Bivariate display of squared bias forecast and unconditional forecast variance relative to the equal weight benchmark for horizons $h = \{1, 2\}$. Each axis represents the difference between the forecast variance [squared bias] of the proposed combination and the forecast variance [squared bias] of the equal weight benchmark. A point for our combination consideration that exist on the blue diagonal represents a combination forecast with the same MSPE as the equal weight combination.

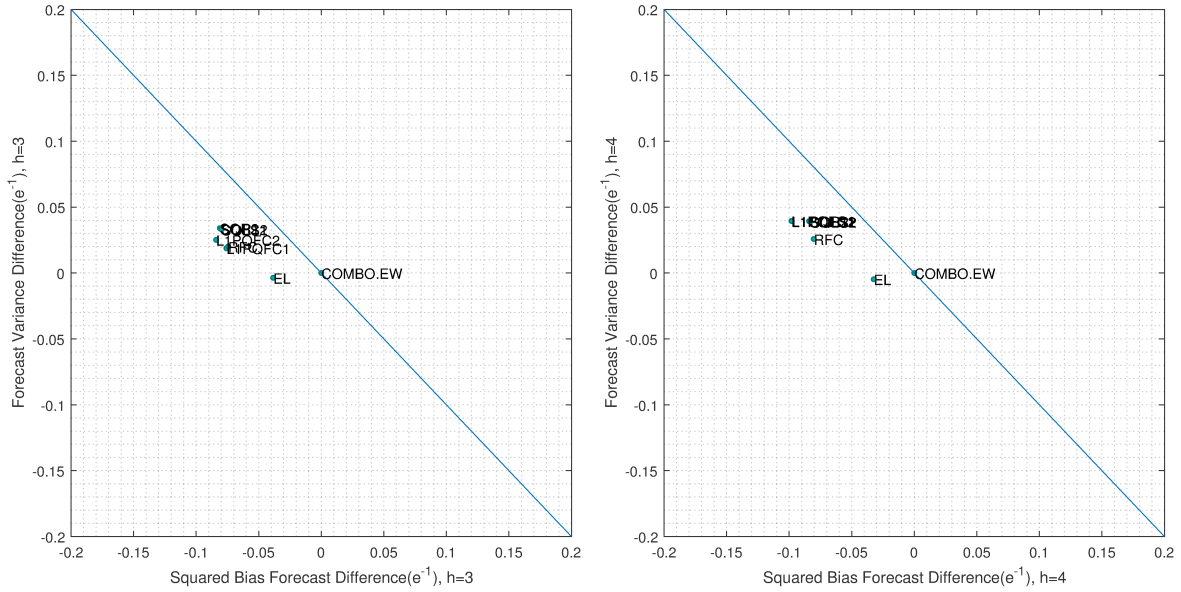


Figure 4.10: Bivariate display of squared bias forecast and unconditional forecast variance relative to the equal weight benchmark for horizons $h = \{3, 4\}$. Each axis represents the difference between the forecast variance [squared bias] of the proposed combination and the forecast variance [squared bias] of the equal weight benchmark. A point for our combination consideration that exist on the blue diagonal represents a combination forecast with the same MSPE as the equal weight combination.

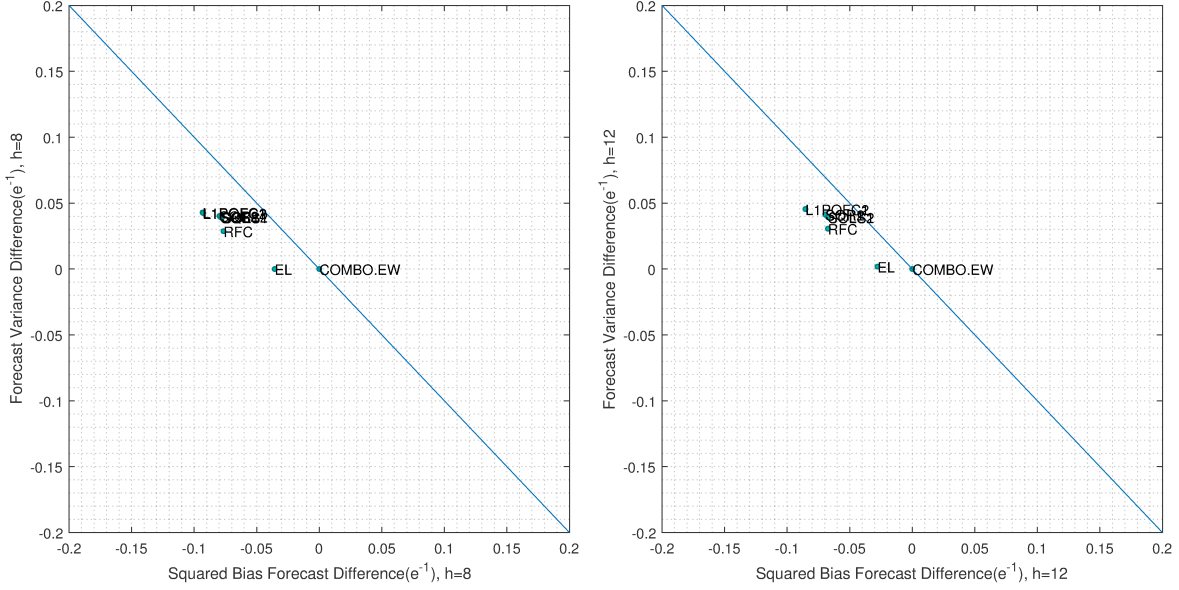


Figure 4.11: Bivariate display of squared bias forecast and unconditional forecast variance relative to the equal weight benchmark for horizons $h = \{8, 12\}$. Each axis represents the difference between the forecast variance [squared bias] of the proposed combination and the forecast variance [squared bias] of the equal weight benchmark. A point for our combination consideration that exist on the blue diagonal represents a combination forecast with the same MSPE as the equal weight combination.

As in [44], any forecasting combination consideration that results in a decrease in squared bias while mitigating minimally increasing variance will surely lead to improved forecast accuracy of an equal weight combination. We also report the frequency of models selected over the out-of-sample period of the L1 penalized quantile forecast combination method across all six forecast horizons. Table 4.4 shows the frequency with which each explanatory model used to compute the L1PLQFC forecast over the OOS period, 2014:M1 – 2017:M12, for $h = \{1, 2, 3, 4, 8, 12\}$ given τ . Recalling the L1PLQFC forecast example in Section 4.3.2, we defined the notion of a “fully weak”, “partially weak”, and “strong” underlying forecasting models in the population. For a given $\tau \in \tau$, if a forecasting model (the point forecasts of one of the univariate linear forecasting models) predicts no quantile, it is considered fully weak; yet if a forecasting model helps all quantiles it is considered “strong”. If a forecasting model helps predicts some, but not all quantiles in τ , then it is considered “partially weak”. Table 4.4 highlights varied observations regarding the most effective individual models composed in the L1PLQFC formulation across the different horizons. For all horizons, we have an assortment of strong, partially weak, and weak predictors across all forecast horizons with varying degrees of selection frequency. Most interestingly, we have at least one model considered as strong predictors across all forecast horizons. For completeness, we will explain the notion of the models’ “strength assessment” for the underlying forecasts for $h = 1$ of annualized monthly inflation gap used in our L1PLQFC1 model. The underlying models AR(1), AR(2), RW-AO, PC, and historical average models are considered “fully weak” as they do help to adequately predict our response variable at any level in τ .

Table 4.4: Frequency of models selected over OOS: January 2014 to December 2017

h=1									
τ	AR(1)	AR(2)	RW-AO	PC	PC-Ext	PC-BF	PC-BF-Ext	SBF	Hist-Avg
30th					31%		69%		
40th						15%	71%	15%	
50th						25%	33%	42%	
60th							100%		
70th						69%	100%		
h=2									
τ	AR(1)	AR(2)	RW-AO	PC	PC-Ext	PC-BF	PC-BF-Ext	SBF	Hist-Avg
30th				100%					
40th				83%		17%			
50th		2%		69%			29%		
60th				71%			29%		
70th		8%		65%			27%		
h=3									
τ	AR(1)	AR(2)	RW-AO	PC	PC-Ext	PC-BF	PC-BF-Ext	SBF	Hist-Avg
30th	8%		13%		100%		10%		100%
40th					100%		63%		
50th					100%		31%		
60th					88%		44%		
70th		44%	100%		13%		15%	27%	
h=4									
τ	AR(1)	AR(2)	RW-AO	PC	PC-Ext	PC-BF	PC-BF-Ext	SBF	Hist-Avg
30th					100%				21%
50th					100%		50%		
70th			8%		100%		100%		
h=8									
τ	AR(1)	AR(2)	RW-AO	PC	PC-Ext	PC-BF	PC-BF-Ext	SBF	Hist-Avg
30th		100%			100%		8%		
40th		100%			100%	25%	100%		
50th		100%			100%	58%	100%		
60th		81%			100%	19%	79%		
70th			65%		100%		92%		
h=12									
τ	AR(1)	AR(2)	RW-AO	PC	PC-Ext	PC-BF	PC-BF-Ext	SBF	Hist-Avg
30th	17%	35%	6%		100%				35%
40th	10%	17%			100%		94%		29%
50th	13%	63%			100%		83%		69%
60th	79%	4%	6%		100%		21%		94%
70th	85%	10%	83%	58%	100%		58%		88%

However, as PC-ext helps to predict $\tau = 0.30$ solely, it is considered to be “partially weak”, yet the PC-BF and SBF models exhibit partial weakness differently as they help to predict $\tau = (0.4, 0.5, 0.7)$ and $\tau = (0.4, 0.5)$, respectively. The PC-BF-Ext model exhibits its indispensability in our L1PLQFC1 model as it predicts across every quantile in τ (being selected at *every* month in our OOS epoch for annualized inflation gap when considering the 6th and 7th decile). For the last three horizons furthest out in time ($h = \{4, 8, 12\}$), the linear model PC-Ext was selected 100% of the time across the set of quantiles in a given τ . As the PC-Ext model is constructed via the use of ‘adaptive expectations’ of monthly annualized inflation gap and monthly annualized unemployment rate, these two individual variables seem to be extremely relevant in properly forecasting out-of-sample monthly inflation gap specifically in our penalized quantile consideration combination. Relative to the other univariate linear models considered in our L1PLQFC forecast combinations, it seems as if the AR(1), PC, RW-AO, and the SBF models are not quite as useful for forecasting the OOS monthly annualized inflation gap, being selected quite infrequently in our L1PLQFC combinations across horizons h . Lastly, regarding the utility of both the inclusion of partially weak models as well as the sheer breadth of the selection frequency of models, the point forecasts produced by our models 12 months ahead ($h = 12$) is ranked highest amongst all other horizons considered.

Finally, as a companion to the results presented in Table 4.4, we demonstrate how our proposed L1PQFC model selection procedure’s temporal effectiveness across the entire OOS period. Figures 4.12-4.17 detail which models are selected on a month-by-month basis across all relevant quantiles corresponding to the L1PQFC model. For each chart across all quantiles, we list the relevant models on the vertical axes, and the months corresponding to the OOS period on the horizontal axes. The indicator $+$ for each chart corresponds to the model selected at a specified month in time for forecasting each quantile. We will take a look at a specific case for one horizon. As mentioned previously, the L1PQFC1 model has the superior out-of-sample performance the point forecasts constructed from the various models for the $h = 4$ horizon. By observing Figure 4.15, we see that the PC-ext model has substantial predictive power for forecasting monthly annualized inflation gap across the entire OOS period 2014:M1 – 2017:M12 for $\tau = 0.3$. However, assessing the same quantile ($\tau = 0.3$), we observe that the historical average model has virtually no predictive power until 2016:M5 which lasts until 2016:M11 and only returns to predictive prominence by our selection procedure for 3 months in 2017. For $\tau = 0.7$ given $h = 4$, we note the predictive power for the PC-Ext and PC-BF-Ext via the L1-penalized selection procedure in across the entire span of the OOS period; however, the RW-AO model does have a brief spurt of significant predictive power for four consecutive months in 2015 (2015:M5-2015:M9). As the RW-AO model for $\tau = 0.7$ only exhibits predictive relevance for these specified quantiles and not the others, its partially weak classification as a forecaster in our L1PQFC2 combination procedure is clear. The other charts corresponding to covariates based off the remaining horizons for each quantiles can be interpreted the same way. Therefore, these figures can graphically show the relevance of partially weak predictors at a specific points in time in our OOS period.

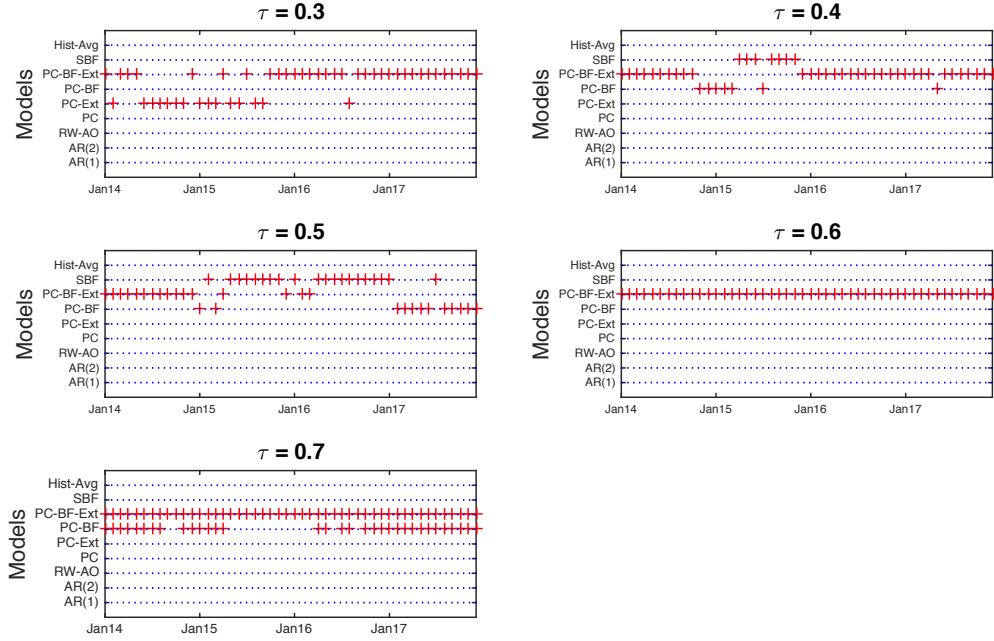


Figure 4.12: Models Selected: $h=1$

Models selected by PLQC for quantile levels : $\tau = (0.3, 0.4, 0.5, 0.6, 0.7)$ over OOS 2014:1â€¦2017:12. The five charts, one for each quantile used in the PLQC forecast, display the selected forecasters at each time point t over the out-of-sample period, 2014:M1 - 2017:M12 for $h=1$.

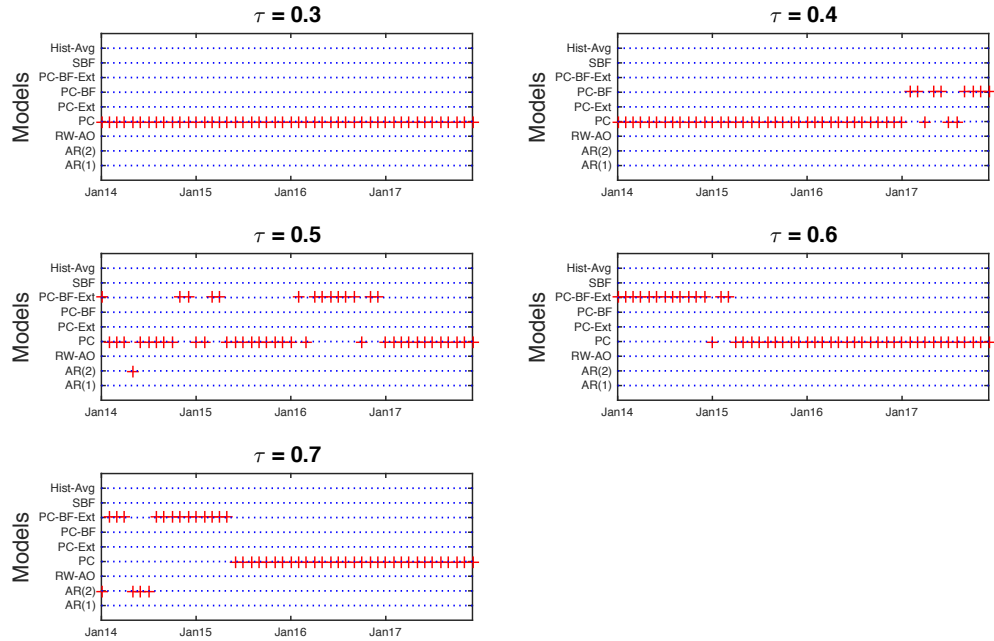


Figure 4.13: Models Selected: $h=2$

Models selected by PLQC for quantile levels : $\tau = (0.3, 0.4, 0.5, 0.6, 0.7)$ over OOS 2014:M1 - 2017:M12. The five charts, one for each quantile used in the L1PQFC2 forecast, display the selected forecasters at each time point t over the out-of-sample period, 2014:M1 - 2017:M12 for $h=2$.

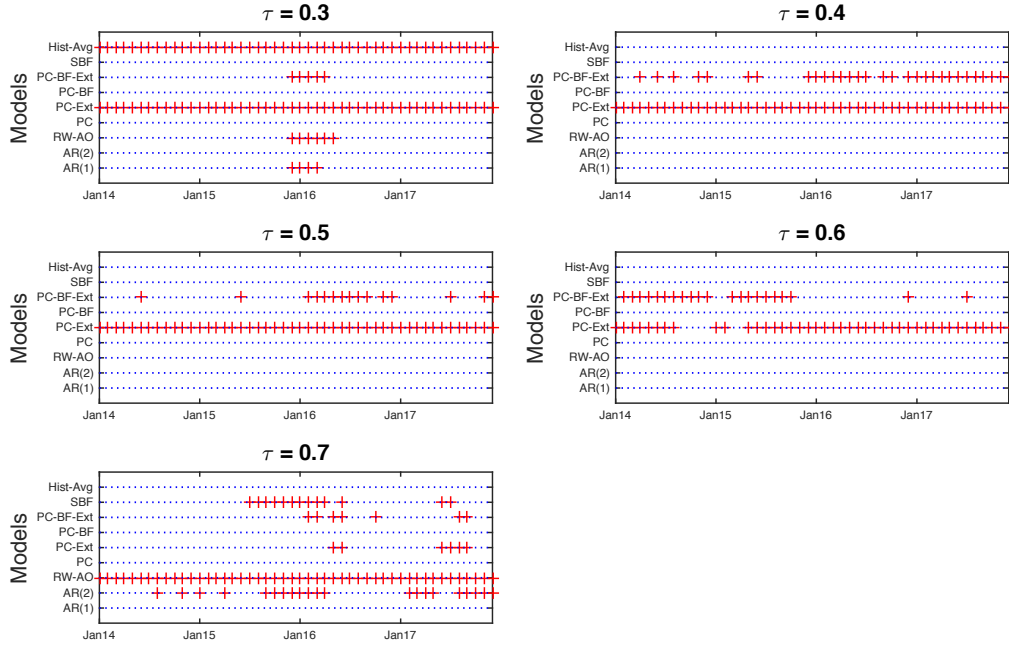


Figure 4.14: Models Selected: $h=3$

Models selected by PLQC for quantile levels : $\tau = (0.3, 0.4, 0.5, 0.6, 0.7)$ over OOS 2014:M1 - 2017:M12. The five charts, one for each quantile used in the L1PQFC2 forecast, display the selected forecasters at each time point t over the out-of-sample period, 2014:M1 - 2017:M12 for $h=3$.

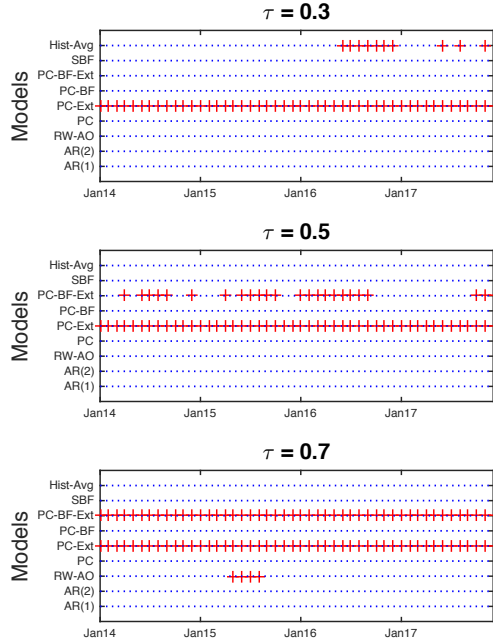


Figure 4.15: Models Selected: $h=4$

Models selected by PLQC for quantile levels : $\tau = (0.3, 0.5, 0.7)$ over OOS 2014:M1 - 2017:M12. The three charts, one for each quantile used in the L1PQFC1 forecast, display the selected forecasters at each time point t over the out-of-sample period, 2014:M1 - 2017:M12 for $h=4$.

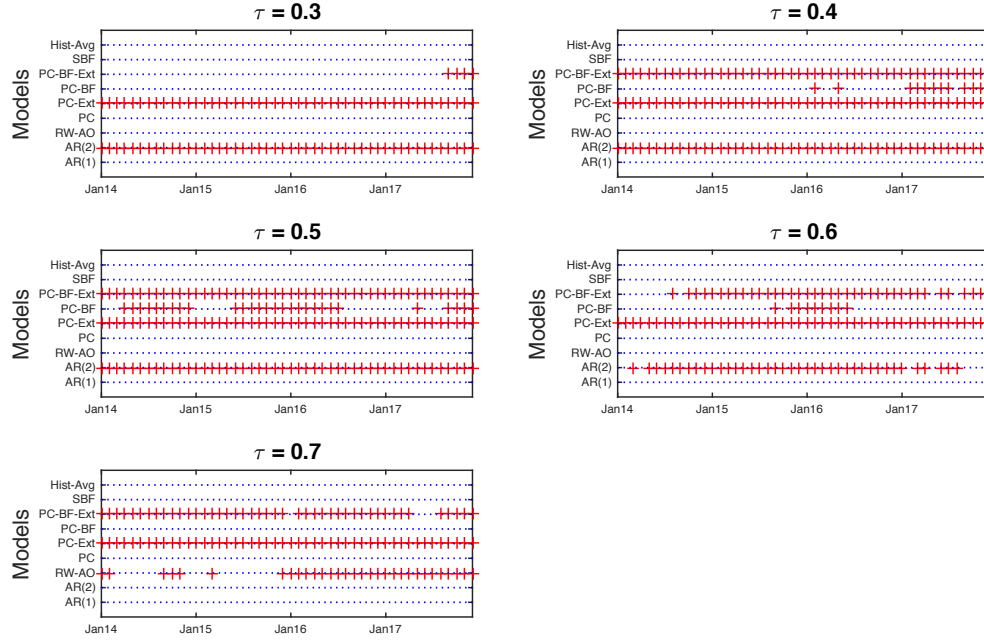


Figure 4.16: Models Selected: $h=8$

Models selected by PLQC for quantile levels : $\tau = (0.3, 0.4, 0.5, 0.6, 0.7)$ over OOS 2014:M1 - 2017:M12 . The five charts, one for each quantile used in the LIPQFC2 forecast, display the selected forecasters at each time point t over the out-of-sample period, 2014:M1 - 2017:M12 for $h=8$.

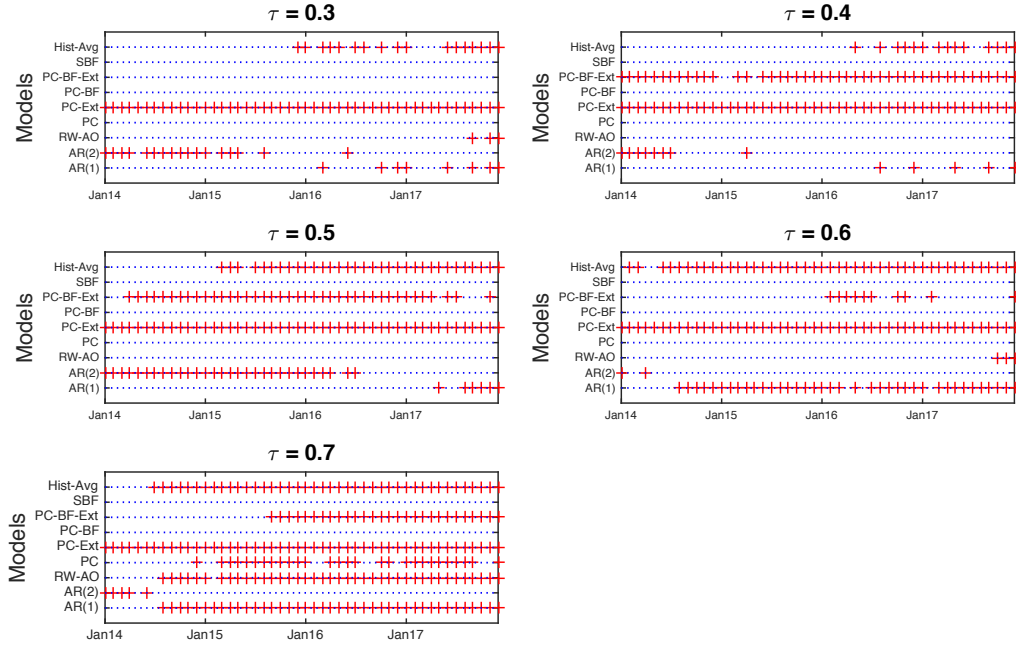


Figure 4.17: Models Selected: $h=12$

Models selected by PLQC for quantile levels : $\tau = (0.3, 0.4, 0.5, 0.6, 0.7)$ over OOS 2014:M1 - 2017:M12. The five charts, one for each quantile used in the LIPQFC2 forecast, display the selected predictor(s) at each time point t over the out-of-sample period, 2014:M1 - 2017:M12 for $h=12$.

4.6 Discussion

Summary: Inspired directly by [20], [26], and [24], we have used fact that the optimal conditional mean solution from Granger-Ramanathan’s proposed regression, under the MSE loss function, can be approximated by a combination of discrete quantiles in order to develop an optimal combination solution that now discards underlying forecasters that add little to any information in forecasting our target response of interest via a penalty. This chapter offers an in-depth analysis using the proposed forecasting method to study monthly annualized inflation gap using 248 observations spanning 1998:M1-2017:M12. The analyses consists of two major parts. In the first part of our analysis, we construct forecasts from nine univariate, linear forecasting models (in addition to an equal weight benchmark) over six varying horizons $h = \{1, 2, 3, 4, 8, 12\}$ for the monthly inflation gap at time $t+h$. In our next step, we construct a two variant, combination model selection procedures, the L1PQFC models, to assess the OOS forecasting performance of the monthly annualized inflation gap at time $t+h$ using point forecasts constructed under the previous six scenarios. The model selection procedure is used to identify strong, partially weak, and weak models which is then used to make MSPE forecasts based on an averaging process applied to arbitrary quantiles. The L1PQFC method is compared to various other combination procedures where each resulting combination model is compared to the previously constructed combination benchmark. The L1PQFC not only soundly outperforms our benchmark, but it is also super to the other combination procedures relative to the benchmark model based on various OOS diagnostics. We can attribute the superiority of out-of-sample performance from the L1PQFC procedure primarily to its ability to utilize partially weak predictors across all quantiles with the added benefit of circumventing estimation errors caused by outliers. In all cases, the L1PQFC is the top performer in all cases.

References

- [1] J. H. Stock and M. W. Watson, “Forecasting with many predictors,” *Handbook of economic forecasting*, vol. 1, pp. 515–554, 2006. [121](#), [125](#)
- [2] X. Yabo, W. Tong, S. Bingxue, and X. Yushu, “Forecasting of production safety situation by combination model,” *Procedia Engineering*, vol. 43, pp. 132–136, 2012. [121](#)
- [3] J. R. Trapero, M. Cardós, and N. Kourentzes, “Quantile forecast optimal combination to enhance safety stock estimation,” *International Journal of Forecasting*, vol. 35, no. 1, pp. 239–250, 2019. [121](#)
- [4] B. Men, R. Long, Y. Li, H. Liu, W. Tian, and Z. Wu, “Combined forecasting of rainfall based on fuzzy clustering and cross entropy,” *Entropy*, vol. 19, no. 12, p. 694, 2017. [121](#)
- [5] R. T. Clemen, “Combining forecasts: A review and annotated bibliography,” *International journal of forecasting*, vol. 5, no. 4, pp. 559–583, 1989. [121](#)
- [6] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969. [121](#), [122](#)
- [7] A. Timmermann, “Forecast combinations,” *Handbook of economic forecasting*, vol. 1, pp. 135–196, 2006. [121](#), [125](#)
- [8] D. W. Bunn, “Statistical efficiency in the linear combination of forecasts,” *International Journal of Forecasting*, vol. 1, no. 2, pp. 151–163, 1985. [121](#)
- [9] F. X. Diebold and P. Pauly, “Structural change and the combination of forecasts,” *Journal of Forecasting*, vol. 6, no. 1, pp. 21–40, 1987. [121](#)
- [10] F. X. Diebold and P. Pauly, “The use of prior information in forecast combination,” *International Journal of Forecasting*, vol. 6, no. 4, pp. 503–508, 1990. [121](#), [125](#)
- [11] D. N. Sessions and S. Chatterjee, “The combining of forecasts using recursive techniques with non-stationary weights,” *Journal of Forecasting*, vol. 8, no. 3, pp. 239–251, 1989. [121](#), [125](#)
- [12] L. M. De Menezes, D. W. Bunn, and J. W. Taylor, “Review of guidelines for the use of combined forecasts,” *European Journal of Operational Research*, vol. 120, no. 1, pp. 190–204, 2000. [121](#)
- [13] Y. Yang, “Combining forecasting procedures: some theoretical results,” *Econometric Theory*, vol. 20, no. 1, pp. 176–222, 2004. [121](#)
- [14] W. P. Gaglianone and J. V. Issler, “Microfounded forecasting,” *Ensaaios Econômicos*; 766, 2015. [121](#), [127](#), [159](#)

- [15] S. Simkins, “Forecasting with vector autoregressive (var) models subject to business cycle restrictions,” *International Journal of Forecasting*, vol. 11, no. 4, pp. 569–583, 1995. [122](#)
- [16] B. Wild, M. Eichler, H.-C. Friederich, M. Hartmann, S. Zipfel, and W. Herzog, “A graphical vector autoregressive modelling approach to the analysis of electronic diary data,” *BMC medical research methodology*, vol. 10, no. 1, p. 28, 2010. [122](#)
- [17] F. Ögünç, K. Akdoğan, S. Başer, M. G. Chadwick, D. Ertuğ, T. Hülügü, S. Kösem, M. U. Özmen, and N. Tekath, “Short-term inflation forecasting models for turkey and a forecast combination analysis,” *Economic Modelling*, vol. 33, pp. 312–325, 2013. [122](#)
- [18] K. Hubrich and F. Skudelny, “Forecast combination for euro area inflation: a cure in times of crisis?,” *Journal of Forecasting*, vol. 36, no. 5, pp. 515–540, 2017. [122](#)
- [19] C. Jiang, J. Zhang, and F. Song, “Selecting single model in combination forecasting based on cointegration test and encompassing test,” *The Scientific World Journal*, vol. 2014, 2014. [122](#)
- [20] C. W. Granger and R. Ramanathan, “Improved methods of combining forecasts,” *Journal of Forecasting*, vol. 3, no. 2, pp. 197–204, 1984. [122](#), [123](#), [126](#), [127](#), [154](#)
- [21] G. G. Judge, “Introduction to the theory and practice of econometrics,” [122](#), [129](#)
- [22] J. W. Taylor, “Forecasting daily supermarket sales using exponentially weighted quantile regression,” *European Journal of Operational Research*, vol. 178, no. 1, pp. 154–167, 2007. [122](#)
- [23] L. Ma and L. Pohlman, “Return forecasts and optimal portfolio construction: a quantile regression approach,” *The European Journal of Finance*, vol. 14, no. 5, pp. 409–425, 2008. [122](#)
- [24] L. R. Lima and F. Meng, “Out-of-sample return predictability: A quantile combination approach,” *Journal of Applied Econometrics*, vol. 32, no. 4, pp. 877–895, 2017. [122](#), [123](#), [124](#), [129](#), [131](#), [154](#), [160](#)
- [25] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [123](#), [160](#)
- [26] A. Belloni, V. Chernozhukov, *et al.*, “ λ_1 -penalized quantile regression in high-dimensional sparse models,” *The Annals of Statistics*, vol. 39, no. 1, pp. 82–130, 2011. [123](#), [130](#), [154](#), [160](#)
- [27] J. Faust and J. H. Wright, “Forecasting inflation,” in *Handbook of economic forecasting*, vol. 2, pp. 2–56, Elsevier, 2013. [123](#), [134](#)
- [28] A. Timmermann, *Forecast Combinations*, vol. 1 of *Handbook of Economic Forecasting*, ch. 4, pp. 135–196. Elsevier, February 2006. [124](#), [125](#)

- [29] W. P. Gaglianone and L. R. Lima, “Constructing optimal density forecasts from point forecast combinations,” *Journal of Applied Econometrics*, vol. 29, no. 5, pp. 736–757, 2014. [124](#), [126](#), [127](#), [134](#), [135](#), [159](#)
- [30] J. M. Bates and C. W. Granger, “The combination of forecasts,” *Journal of the Operational Research Society*, vol. 20, no. 4, pp. 451–468, 1969. [124](#)
- [31] M. Aiolfi and A. Timmermann, “Persistence in forecasting performance and conditional combination strategies,” *Journal of Econometrics*, vol. 135, no. 1-2, pp. 31–53, 2006. [125](#)
- [32] A. Timmermann, “Forecast combinations,” 2005. [125](#), [136](#)
- [33] T.-H. Lee, “Combining forecasts with many predictors,” [125](#)
- [34] R. F. Engle, C. W. Granger, and D. Kraft, “Combining competing forecasts of inflation using a bivariate arch model,” *Journal of economic dynamics and control*, vol. 8, no. 2, pp. 151–165, 1984. [125](#)
- [35] C. Granger and P. Newbold, “Forecasting economic time series,” tech. rep., Elsevier, 1986. [126](#)
- [36] A. J. Patton and A. Timmermann, “Properties of optimal forecasts under asymmetric loss and nonlinearity,” *Journal of Econometrics*, vol. 140, no. 2, pp. 884–918, 2007. [126](#), [159](#)
- [37] F. Peracchi, “On estimating conditional quantiles and distribution functions,” *Computational statistics & data analysis*, vol. 38, no. 4, pp. 433–447, 2002. [127](#)
- [38] R. Koenker, S. Leorato, and F. Peracchi, “Distributional vs. quantile regression,” 2013. [127](#)
- [39] R. Koenker, A. Chesher, and M. Jackson, *Quantile Regression*. Econometric Society Monographs, Cambridge University Press, 2005. [129](#)
- [40] L. Meligkotsidou, E. Panopoulou, I. D. Vrontos, and S. D. Vrontos, “A quantile regression approach to equity premium prediction,” *Journal of Forecasting*, vol. 33, no. 7, pp. 558–576, 2014. [129](#), [133](#)
- [41] J. V. Issler and L. R. Lima, “A panel data approach to economic forecasting: The bias-corrected average forecast,” *Journal of Econometrics*, vol. 152, no. 2, pp. 153–164, 2009. [130](#)
- [42] V. Fonti and E. Belitser, “Feature selection using lasso,” 2017. [130](#)
- [43] S. Manzan, “Forecasting the distribution of economic variables in a data-rich environment,” *Journal of Business & Economic Statistics*, vol. 33, no. 1, pp. 144–164, 2015. [130](#), [160](#)
- [44] W. Lin and J. M. Wooldridge, “Testing and correcting for endogeneity in nonlinear unobserved effects models,” 2017. [132](#), [148](#)
- [45] J. H. Stock and M. W. Watson, “Forecasting inflation,” *Journal of Monetary Economics*, vol. 44, no. 2, pp. 293–335, 1999. [134](#)

- [46] R. Duncan and E. Martínez-García, “New perspectives on forecasting inflation in emerging market economies: An empirical assessment,” *New Perspectives on Forecasting Inflation in Emerging Market Economies: An Empirical Assessment (2018-01-01)*. Globalization and Monetary Policy Institute Working Paper, no. 338, 2018. [134](#)
- [47] W. E. Weber, L. E. Ohanian, and J. A. Schmitz Jr, “The quarterly review is published by the research department of the federal reserve bank of minneapolis. subscriptions are available free of charge. quarterly review articles that are reprints or revisions of papers published elsewhere may not be reprinted without the written,” *Quarterly Review*, vol. 23, no. 1, 1999. [135](#)
- [48] F. Bliemel, “Theil’s forecast accuracy coefficient: A clarification,” *Journal of Marketing Research*, vol. 10, no. 4, pp. 444–446, 1973. [137](#)
- [49] F. C. Palm and A. Zellner, “To combine or not to combine? issues of combining forecasts,” *Journal of Forecasting*, vol. 11, no. 8, pp. 687–701, 1992. [139](#)
- [50] F. X. Diebold and M. Shin, “Egalitarian lasso for combining economic forecasts,” 2018. [140](#), [145](#)
- [51] D. E. Rapach, J. K. Strauss, and G. Zhou, “Out-of-sample equity premium prediction: Combination forecasts and links to the real economy,” *The Review of Financial Studies*, vol. 23, no. 2, pp. 821–862, 2010. [146](#)
- [52] P. F. Christoffersen and F. X. Diebold, “Optimal prediction under asymmetric loss,” *Econometric theory*, vol. 13, no. 6, pp. 808–817, 1997. [159](#)
- [53] F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” *Journal of Business & economic statistics*, vol. 20, no. 1, pp. 134–144, 2002. [162](#)
- [54] A. Thomadakis, “Do combination forecasts outperform the historical average? economic and statistical evidence,” 2016. [162](#)

Appendix

Proof of Proposition 1

This is similar to that shown by [52],[36], [29], and [14]. We, first, seek to find the form of the optimal predictor that solves the combination problem below. Given the data generating process in (1) as well as the homogeneity of the loss function of the h-step ahead prediction/forecast error $e_{t+h,t}$, we have:

$$\begin{aligned}
\hat{y}_{t+h,t}^i &= \arg \min_{\hat{y}} \int_{-\infty}^{\infty} L^i(y - \hat{y}) dF_{t+h,t}(y) \\
&= \arg \min_{\hat{y}} \int_{-\infty}^{\infty} \left[g\left(\frac{1}{\mathbf{y}_{t+h,t}' \boldsymbol{\phi}}\right) \right]^{-1} L^i\left(\frac{1}{\mathbf{y}_{t+h,t}' \boldsymbol{\phi}}(y - \hat{y})\right) dF_{t+h,t}(y) \\
&= \arg \min_{\hat{y}} \int_{-\infty}^{\infty} \left[g\left(\frac{1}{\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j}\right) \right]^{-1} L^i\left(\frac{1}{\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j}(y - \hat{y})\right) dF_{t+h,t}(y) \\
&= \arg \min_{\hat{y}} \int_{-\infty}^{\infty} L^i\left(\frac{1}{\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j}(y - \hat{y})\right) dF_{t+h,t}(y) \\
&= \arg \min_{\hat{y}} \int_{-\infty}^{\infty} L^i\left(\frac{\beta_0 + \sum_{j=1}^{k-1} \beta_j \hat{y}_{t+h,t}^j + \eta_{t+h}\left(\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j\right) - \hat{y}}{\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j}\right) dF_{t+h,t}(y)
\end{aligned}$$

Let us define a forecast as for an arbitrary horizon as : $\hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\beta} + (\hat{\mathbf{y}}_{t+h,t}' \boldsymbol{\phi}) \phi_h^i = \beta_0 + \beta_1 \hat{y}_{t+h,t}^1 + \dots + \beta_{k-1} \hat{y}_{t+h,t}^{k-1} + (\phi_0 + \phi_1 \hat{y}_{t+h,t}^1 + \dots + \phi_{k-1} \hat{y}_{t+h,t}^{k-1}) \hat{\phi}_{t+h,t}^i$

Then it follows that

$$\begin{aligned}
\hat{y}_{t+h,t}^i &= \beta_0 + \beta_1 \hat{y}_{t+h,t}^1 + \dots + \beta_{k-1} \hat{y}_{t+h,t}^{k-1} + (\phi_0 + \phi_1 \hat{y}_{t+h,t}^1 + \dots + \phi_{k-1} \hat{y}_{t+h,t}^{k-1}) \hat{\phi}_{t+h,t}^i. \\
\arg \min_{\hat{\phi}} \int_{-\infty}^{\infty} L^i\left(\frac{\beta_0 + \sum_{j=1}^{k-1} \beta_j \hat{y}_{t+h,t}^j + \eta_{t+h}\left(\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j\right)}{\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j} - \frac{\left(\beta_0 + \sum_{j=1}^{k-1} \beta_j \hat{y}_{t+h,t}^j\right) + \left(\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j\right) \hat{\phi}}{\phi_0 + \sum_{j=1}^{k-1} \phi_j \hat{y}_{t+h,t}^j}\right) dF_{\eta,h}(\eta) \\
&= \beta_0 + \beta_1 \hat{y}_{t+h,t}^1 + \dots + \beta_{k-1} \hat{y}_{t+h,t}^{k-1} + (\phi_0 + \phi_1 \hat{y}_{t+h,t}^1 + \dots + \phi_{k-1} \hat{y}_{t+h,t}^{k-1}) \arg \min_{\hat{\phi}} \int_{-\infty}^{\infty} L^i(\eta_{t+h} - \hat{\phi}) dF_{\eta,h}(\eta) \\
&= \beta_0 + \phi_0 \phi_h^i + \beta_1 \hat{y}_{t+h,t}^1 + \dots + \beta_{k-1} \hat{y}_{t+h,t}^{k-1} + \phi_1 \hat{y}_{t+h,t}^1 \phi_h^i + \dots + \phi_{k-1} \hat{y}_{t+h,t}^{k-1} \phi_h^i \\
&= \beta_0 + \phi_0 \phi_h^i + \beta_1 \hat{y}_{t+h,t}^1 + \phi_1 \hat{y}_{t+h,t}^1 \phi_h^i + \dots + \beta_{k-1} \hat{y}_{t+h,t}^{k-1} + \phi_{k-1} \hat{y}_{t+h,t}^{k-1} \phi_h^i \\
&= (\beta_0 + \phi_0 \phi_h^i) + (\beta_1 + \phi_1 \phi_h^i) \hat{y}_{t+h,t}^1 + \dots + (\beta_{k-1} + \phi_{k-1} \phi_h^i) \hat{y}_{t+h,t}^{k-1} \\
&= \gamma_{i,0} + \gamma_{i,1} \hat{y}_{t+h,t}^1 + \dots + \gamma_{i,k-1} \hat{y}_{t+h,t}^{k-1}
\end{aligned}$$

Finding λ for our proposed L1PQFC

Note, the ℓ_1 -norm is the regularization technique following the principal contribution of [25]. As given in [26] and [43], where [26] uses this method in dimension reduction for deriving quantile regression estimators in high-dimensional sparse settings, this LASSO method will first select the individual forecasting models from our information in $\hat{\mathbf{y}}_{t+h,t}$ from our conditional models for each specified quantile at each time period t . We briefly highlight the method below.

For notational simplicity, let x_{ij} represents a covariate in row i for column covariate j . For λ (the penalty), a random variable is defined as Λ as

$$\Lambda = n \sup_{u \in \mathcal{U}} \max_{1 \leq j \leq p} \left| \mathbb{E} \left[\frac{x_{ij}(u - 1\{u_j \leq u\})}{\hat{\sigma}_j \sqrt{u(1-u)}} \right] \right|$$

where $u_1, \dots, u_n$ are i.i.d. $U(0, 1)$ random variables which are independent and identically distributed from the covariates, and $\hat{\sigma}_j = \sqrt{\mathbb{E}_n[x_{ij}^2]}$ for the j th covariate. This random variable has a pivotal distribution conditional on the covariates $X = [x_1, \dots, x_n]$. Setting

$$\lambda = c \cdot \Lambda(1 - \alpha|X),$$

where $\Lambda(1 - \alpha|X) := (1 - \alpha)$ -quantile of Λ conditional on X , with $1 - \alpha = 0.95$ and $c = 2$, we find the proper choice of the penalty λ .

From a programming perspective given [26] and [24], we employ a function that computes the empirical distribution of a pivotal quantity (Λ) for the choice of λ which takes as inputs 1.) a given τ , 2.) the matrix of covariates at time $t - 1$, and 3.) an arbitrary number of trials. Then, a λ is generated recursively for each τ at each time t to aid in computing the L1-penalized quantile forecasts at time t , resulting in post-LASSO quantile forecasts based on the selected predictors (e.g. post-LASSO penalization).

Relationship between LASSO and Egalitarian LASSO

We will now show how the Egalitarian LASSO objective function can be adapted into the R software. Recall the objective function for the Egalitarian LASSO utilized in our combination consideration is as follows

$$\text{EL} = \arg \min_{\beta} \left(\sum_t \left(y_{t+h} - \sum_{i=1} \beta_i \hat{y}_{t+h,t}^i \right)^2 + \lambda \sum_{i=1}^K \left| \beta_i - \frac{1}{K} \right| \right)$$

We modify the minimization objection function as follows

$$\begin{aligned}
\sum_{t=1}^T \left(y_{t+h} - \sum_{i=1}^K \beta_i \hat{y}_{t+h,t}^i \right)^2 + \lambda \sum_{i=1}^K \left| \beta_i - \frac{1}{K} \right| &= \sum_{t=1}^T \left(y_{t+h} - \bar{y}_t + \bar{y}_t - \sum_{i=1}^K \beta_i \hat{y}_{t+h,t}^i \right)^2 + \lambda \sum_{i=1}^K \left| \beta_i - \frac{1}{K} \right| \\
&= \sum_{t=1}^T \left((y_{t+h} - \bar{y}_t) + \sum_{i=1}^K \left(\frac{1}{K} - \beta_i \right) \hat{y}_{t+h,t}^i \right)^2 + \lambda \sum_{i=1}^K \left| \beta_i - \frac{1}{K} \right| \\
&= \sum_{t=1}^T \left((y_{t+h} - \bar{y}_t) + \sum_{i=1}^K \delta_i \hat{y}_{t+h,t}^i \right)^2 + \lambda \sum_{i=1}^K |\delta_i|
\end{aligned}$$

where $\delta_i = \beta_i - \frac{1}{K}$ and $\bar{y}_t = \frac{1}{K} \sum_{i=1}^K \hat{y}_{t+h,t}^i$

Therefore, Egalitarian LASSO for model selection is obtained via the EL regression of y_{t+h} on $(\hat{y}_{t+h,t}^1, \dots, \hat{y}_{t+h,t}^K)$ by running a standard LASSO of $(y_{t+h} - \bar{y}_t)$ on $(\hat{y}_{t+h,t}^1, \dots, \hat{y}_{t+h,t}^K)$.

OOS diagnostics

Here is an example of the OOS accuracy diagnostics. We define the accuracy diagnostics as:

- ME (Mean Error): $ME = \frac{\sum_{t=1}^n (y_t - \hat{y}_{t+h})}{n}$
- MSE (Mean Squared Error): $MSE = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_{t+h})^2$
- RMSE (Root Mean Squared Error): $RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_{t+h})^2}$
- MAE (Mean Absolute Error): $MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_{t+h}|$
- MPE (Mean Percentage Error): $MPE = \frac{100\%}{n} \frac{\sum_{t=1}^n (y_{t+h} - \hat{y}_t)}{y_t}$
- MAPE (Mean Absolute Percentage Error): $MAPE = \frac{100\%}{n} \left| \frac{\sum_{t=1}^n (y_{t+h} - \hat{y}_t)}{y_t} \right|$

Table 4.5: Out of Sample Accuracy Metrics

	h=1							h=4					
	ME	RMSE	MSE	MAE	MPE	MAPE		ME	RMSE	MSE	MAE	MPE	MAPE
AR(1)	0.01604678	0.71710779	0.51424359	0.50075543	84.440619	111.831847	AR(1)	0.02489447	1.16334215	1.35336496	0.80579085	105.1253	107.223695
AR(2)	0.00806693	0.43723929	0.1911782	0.30878879	25.5722786	160.076448	AR(2)	0.03290495	1.10130996	1.21288362	0.7369386	170.701849	230.399029
RW-AO	0.02291933	0.88200291	0.77792914	0.59778441	163.32664	289.865348	RW-AO	0.04896448	1.4697161	2.16006541	1.01022429	160.509355	400.357577
PC	0.01879566	0.50941369	0.25950231	0.34870634	77.3901767	155.572991	PC	0.02427197	0.52544803	0.27609563	0.35241913	79.5074575	155.386357
PC-Ext	-0.0056905	0.44444143	0.19752818	0.30992473	57.1555905	165.52416	PC-Ext	-0.0025703	0.45977947	0.21139716	0.3228605	60.2728369	162.95237
PC-BF	0.04468001	0.51278264	0.26294604	0.35146353	53.8354805	165.572528	PC-BF	0.04805436	0.53526526	0.2865089	0.35847771	63.8500823	163.422157
PC-BF-Ext	0.04117137	0.51381526	0.26400612	0.35651561	67.3078574	170.656421	PC-BF-Ext	0.04663519	0.53481443	0.28602647	0.35995497	75.9600335	166.54673
SBF	-0.0304451	1.16063306	1.3470691	0.8051073	130.02432	136.864928	SBF	-0.0204919	1.17352788	1.37716769	0.81730794	124.977759	129.792903
Hist Avg	0.0191897	1.16536361	1.35807235	0.81128629	100.490317	105.521848	Hist Avg	0.01905515	1.18444471	1.40290928	0.82513345	100.069327	106.385855
Combo EW	0.01497047	0.58781566	0.34552725	0.40315405	84.3936976	130.61807	Combo EW	0.02463538	0.78704049	0.61943273	0.53507791	104.552667	139.562823
	h=2							h=8					
	ME	RMSE	MSE	MAE	MPE	MAPE		ME	RMSE	MSE	MAE	MPE	MAPE
AR(1)	0.02270793	1.00568916	1.0114107	0.69624867	115.555803	125.248867	AR(1)	0.02260985	1.19198559	1.42082964	0.83289904	97.1686277	110.183607
AR(2)	0.01734455	0.76083449	0.57886912	0.53411631	151.312326	256.908805	AR(2)	0.05075972	1.24256694	1.54397259	0.87240212	60.7995898	171.62149
RW-AO	0.03280579	1.12316014	1.2614887	0.7636045	195.920995	362.718803	RW-AO	0.06366666	1.86340682	3.47228498	1.31067618	433.08752	678.589818
PC	0.02058464	0.5216185	0.27208586	0.35137872	78.6007245	156.277036	PC	0.02189056	0.52849755	0.27930966	0.35687205	79.7042951	154.26533
PC-Ext	-0.0054459	0.45329173	0.20547339	0.3166534	58.8547531	164.054508	PC-Ext	-0.0019954	0.4631211	0.21448115	0.3325456	61.9854098	166.585027
PC-BF	0.04617551	0.53124557	0.28222185	0.35814696	57.5801551	166.325065	PC-BF	0.037522	0.54315249	0.29501462	0.36487622	70.0442099	157.20352
PC-BF-Ext	0.0430593	0.53167738	0.28268083	0.36211069	70.3127566	171.233243	PC-BF-Ext	0.04007425	0.5431413	0.29500247	0.36763946	82.3013501	162.519517
SBF	-0.0266961	1.16604366	1.35965781	0.81012235	128.700666	134.939367	SBF	-0.0125891	1.17881558	1.38960618	0.82427138	119.95654	121.341529
Hist Avg	0.01931303	1.17328591	1.37659982	0.81696973	100.835746	106.134499	Hist Avg	0.01671517	1.19021981	1.41662319	0.83181474	98.7855201	108.864926
Combo EW	0.01887208	0.68667771	0.47152628	0.46681696	106.408214	141.357023	Combo EW	0.02651709	0.8642781	0.74697663	0.60270551	122.64802	146.344133
	h=3							h=12					
	ME	RMSE	MSE	MAE	MPE	MAPE		ME	RMSE	MSE	MAE	MPE	MAPE
AR(1)	0.02471601	1.11901125	1.25218618	0.77442344	106.368505	116.555272	AR(1)	0.01919309	1.18265386	1.39867015	0.82837615	92.6187981	114.671454
AR(2)	0.02564158	0.97406346	0.94879963	0.67309594	88.3762865	294.949306	AR(2)	0.05393818	1.21854645	1.48485545	0.8705261	95.0719443	129.410629
RW-AO	0.0417721	1.31440748	1.72766703	0.89807676	186.359873	390.483225	RW-AO	0.06759811	1.87970507	3.53329116	1.35170169	423.563881	733.628309
PC	0.02279033	0.52433894	0.27493132	0.3516711	79.0669856	156.650643	PC	0.01259583	0.51539057	0.26562744	0.3510367	80.6820918	149.00498
PC-Ext	-0.0039962	0.45773509	0.20952141	0.31959498	58.0687467	165.966988	PC-Ext	-0.0035937	0.45859412	0.21030856	0.33412459	64.9732883	167.874438
PC-BF	0.04768641	0.53468755	0.28589078	0.35895057	60.5382316	166.15737	PC-BF	0.01811138	0.5278424	0.2786176	0.35911735	73.1539322	149.713756
PC-BF-Ext	0.04531497	0.53449331	0.28568309	0.36246749	72.3012116	170.271823	PC-BF-Ext	0.02307068	0.52926067	0.28011685	0.3636112	86.683407	153.752047
SBF	-0.0233765	1.17027209	1.36953676	0.81407549	126.770199	132.429825	SBF	-0.0089365	1.17261737	1.37503149	0.82128607	110.127577	113.05331
Hist Avg	0.01927548	1.17963269	1.39153329	0.82147826	100.4664	106.320917	Hist Avg	0.01289238	1.17843801	1.38871613	0.82485157	92.3048534	114.009534
Combo EW	0.02220269	0.7476392	0.55896437	0.50794835	97.5907155	154.644268	Combo EW	0.02165216	0.85330621	0.7281315	0.59975121	124.353308	140.03078

Diebold-Mariano (DM) Test: Non-nested models

We follow [53]. Assume two competing point forecasts for two models (indexed by $\{i, j\}$) with associated sets forecast errors $(\epsilon_{i1}, \dots, \epsilon_{iT})$ and $(\epsilon_{j1}, \dots, \epsilon_{jT})$ up to time T ; the forecast quality is evaluated at an arbitrary loss function $h(\cdot)$ of the error of the forecasts. The null hypothesis for homogeneous predictive accuracy is $H_0 : E[d_t] = 0$, for all t , with $d_t = h(\epsilon_{iT}) - h(\epsilon_{jT})$. Note, $\frac{\bar{d}}{\sqrt{\hat{v}/T}}$ has an limiting Gaussian distribution (e.g. $N \sim (0, 1)$) under H_0 under given a set of regularity conditions. Note, $\bar{d}$ is the sample mean of the loss differential series $d_1, \dots, d_T$, where $\hat{v}$ is a consistent estimator of the asymptotic variance of $\sqrt{T}\bar{d}$.

Adjusted Clark and West Test

We follow [54]. Given the forecasts of annualized monthly inflation gap, y_{t+h} and the forecasts from the equal weight combination benchmark using predictors across all horizons, $\bar{y}_{t+h,t}$, the predictive combination models using predictors across all horizons, $\hat{r}_{t+h,t}^i$, we can define the adjusted MSPE statistic, for our 6 scenarios/horizons, as

$$z_{t+h} = (y_{t+h} - \bar{y}_{t+h,t})^2 - [(y_{t+h} - \hat{r}_{t+h,t}^i)^2 - (\bar{y}_{t+h,t} - \hat{r}_{t+h,t}^i)^2] \quad i = 1, \dots, 8 \quad h = \{1, 2, 3, 4, 8, 12\}$$

The additional term $(\bar{y}_{t+h,t} - \hat{r}_{t+h,t}^i)^2$ is an “adjustment term” that adjusts for the upward bias in MSPE by estimating parameters that are necessarily zero under the under $H_0 : R_{OS}^2 \leq 0$ as $[(y_{t+h} - \hat{r}_{t+h,t}^i)^2 -$

adjustment] $< (y_{t+h} - \hat{r}_{t+h,t}^i)^2$. Thus, the adjusted MSPE is the t-statistic corresponding to a zero constant with a one-sided p-value corresponding to the alternative $H_A : R_{OS}^2 > 0$ is obtained under the standard normal distribution.

CHAPTER 5

Conclusion

In concluding my three-chapter dissertation, I will now end by offering various limitations and opportunities for future avenues of research aligning with my contributions.

Chapter 2: Limitations/Avenues for future research

Although the results presented from our robust, idiosyncratic endogeneity test for nonlinear unbalanced panels were written exclusively for one endogenous binary covariate, y_{it2} , these results should be easily intuitively extendable to a vector of non-continuous endogenous covariates $\mathbf{y}_{it2}$, provided that one has appropriate relevant and excluded time-varying and time-varying instrumental variables. There likely would be no restrictions on the form on any of the endogeneous variables present in $\mathbf{y}_{it2}$ (e.g. a mix of continuous, binary, or discrete). A direct extension to a multinomial endogenous regressor (or vector of regressors) could potentially be explored as well. Additionally, although our Monte Carlo simulation designs are very convincing across all principal metrics of interest, we did not explore a rigorous power analysis as it relates to manipulating sample size (nor arbitrary levels of α), nor negative values of induced correlation representing the endogeneity in terms of assessing rejection rates. Regardless, we would expect results in line with our previous results. Also, there exists some discussion in the previous decade regarding the choice of simulating instrument strength for the indicator function forming the binary endogenous regressor in terms of directly specifying “concentration parameters” for the instrument strength versus directly specifying the total proportion of variance in the indicator function that relates the variance of the instrument vector and associated coefficients relative to the variance of the idiosyncratic error ([1]). Inspired by the work of [2],[3], and [4], in assessing endogeneity cross-sectionally for nonlinear conditional mean functions, these authors opt for a concentration parameter approach as I have for instrument coefficients in their simulations, which follows through without consequence for nonlinear structural equations. An interesting next step would be to compare both approaches of instrument strength manipulation in finite sample sizes for tests of endogeneity in our scenario for panel data.

Regarding our application, although we are comfortable with our choice of instruments, there are other quantitative avenues that could be explored regarding instrument construction (e.g. aggregated amount of contemporaneously redeemed coupons in time t). Thus, we do not claim our instrument set is fully exhaustive regarding the associated idiosyncratic endogeneity test. Also, our nonlinear, IV-GMM solution, as it relates to choosing endogenous lags of Type A campaign is strongly dependent on the lack of serial correlation in the time-varying unobservables. An assessment of finite-sample properties of this IV-GMM solution via simulation can also be a necessary extension. Additionally, and to be clear, the resulting unbalancedness in our study occurs specifically due to the aggregation scheme for which the designers of the loyalty promotional program sent campaigns that contain redeemable coupons within a designated period over the two years. To reiterate, we do not know the exact logic as to why certain households are prone to any exposure nor why they are exposed for the temporal duration for which they are exposed. To this end, “attrition” is seemingly at play as the in-store visits and all other quantitative metrics of interest are aggregated and

directly linked with the time frame for which a campaign is valid. In other words, there is a lower number of households who were observed, for example, 14 periods as opposed to a higher number of households that were observed, for instance, for 4 periods. The scope of these households' observations in certain time frames in our application is linked to the period covariate which represents any time there was a change in exposure effect either due to reversions back to “washout states” (‘Type D’), a change in the exposure effect (‘Type A’ to ‘Type B’), or even any instance in which exposure effects overlap over the same duration of coupon redeemability (e.g. a household received ‘Type A’ and ‘Type B’ conjointly, over the time frame in which they were campaign sets contain coupons that are redeemable, resulting in ‘Type AB’). However, we argue this is not a case of “attrition” in the traditional sense as the resulting unbalancedness is not, at least in part, heavily due to situations in which our cross-sectional units “drop out” and never return to visit a store throughout the loyalty promotional campaign. The “attrition” arises due to our aggregation as a result of campaign/treatment distribution in addition to the selection bias associated with the distribution of the Type A campaign.

Prior to aggregation, and as a quick check regarding our 1584 households who participated in the loyalty program, approximately 70% of the 1584 households (1110 households) had at least one transaction (in-store visit) that occurred at least three-quarters through the entire duration of the program. However, this form of truncation for household behavior for which we have no latter promotional information on in-store visits could be attributed to a variety of different factors that is completely unrelated to any underlying mechanisms of the promotional program or the treatment distribution process, and may be considered a MAR process. If there were a large amount of households who only visited in-store, for example, for the first 30 days of the test cycle, then that would be an issue worthy of further investigation.

Chapter 3: Limitations/Avenues for future research

Although we were able to see very favorable results from our simulation designs, we note that we did not consider any other spline besides other than thin plate regression splines for the interactive nonlinear effects via the factor smooths. There are other splines that could have been attempted; yet, we opted to use the most “optimal” spline offered by the **mgcv** package. However, given this choice and the size of the datasets in all simulations, all of our design simulations were not necessarily computationally efficient. For example, each simulation design attempted across varying parameters took approximately 3.3 hours on average to complete. Positively, however, we did not run into many convergence issue which would require a re-run as a maximum of 250 iterations seemed to be reasonable. Convergence issues occurred in approximately 10%-12% of the 500 replications in any scenario. Also, in our simulations, we chose a moderate amount of observations per individual (e.g. $n_i = 40$ and $n_i = 25 - 40$) as there were definitive moderate bias issues, specifically for ρ , when the number of observations were set to be small ($n_i = 3 - 5$). This also contributes of course to the overall computational inefficiency in terms of estimation in our Monte Carlo simulations. Also, there could have been more considerations for variance parameter pairs ($\sigma_v^2 = 0.75$,

$\rho = 0.75$), chosen functions for $f(\cdot)$ (e.g. more complex polynomials) for robustness of the thin plate regression approach, and the distribution of our response variable. Summarizing, it would certainly be interesting to study finite sample bias performance of hierarchical, complex variance structures where the response could be Gaussian, binary, or follow any other parametric distribution of interest with even more complex, nonlinear test functions (polynomials) associated with levels exceeding two of the parametric factor covariate. Additionally, of interest is perhaps the assessment of variance components for conditionally serially correlated counts for higher-ordered autoregressive specifications beyond the AR(1) processes. Obvious extensions such as multivariate smooths and tensor products, perhaps even augmented by parametric factor smooths, would be an interesting avenue as well for finite sample performance in these semi-parametric mixed model formulations. Also, in our simulations, we relied specifically on the simulation of conditionally serially correlated Poisson counts from [5], whereas our final chosen model was a negative binomial GAMM for our application. While there is precedence for simulating both stationary and non-stationary correlation structures for counts following a negative binomial distribution, the process of estimating the GAMMs via penalized quasi-likelihood estimation would be extremely inefficient as it would require the shape parameter to be known for each replication of our model, which would have required an additional 500 models per scenario to estimate and extract the shape parameter from the `gam()` function in order to utilize it in `gamm()` utilizing the `mgcv` package.

Regarding our application, we chose to include few covariates (in the parametric portion of the GAMM) for simplicity as our main focus were the factor smooth interactions and variance components, exclusively. There also exists avenues to exploit GAMMs in future non-promotional periods (NP3, NP4, etc.) utilizing these methods to assess nonlinear interactions of campaign types received over time.

Chapter 4: Limitations and avenues for future research

The limitations presented in this chapter are few although we are quite confident in our results. Firstly, in the first portion of the production of OOS forecasts (prior to any combination), we do have relatively fewer observations used in the recursive forecasting estimation. That is, the underlying forecasting models are produced from a training set of $1 : (52 - h)$ ¹, $h = \{1, 2, 3, 4, 8, 12\}$ attributed specifically to the limited data availability of inflation expectations of the survey based forecasting model across all horizons². The limited data availability is likely reflected in the uninspired results of the Theil's U statistics in Table 4.1. Secondly, the underlying model constructs we used for our combination considerations, while effective, are arbitrary. While our forecasters are shown to be logical and effective, there exists an array of other univariate, linear time series models that exist beyond ours that could be perhaps used to estimate the monthly annualized inflation gap. Also, the L1 penalty that we used in our proposed L1-penalized quantile forecast combination procedure is only one penalization procedure that could be used. Future work could

¹M1:1998 until M4:2002 - h

²Inflation Expectation from the data source we found begins on M1:1998

explore different penalization procedures for selecting forecasters. Additionally, following from [6], there exists an avenue to consider a weighting scheme that is time-varying if there are reasons to believe that quantiles' relevance in determining optimal forecasts may be non-constant over time which could be assessed via a constrained regression technique. We, however, rely exclusively on a time-invariant weight scheme for L1PQFC1 and L1PQFC2 models. Lastly, it would also be very interesting to observe how well this forecasting procedure could be applied to multivariate time series along the lines of [7] and [8].

References

- [1] K. E. Staub, “Simple tests for exogeneity of a binary explanatory variable in count data regression models,” *Communications in Statistics-Simulation and Computation*, vol. 38, no. 9, pp. 1834–1855, 2009. [165](#)
- [2] H. B. Nguyen, *Estimating Nonlinear Cross Section and Panel Data Models with Endogeneity and Heterogeneity*. Michigan State University. Economics, 2011. [165](#)
- [3] A. Geraci, D. Fabbri, and C. Monfardini, “Testing exogeneity of multinomial regressors in count data models: Does two-stage residual inclusion work?,” *Journal of Econometric Methods*, vol. 7, no. 1. [165](#)
- [4] B. E. Honoré and M. Kesina, “Estimation of some nonlinear panel data models with both time-varying and time-invariant explanatory variables,” *Journal of Business & Economic Statistics*, vol. 35, no. 4, pp. 543–558, 2017. [165](#)
- [5] B. C. Sutradhar, *Dynamic mixed models for familial longitudinal data*. Springer Science & Business Media, 2011. [167](#)
- [6] L. R. Lima and F. Meng, “Out-of-sample return predictability: A quantile combination approach,” *Journal of Applied Econometrics*, vol. 32, no. 4, pp. 877–895, 2017. [168](#)
- [7] N.-J. Hsu, H.-L. Hung, and Y.-M. Chang, “Subset selection for vector autoregressive processes using lasso,” *Computational Statistics & Data Analysis*, vol. 52, no. 7, pp. 3645–3657, 2008. [168](#)
- [8] W. B. Nicholson, I. Wilms, J. Bien, and D. S. Matteson, “High dimensional forecasting via interpretable vector autoregression,” *arXiv preprint arXiv:1412.5250*, 2014. [168](#)

Vita

Amil Jamal Williamson is an applied statistician from Tennessee. He gained a bachelor's degree in statistics in 2010. He also gained a full scholarship for both his master's degree at The University of Tennessee completed in 2012 as well as a full scholarship for the pursuit of his Ph.D. He gained significant exposure to teaching undergraduate courses in statistics at the University of Tennessee while concurrently pursuing his master's degree and in the interim before entering the Ph.D. program in 2015. He earned a Doctor of Philosophy in Management Science with a concentration in Analytics in December 2019 from The University of Tennessee.