TRADE-OFFS IN NON-LINEAR MODELS AND ESTIMATION STRATEGIES

By

Alyssa Helen Carlson

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics — Doctor of Philosophy

2019

ABSTRACT

TRADE-OFFS IN NON-LINEAR MODELS AND ESTIMATION STRATEGIES

By

Alyssa Helen Carlson

This dissertation examines the assumptions presumed throughout the literature to establish valid estimation procedures for non-linear models. The following three chapters addresses issues of identification, consistent and efficient estimation, and incorporating heteroskedasticity and serial correlation for binary response models in cross-sectional and panel data settings.

Chapter 1: Parametric Identification of Multiplicative Exponential Heteroskedasticity

Multiplicative exponential heteroskedasticity is commonly seen in latent variable models such as Probit or Logit where correctly modelling the heteroskedasticity is imperative for consistent parameter estimates. However, it appears the literature lacks a formal proof of point identification for the parametric model. This chapter presents several examples that show the conditions presumed throughout the literature are not sufficient for identification and as a contribution provides proofs of point identification in common specifications.

Chapter 2: Relaxing Conditional Independence in an Endogenous Binary Response Model

For binary response models, control function estimators are a popular approach to address endogeneity. But these estimators utilize a Control Function assumption that imposes Conditional Independence (CF-CI) to obtain identification. CF-CI places restrictions on the

relationship between the latent error and the instruments that are unlikely to hold in an empirical context. In particular, the literature has noted that CF-CI imposes homoskedasticity with respect to the instruments. This chapter identifies the consequences of CF-CI, provides examples to motivate relaxing CF-CI, and proposes a new consistent estimator under weaker assumptions than CF-CI. The proposed method is illustrated in an application, estimating the effect of non-wife income on married women's labor supply.

Chapter 3: Behavior of Pooled and Joint Estimators in Probit Model with Random Coefficients and Serial Correlation

This chapter compares a pooled maximum likelihood estimator (PMLE) to a joint (full) maximum likelihood estimator (JMLE), the dominant estimation method for mixture models, for dealing with potential individual-specific heterogeneity and serial correlation in a binary response Probit Mixture model. The JMLE is more statistically efficient but computationally demanding and the implementation becomes more difficult if one tries to model the serial correlation over time. On the other hand, the PMLE is computationally simple and robust to arbitrary forms of serial correlation. Focusing on the Average Partial Effects, this chapter finds it imperative for the model to allow the individual-specific heterogeneity to be potentially correlated with the covariates (not a standard specification in Mixture models). Moreover, the JMLE can produce quite satisfactory estimates that seem robust to serial correlation even under misspecification of the likelihood function. Results are illustrated in an application, estimating the effects of different interventions on high risk men's behavior, complementing the original study of Blattman, Jamison, and Sheridan (2017).

ACKNOWLEDGMENTS

First and foremost, I would like to thank the chair of my dissertation committee, Jeff Wooldridge, for all of his advice, encouragement, and helpful critiques. I would also like to thank Kyoo Il Kim, Joe Herriges and Nicole Mason for serving on my committee and providing valuable feedback and assistance. I also appreciate the comments of seminar participants at Michigan State University, the Econometrics Reading Group at MSU, Grand Valley State University, the 2018 MEA Conference, the 2018 and 2019 Annual Meeting of the Midwest Econometrics Group and the corresponding Women's Mentoring Workshops, and the 2018 International Association of Applied Econometrics Conference.

I am especially grateful for the financial support I received from the Graduate School and the Department of Economics at Michigan State University, including the Goodman Fellowship, Summer Research Fellowship, and Dissertation Completion Fellowship. I also appreciate the support and advice that Lori Jean Nichols, Steven Haider, and Mike Conlin all gave me as I navigated the graduate program and job market. I am also grateful to my friends and colleagues at Michigan State for making my graduate experience so memorable.

I am truly thankful for my endlessly supportive parents Lance and Chim Carlson, whose love and encouragement helped me at every step of my life. Finally, I am especially grateful to my partner, Thom, for picking up his life and moving across the country to start a new adventure with me (multiple times), as well as all the countless ways he has supported my endeavors over the years.

TABLE OF CONTENTS

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **xi**

Introduction **1**

Chapter 1 Parametric Identification of Multiplicative Exponential Heteroskedasticity **3**
1.1 Introduction 3
1.2 Identification when there is no bijective transformation 7
1.3 No identification when there is a bijective transformation 10
1.4 Identification in a common specification 14
1.5 Conclusion 15

Chapter 2 Relaxing Conditional Independence in an Endogenous Binary Response Model **17**
2.1 Introduction 17
2.2 Background and Motivation 23
2.3 Model Set Up 29
2.4 General Control Function 38
2.4.1 Identification 38
2.4.2 Simulation: General Control Function in the Demand for Premium Cable 43
2.5 Estimation and Interpretation 46
2.5.1 Asymptotic Properties 47
2.5.2 Average Structural Function 49
2.5.3 Average Partial Effects 56
2.5.4 Simulation: ASF Estimates for the Effect of Income on Homeownership 58
2.6 Empirical Example 61
2.7 Extension: Semi-Parametric Distribution Free Estimator 67
2.7.1 Observational Equivalence and Identification 68
2.7.2 Asymptotic Properties 76
2.7.3 Simulation 80
2.8 Conclusion 86

Chapter 3 Behavior of Pooled and Joint Estimators in Probit Model with Random Coefficients and Serial Correlation **89**
3.1 Introduction 89
3.2 Model Set Up 97
3.3 Estimation Methods 98

3.3.1 Mixed Effects Probit . . . 98
3.3.2 Pooled Heteroskedastic Probit . . . 103
3.4 Average Partial Effects . . . 105
3.5 Simulation . . . 108
3.5.1 Computational Results . . . 110
3.5.2 Parameter Estimates . . . 111
3.5.3 Average Partial Effect Estimates . . . 115
3.5.4 ASF . . . 117
3.6 Application . . . 119
3.7 Discussion . . . 127
3.7.1 AR(2) . . . 129
3.7.2 No Random Effects . . . 130
3.7.3 Logit . . . 132
3.8 Conclusion . . . 134

APPENDICES . . . 137
APPENDIX A Figures for Chapter 1 . . . 138
APPENDIX B Proofs and Notation for Chapter 2 . . . 141
APPENDIX C Simulation Details for Chapter 2 . . . 151
APPENDIX D Figures for Chapter 2 . . . 157
APPENDIX E Tables for Chapter 2 . . . 181
APPENDIX F Figures for Chapter 3 . . . 200
APPENDIX G Tables for Chapter 3 . . . 224

BIBLIOGRAPHY . . . 260

LIST OF TABLES

Table E.1: **Summary Statistics** 182

Table E.2: **Comparison of Logit Parameter Estimates** 183

Table E.3: **Comparison of Price Elasticity Estimates** 183

Table E.4: **Comparison of Summary Statistics** 184

Table E.5: **Comparison of Parameter Estimates** 185

Table E.6: **APE Results and Simulated Distribution (True APE = 0.6448)** 186

Table E.7: **Summary Statistics** 187

Table E.8: **Coefficient Estimates for Married Women's LFP** 188

Table E.9: **Wald Test Results** 190

Table E.10:**APE Estimates for Non-Wife Income effect on Wife's LFP** 191

Table E.11:**Logistic Distribution** ($h_o^1 = v_{2i}$) 192

Table E.12:**Uniform Distribution** ($h_o^1 = v_{2i}$) 192

Table E.13:**Student T Distribution** ($h_o^1 = v_{2i}$) 193

Table E.14:**Gaussian Mixture Distribution** ($h_o^1 = v_{2i}$) 193

Table E.15:**Logistic Distribution with Linear GCF** (h_o^2) 194

Table E.16:**Uniform Distribution with Linear GCF** (h_o^2) 194

Table E.17:**Student T Distribution with Linear GCF** (h_o^2) 195

Table E.18:**Gaussian Mixture Distribution with Linear GCF** (h_o^2) 195

Table E.19:**Logistic Distribution with Non-Parametric GCF** (h_o^3) 196

Table E.20:**Uniform Distribution with Non-Parametric GCF** (h_o^3) 196

Table E.21:**Student T Distribution with Non-Parametric GCF (h_o^3)** 197

Table E.22:**Gaussian Mixture Distribution with Non-Parametric GCF (h_o^3)** . 197

Table E.23:**Heteroskedastic Logistic ($h_o^1 = v_{2i}$)** 198

Table E.24:**Heteroskedastic Logistic with Linear GCF (h_o^2)** 198

Table E.25:**Heteroskedastic Logistic with Non-Parametric GCF (h_o^3)** 199

Table G.1: **Estimation Times for DGP 1** 225

Table G.2: **Estimation Times for DGP 2** 226

Table G.3: **Estimation Times for DGP 3** 227

Table G.4: **Bias and Std Deviation of De-scaled ME Probit Estimates for DGP 1** 228

Table G.5: **Bias and Std Deviation of De-scaled ME Probit Estimates for DGP 2** 229

Table G.6: **Bias and Std Deviation of De-scaled ME Probit Estimates for DGP 3** 230

Table G.7: **Bias and Std Deviation of Scaled Coefficient Estimates for DGP 1** 231

Table G.8: **Bias and Std Deviation of Scaled Coefficient Estimates for DGP 2** 232

Table G.9: **Bias and Std Deviation of Scaled Coefficient Estimates for DGP 3** 233

Table G.10:**Root Mean Square Error of $\hat{\beta}_{2\sigma}$ for Specification (2)** 234

Table G.11:**Bias and Std Deviation of Variance Component σ_2^2 for Specification (2)** 235

Table G.12:**Bias and Std Deviation ($\times 10$) of APE Estimates for DGP 1** . . . 236

Table G.13:**Bias and Std Deviation ($\times 10$) of APE Estimates for DGP 2** . . . 237

Table G.14:**Bias and Std Deviation ($\times 10$) of APE Estimates for DGP 3** . . . 238

Table G.15:**Comparison of APE and PEA** 239

Table G.16:**Select Baseline Summary Statistics** 240

Table G.17:**Preliminary OLS Estimates** 242

Table G.18:**Scaled Probit Coefficient Estimates for Selling Drugs** 244

Table G.19:**Scaled Probit Coefficient Estimates for being Arrested** 245

Table G.20:**Scaled Probit Coefficient Estimates for Illicit Activity** 246

Table G.21:**ATE Estimates** 247

Table G.22:**Bias and Std Deviation of Scaled Coefficient Estimates under AR(2)** 248

Table G.23:**Bias and Std Deviation ($\times 10$) of APE Estimates under AR(2)** . . 249

Table G.24:**Failure Count under no Random Coefficients** 249

Table G.25:**Estimation Times under no Random Coefficients** 250

Table G.26:**Bias and Std Deviation ($\times 10$) of APE Estimates under no Random Coefficients** 251

Table G.27:**Variance Component σ_1^2 Estimates under no Random Coefficients** 252

Table G.28:**Variance Component σ_2^2 Estimates under no Random Coefficients** 252

Table G.29:**Rejection Rate of LR Test for Random Coefficients** 253

Table G.30:**Bias and Std Deviation of De-scaled ME Logit Estimate under a Conditional Logistic AR(1) Process** 254

Table G.31:**Bias and Std Deviation of De-scaled ME Logit Estimate under a Marginal Logistic AR(1) Process** 255

Table G.32:**Bias and Std Deviation of Scaled Coefficient Estimates under a Conditional Logistic AR(1) Process** 256

Table G.33:**Bias and Std Deviation of Scaled Coefficient Estimates under a Marginal Logistic AR(1) Process** 257

Table G.34: **Bias and Std Deviation ($\times 10$) of APE Estimates under a Conditional Logistic AR(1) Process** 258

Table G.35: **Bias and Std Deviation ($\times 10$) of APE Estimates under a Marginal Logistic AR(1) Process** 259

LIST OF FIGURES

Figure A.1: **Visual representation of bijective transformations** 139

Figure A.2: **Parameter estimates from two observationally equivalent models** 140

Figure D.1: **Effect of Heteroskedasticity on Parameter Estimate** 158

Figure D.2: **ASF for Income equal to $85,000** 159

Figure D.3: **ASF Estimates for Misspecified Models** 160

Figure D.4: **Consequence of CF-LI Assumption on ASF Estimates** 161

Figure D.5: **Comparison of ASF for Families with No Children** 162

Figure D.6: **Comparison of ASF for Families with Young Children Only** . . . 163

Figure D.7: **Comparison of ASF for Families with Old Children Only** 164

Figure D.8: **Comparison of ASF for Families with Both Young and Old Children** 165

Figure D.9: **Logistic Distribution** ($h_o^1 = v_{2i}$) 166

Figure D.10: **Uniform Distribution** ($h_o^1 = v_{2i}$) 167

Figure D.11: **Student T Distribution** ($h_o^1 = v_{2i}$) 168

Figure D.12: **Gaussian Mixture Distribution** ($h_o^1 = v_{2i}$) 169

Figure D.13: **Logistic Distribution with Linear GCF** (h_o^2) 170

Figure D.14: **Uniform Distribution with Linear GCF** (h_o^2) 171

Figure D.15: **Student T Distribution with Linear GCF** (h_o^2) 172

Figure D.16: **Gaussian Mixture Distribution with Linear GCF** (h_o^2) 173

Figure D.17: **Logistic with Non-Parametric GCF** (h_o^3) 174

Figure D.18: **Uniform with Non-Parametric GCF** (h_o^3) 175

Figure D.19: **Student T with Non-Parametric GCF (h_o^3)** 176

Figure D.20: **Gaussian Mixture with Non-Parametric GCF (h_o^3)** 177

Figure D.21: **Heteroskedastic Logistic ($h_o^1 = v_{2i}$)** 178

Figure D.22: **Heteroskedastic Logistic with Linear GCF (h_o^2)** 179

Figure D.23: **Heteroskedastic Logistic with Non-Parametric GCF (h_o^3)** 180

Figure F.1: **Distribution of $\hat{\sigma}_1^2$ for T=5 under DGP1** 200

Figure F.2: **Distribution of $\hat{\sigma}_1^2$ for T=10 under DGP1** 201

Figure F.3: **Distribution of $\hat{\sigma}_1^2$ for T=20 under DGP1** 202

Figure F.4: **Distribution of $\hat{\sigma}_1^2$ for T=5 under DGP2** 203

Figure F.5: **Distribution of $\hat{\sigma}_1^2$ for T=10 under DGP2** 204

Figure F.6: **Distribution of $\hat{\sigma}_1^2$ for T=20 under DGP2** 205

Figure F.7: **Distribution of $\hat{\sigma}_1^2$ for T=5 under DGP3** 206

Figure F.8: **Distribution of $\hat{\sigma}_1^2$ for T=10 under DGP3** 207

Figure F.9: **Distribution of $\hat{\sigma}_1^2$ for T=20 under DGP3** 208

Figure F.10: **ASF Estimates for T=5 under DGP1** 209

Figure F.11: **ASF Estimates for T=10 under DGP1** 210

Figure F.12: **ASF Estimates for T=20 under DGP1** 211

Figure F.13: **ASF Estimates for T=5 under DGP2** 212

Figure F.14: **ASF Estimates for T=10 under DGP2** 213

Figure F.15: **ASF Estimates for T=20 under DGP2** 214

Figure F.16: **ASF Estimates for T=5 under DGP3** 215

Figure F.17: **ASF Estimates for T=10 under DGP3** 216

Figure F.18: **ASF Estimates for T=20 under DGP3** 217

Figure F.19: **ATE for Selling Drugs** 218

Figure F.20: **ATE for Being Arrested** 219

Figure F.21: **ATE for Engaging in Illicit Activities** 220

Figure F.22: **Distribution of $\hat{\sigma}_1^2$ for T=5 under AR(2)** 221

Figure F.23: **Distribution of $\hat{\sigma}_1^2$ for T=10 under AR(2)** 222

Figure F.24: **Distribution of $\hat{\sigma}_1^2$ for T=20 under AR(2)** 223

Introduction

When the outcome has restricted support – non-negative, binary, discrete, etc. – non-linear models are used to better capture the underlying data generating process in which a linear model can only at best approximate. A common example is a binary response model where the threshold latent variable set-up is more reasonable than a linear approximation where the predicted outcomes could fall outside the 0 and 1 bound for probabilities. But unlike the linear regression, the non-linear models are not as well understood when the standard assumptions fail to hold. This dissertation addresses three settings in which the standard assumptions fail to hold: heteroskedasticity, endogeneity, and a panel setting with random coefficients and serial correlation.

In the first chapter, I address the identification of a multiplicative exponential heteroskedastic model. Although it is presented in a general setting, introducing heteroskedasticity in a binary response model is usually done through a multiplicative exponential heteroskedasticity. Unlike the linear regression, heteroskedasticity – where the variance of the latent error depends on the covariates – does not just influence the calculation of the asymptotic variance (and consequently the standard error estimates), but it changes the conditional mean function in the log-likelihood. This means that ignoring heteroskedasticity in a binary response model will result in inconsistent parameter estimates, not just inaccurate standard

error estimates (see Figure D.1). Moreover, introducing a heteroskedastic specification can capture more flexibility in the conditional mean function as see in Chapters 2 and 3. In chapter 2, I utilize an observational equivalence result from Khan (2013) that essentially implies flexibly specified heteroskedasticity allows for distributional misspecification in the latent error. In Chapter 3, a pooled probit estimator allows for random coefficients through a heteroskedastic specification. Both of these examples provides further motivation for the utility of a heteroskedastic binary response mode.

But I find the assumptions in the literature are not sufficient to guarantee identification of a multiplicative exponential heteroskedastic model. I provide a proof of identification for a linear in parameters specification that will be utilized in the later chapters.

The next two chapters propose and compare estimation procedures for specific binary response settings. In both cases, the alternative estimators from the literature considered are built upon a set of fairly restrictive assumptions. I examine the assumptions underlying the estimators and ask if they are realistic empirically. In both cases, I find simply scenarios in which the underlying assumptions would be violated. In the second chapter, the conditional independence assumption for the control function estimators are violated when there is heteroskedasticity. In the third chapter, the joint maximum likelihood estimator is inconsistent under the presence of serial correlation. Given these limitations, an alternative estimation procedure is proposed. This dissertation aims to supply empirical economists with estimation tools they would need to address these complex issues that commonly arise in binary response estimation.

Chapter 1

Parametric Identification of Multiplicative Exponential Heteroskedasticity

1.1 Introduction

Multiplicative exponential heteroskedasticity was first proposed by Harvey (1976) in the context of a linear conditional mean model. Estimation is undertaken in two stages, requiring first an argument that the conditional variance parameters are identified and then showing that a weighted least squares estimator identifies the conditional mean parameters. More recently, multiplicative exponential functions are used to model heteroskedasticity in the latent errors of binary response models. However, with the cases of heteroskedastic Logit and Probit, the parameters in the conditional variance function are estimated concurrently with the coefficients of interest requiring joint identification of the parameters. Standard textbooks such as Greene (2011) and Wooldridge (2010) state that these models are estimable under fairly standard conditions and are more flexible in the specification of the conditional mean function compared to standard Probit and Logit models leading to widespread use in empirical work. However the literature has yet to provide proofs of parametric identification.

To fill the gap in the literature, this chapter explores the issues of identifications in models with exponential heteroskedasticity.

Although the results may be applied for any model with a multiplicative exponential component, I will use the example of a heteroskedastic binary response model throughout the chapter to give the identification proofs some context. Consider the standard latent binary response model set up:

$$Y = 1\{X\beta_o - U > 0\}$$

where U is heteroskedastic with conditional variance: $Var(U|Z) = \exp(2Z\delta_o)$ where Z may include functions X. Presuming that there is no endogeneity in the usual sense, $E(U|X,Z) = 0$, and the scaled latent error, $U/\exp(Z\delta_o)$, is independent of (X,Z), then the heteroskedastic binary response model has the following conditional probability distribution,

$$f(y|X,Z,\theta_o) = \left(\Phi\left(\frac{X\beta_o}{\exp(Z\delta_o)}\right)\right)^y \left(1 - \Phi\left(\frac{X\beta_o}{\exp(Z\delta_o)}\right)\right)^{(1-y)} \tag{1.1}$$

where Φ is the known cumulative distribution function for $\exp(Z\delta_o)U$, is monotonic, and has support on the real line. If we assume a normal distribution then this is the individual likelihood for a heteroskedastic Probit model and if we assume a logistic distribution this is the individual likelihood for the heteroskedastic Logit model.

The following restates the identification definition from Newey and McFadden (1994) for MLE.[1]

Definition 1.1.1 (Identification)**.** *Let $f(y|X,Z,\theta_o)$ be the conditional probability distribution of Y defined over the measures of X and Z with positive probability. If $\theta \neq \theta_o$ in the parameter space Θ implies $P(f(y|X,Z,\theta) \neq f(y|X,Z,\theta_o)) > 0$ over the measures of X and*

[1]This discussion can easily be extended to the cases of NLLS or GMM.

Z then θ_o is point identified.

If one were to assume that $E(X'X)$ is non-singular and β_o is non-zero, then identification requires

$$\text{For } (\beta, \delta) \in \Theta, \text{ if } X(\beta_o - \exp(Z(\delta - \delta_o))\beta) = 0 \text{ w.p. } 1, \text{ then } (\beta, \delta) = (\beta_o, \delta_o) \tag{1.2}$$

where Θ is the joint parameter space. The above statement captures the fundamental identification requirement for models with exponential heteroskedasticity. This chapter aims to clarify when this statement holds and under what necessary or sufficient conditions.

The simplest case of identification is when X and Z are not a bijective transformation of each other in the sense that the variation in Z cannot be entirely explained by the X or visa versa. One of the main contributions of this chapter is to provide a formal proof of identification in this scenario under the standard conditions of the literature. A sufficient condition for X and Z to not be a bijective transformation of one another is to impose an exclusion restriction (in either X or Z). An exclusion restriction would require that one of the random variables in the vector X is not included in the vector Z or visa versa. By doing so, variation in one of the random vectors has been introduced that cannot be perfectly explained by the other random vector. But when one allows for an arbitrary relationship between X and Z, the standard conditions are no longer sufficient for identification.

To provide some intuition, when X and Z are a bijective transformation, then showing identification is difficult due to the non-linear nature of the problem. Noted in Lewbel (forthcoming), non-linearity can allow for multiple solutions to the statement in (1.2). Specifically, if the relationship between X and Z allows $X\beta_o$ to be equal to a scaling of $X\beta$ by $\exp(Z(\delta - \delta_o))$, then the model is not identified. This chapter will look at two ways the scaling of $X\beta$ by $\exp(Z(\delta - \delta_o))$ can be manipulated: through the joint support of (X, Z) and

through the functional form of the heteroskedasticity, $\exp(Z\delta_o)$. Section 3 discusses several counter-examples in which the conditions presupposed in the literature are not sufficient for identification.

The non-identification result is section three can be compared to the literature on identification in a binary response model. Identification in this setting has been well-studied in several papers by Manski (1985, 1988). In the earlier paper, Manski looks at identification of a binary response model under a median restriction. This method allowed for arbitrary heteroskedasticity in the latent error but at most, would identify the scaled parameters $\beta_o/||\beta_o||$. Simply put, to obtain identification in his framework, for every β in the parameter space such that $P(\text{sgn}(X\beta) \neq \text{sgn}(X\beta_o)) > 0$ then $\beta/||\beta|| \neq \beta_o/||\beta_o||$. Consequently, he provided non-identification results depending on the support of X. For instance, if $X\beta$ had bounded support away from 0 (for all values of β in the parameter space) then β_o is not identified. However in our setting, the identification definition is based on the entire likelihood, rather than just the sign of the linear index. So to be clear, non-identification results presented in this chapter are consequences of the highly non-linear specification of exponential multiplicative heteroskedasticity as opposed to limited information in median restriction framework in Manski (1985).

Manski (1988) looks at identification of the scaled parameters β_o in conjunction with the non-parametric identification of the conditional cumulative distribution of the latent error, $F_{U|X}(\cdot)$. Manski is able to show identification in the case of a known cumulative distribution function and statistical independence between U and X. But since in our setting there is heteroskedasticity, statistical independence does not hold. Manski also provides a non-identification result in the case of conditional mean independence and an unknown cumulative distribution function. Although in our setting, the conditional distribution of

the latent error is parametrically specified, the non-identification result would suggest that there should exist some conditional distribution specification in which identification does not hold. Therefore it is unsurprising that even in this parametric setting, we can construct counter-examples in which identification is lost.

However, the non-identification results should not discourage the utilization of models with multiplicative heteroskedasticity in empirical work. The counter-examples provided are trivial in nature and are meant to highlight the non-existence of a general identification theorem for these models. As a helpful contribution, this chapter ends with a corollary that provides identification with a bijective transformation relationship between the random vectors for possibly the most commonly used specification.

1.2 Identification when there is no bijective transformation

Continuing with the example of a heteroskedastic binary response model described in equation (1.1.1), standard textbooks such as Greene (2011) and Wooldridge (2010) imply that the parameters are estimable[2] under the following conditions,

Condition 1. *Z does not contain a constant.*

Condition 2. *$E(X'X)$ is non-singular.*

Condition 3. *$E(Z'Z)$ is non-singular.*

[2]In this context, estimable is interpretively synonymous with point identified. However, neither text provides proofs of identification nor explicitly state that the models are point identified. Therefore the term "estimable" emphasizes the lack of rigorous treatment in the literature for identification in a parametric model.

Condition 1 implies the model is only identified up to scale. Alternatively, one could assume the normalization that one of the coefficients on X is equal to 1. Conditions 2 and 3 are needed in order to show $X\beta_o = X\beta$ and $Z\delta_o = Z\delta$ implies $\beta_o = \beta$ and $\delta_o = \delta$ respectively. Additionally, although not commonly stated, identification requires

Condition 4. *β_o is non-zero.*

Without this assumption, δ_o is not identified.[3] This can easily be addressed by assuming a non-zero intercept as a location normalization. Under these assumptions, identification holds when X and Z are not bijective transformations of each other stated in the following theorem.

Theorem 1.2.1. *If Conditions 1-4 hold, and X and Z are not bijective transformations of each other, then the parameters β_o and δ_o are point identified.*

Before providing the proof, I will formally characterized 'bijective transformation'.

Definition 1.2.1 (Bijective Transformation)**.** *X is a bijective transformation of Z (and equivalently Z is a bijective transformation of X) if there exists a bijective function f such that*

$$X = f(Z)$$

$$Z = f^{-1}(X)$$

where f^{-1} denotes the inverse of f.

[3] Suppose $\beta_O = 0$ then $X\beta_O$ is zero with probability 1, so as long as $\beta = 0$, then any $\delta \neq \delta_O$ satisfies

$$X(\beta_O - \exp(Z(\delta - \delta_O))\beta) = 0$$

therefore δ_O is not identified. This has fairly minor consequences since in empirical work, researchers tend to be more interested in the coefficient parameters β_O.

This definition can also be understood in terms of the support of X and Z. A bijective transformation would require that for every x in the support of X, there exists a unique z in the support of Z such that (x, z) occurs with positive probability in the joint support of (X, Z) and for any $z' \neq z$ in the support of Z, (x, z') occurs with probability 0 in the joint support. Conversely, for every z in the support of Z there exists a unique x in the support of X such that (x, z) occurs with positive probability in the joint support of (X, Z) and for any $x' \neq x$ in the support of X, (x', z) occurs with probability 0 in the joint support. This implies that the variation in X can be perfectly described by variation in Z. Figure A.1 visually shows what is implied by a bijective transformation and examples in which a bijective transformation does not hold.

Proof. Suppose there exists a $(\beta, \delta) \in \Theta$ such that

$$X(\beta_o - \exp(Z(\delta - \delta_o))\beta) = 0 \tag{1.3}$$

holds for almost all X and Z in their support. Since β_o is non-zero and $E(X'X)$ is non-singular, $X\beta_o$ (and similarly $X\beta$) is non-zero with positive probability. Consequently, $X\beta_o/X\beta$ exists and is strictly positive with positive probability.[4] Rearranging the equation above for a realization (x, z),

$$z(\delta_o - \delta) = \ln(x\beta_o/x\beta) \tag{1.4}$$

where the realizations are in the following restricted support $\{(x, z) \in \text{support}(X, Z) : x\beta_o \text{ and } x\beta \text{ are non-zero}\}$. Since X and Z are not bijective transformations of each other, there must exist variation in either X or Z that cannot be explained by the other. When there is variation in Z not explained by X, there exists a realization in X in which there are

[4] For equation (1.3) to hold, $\text{sign}(X\beta_o) = \text{sign}(X\beta)$

more than one realizations of Z that occur with positive probability in the joint support. This would allow for different realizations on the left hand side of the above equation while the right hand side is fixed at one possible realization. Since $E(Z'Z)$ is non-singular, the above equation can only hold when $\delta = \delta_o$ and consequently $\beta = \beta_o$. Similar conclusions follow when there is variation in X not explained by Z.□

1.3 No identification when there is a bijective transformation

However confining to the case X and Z are not bijective transformations of one another is fairly restrictive. Return to the heteroskedastic binary response example where one is interested in modelling the mean of Y conditional on X. Let $\sigma(X)$ denote the conditional standard deviation of the latent error where it is reasonable to assume a double index model such that,

$$E(Y|X) = \Phi\left(\frac{X\beta_o}{\sigma(X)}\right) = \Phi\left(\frac{X\beta_o}{\exp(Z\delta_o)}\right) \tag{1.5}$$

where Z consists of bijective transformations of the elements in X.[5] As mentioned before, to get around X and Z being bijective transformations, one could consider imposing an exclusion restriction. But this would require prior knowledge of which elements in X effect the conditional variance and which would not. Nevertheless, more generally, identification is not obtainable in the case of bijective transformation under the previously stated conditions. The following two counter-examples provide settings in which identification fails.

[5]Klein and Vella (2009) discuss identification in the semi-parametric case where they use a re-indexing approach following Ichimura and Lee (1991).

Counter-example: binary support

Suppose $X = (1, Z)$ where Z is a binary variable. Then the first part of statement (1.2) can be decomposed to,

$$X(\beta_o - \exp(Z(\delta - \delta_o))\beta) = \begin{cases} \beta_{1o} - \beta_1 & \text{if } Z = 0 \\ \beta_{1o} + \beta_{2o} - \exp(\delta - \delta_o)(\beta_1 + \beta_2) & \text{if } Z = 1 \end{cases} \tag{1.6}$$

The first part implies $\beta_1 = \beta_{1o}$. Plugging into the second part, equation (1.2) holds if $\beta_2 = \exp(\delta_o - \delta)\beta_{2o} - \beta_{1o}(1 - \exp(\delta_o - \delta))$ which does not imply $\delta = \delta_o$ or $\beta_2 = \beta_{2o}$. Even though Conditions 1-4 are satisfied, identification is lost because under the binary support of Z, the parameters β_{2o} and δ_o are inherently linked. Obviously with binary support there is no possible way to separately identify a non-linear (the exponential component) effect from a linear effect. Therefore specifying an exponential multiplicative heteroskedastic model is naive and illogical in the binary support setting. In fact, it is not possible to discern any non-negative scale function as heteroskedasticity as opposed to the linear mean function, $X\beta_o$.[6] Nevertheless, this concern needs to be addressed when determining conditions for identification.

[6] For any two non-negative scale functions $g_o(Z)$ and $g(Z)$

$$X\left(\beta_o - \frac{g_o(Z)}{g(Z)}\beta\right) = \begin{cases} \beta_{1o} - \frac{g_o(0)}{g(0)}\beta_1 & \text{if } Z = 0 \\ \beta_{1o} + \beta_{2o} - \frac{g_o(1)}{g(1)}(\beta_1 + \beta_2) & \text{if } Z = 1 \end{cases}$$

which implies $\beta_1 = \frac{g_o(0)}{g(0)}\beta_{1o}$ and the second part holds as long as $\beta_2 = \frac{g(1)}{g_o(1)}\beta_{2o} + \beta_{1o}\left(\frac{g(1)}{g_o(1)} - \frac{g_o(0)}{g(0)}\right)$ which does not imply $g(Z) = g_o(Z)$ or $\beta = \beta_o$. Consequently any non-negative scale function cannot be identified as heteroskedasticity separately from a mean effect.

Counter-example: exponential transformation

Unlike the previous counter-example which manipulates the support of (X, Z), this counter-example takes advantage of the functional form of the heteroskedasticity. Suppose $X = (1, \exp(Z))$ and Z is univariate and continuous, then the first part of statement (1.2) becomes,

$$X(\beta_o - \exp(Z(\delta - \delta_o))\beta) = \beta_{1o} + \exp(Z)\beta_{2o} - \exp(Z(\delta - \delta_o))\beta_1 - \exp(Z(1 + \delta - \delta_o))\beta_2$$

If $\delta - \delta_o = -1$ and $\beta_1 = \beta_{2o} = 0$, then any values $\beta_{1o} = \beta_2$ make the above equation equal to 0 for all values of X_1. Alternatively if $\delta - \delta_o = 1$ and $\beta_{1o} = \beta_2 = 0$, then any values $\beta_1 = \beta_{2o}$ also make the above equation equal to 0. This only holds for the exponential transformation because the heteroskedasticity is of exponential form. By imposing the same transformation in the mean term as in the heteroskedastic term, it becomes difficult to differentiate between the mean effect $X\beta_o$ and the heteroskedastic effect $\exp(Z\delta_o)$.

Non-identification in simulation

To illustrate the consequence of non-identification in estimation, the following simulation exercise uses the second counter-example to construct two observationally equivalent data generating processes for a heteroskedastic Probit model. Let $Z \sim N(0, 1)$ and $X = \exp(Z)$, then consider the following two data generating processes:

$$Y_1 = 1\{0 + 0.5X + U_1\}, \qquad \text{where } U_1 \sim N(0, \exp(4Z)) \tag{1.7}$$

$$Y_2 = 1\{0.5 + U_2\}, \qquad \text{where } U_2 \sim N(0, \exp(2Z)) \tag{1.8}$$

According to the analysis given above, these two models are observationally equivalent. The simulation randomly draws a sample of (X, Z) and then computes two different outcomes Y_1,

and Y_2 for the same independent variable sample. Then using the `hetprobit` command in Stata, two estimations are performed, one using the outcomes Y_1 from the first specification and the other uses the outcomes Y_2 from the second specification.

Figure A.2 show the empirical distributions of the parameter estimates for a sample size of 1,000. This plainly demonstrates that the estimator cannot distinguish between the different parameters values that construct the two outcomes. Because the distribution of the estimates for Specification 1 and Specification 2 look identical, one could think there is not divergence in the parameter estimates within a sample between the two data generating processes but looking at the difference of the two parameter estimates (Difference), there appears to be a trimodal distribution. The mass around 0 implies that the outcomes in the two data-generating processes are similar enough that the estimation procedure calculates the same parameter values when the data generating process is formed using two different parameter values. The mass around -0.5 in the first figure and the mass around 0.5 in the second figure show that in some of the samples, the estimator correctly matches the 'true parameter value' to the data generating process. However the remaining mode that occurs symmetrically across 0 shows that the estimator can also incorrectly match the parameter estimates to the alternate data generating process.

But again, this example is trivial in nature in which an empirical researcher may sidestep by flexibly specifying the conditional mean as $W = (1, 1/X)$ with a homogeneous latent error. This specification is observationally equivalent and is identified. The concern is how could one generally show identification in the case of bijective transformed variables that excludes these types of counter-examples.

The two counter-examples show that Conditions 1-4 are not sufficient for identification. They manipulate the support of the random variables and the form of the heteroskedasticity

to lose identification. To better understand why, it is best to re-examine equation (1.4). The left hand side is linear in Z while the right hand side is a logarithmic function of a ratio of X. If there is a defined relationship between X and Z such that the logarithmic function in X is equivalent to a linear function of Z then identification does not hold. In the first example, the logarithmic function of X is necessarily linear because of the binary support. In the second example, the transformation undoes the logarithmic function resulting in a linear function (for specific values of the parameters).

1.4 Identification in a common specification

The previous section provides justification as to why there is no general result on identification for the case of bijective transformations. The concern is that the most prevalent use of multiplicative exponential heteroskedastic models is when there is a bijective transformation between the random vectors. This would require showing identification prior to estimation for every variation of a specification. To provide some assistance in that front, the following shows identification in a general (although not completely general) and commonly used specification.

Example: polynomial transformations

Wanting to allow for a flexibly specified conditional variance function, one might consider polynomial functions as an approximation of the variance function[7]. The following corollary obtains identification in this commonly used specification.

[7]Khan (2013) shows that a heteroskedastic Probit model with a non-parametric conditional variance function is observationally equivalent to a model with median restriction and no distributional assumptions on the latent error. This would motivate flexible specification of the variance function as way to allow flexibility in the latent error distribution.

Corollary 1.4.1. *If $X = (1, X_2)$ in which X_2 is a univariate continuous random variable and $Z = (X_2, X_2^2, ..., X_2^p)$ then under Conditions 1, 3, and 4, the parameters β_o and δ_o are identified.*

Proof. Suppose there exists a $(\beta, \delta) \in \Theta$ such that

$$X(\beta_o - \exp(Z(\delta - \delta_o))\beta) = 0 \tag{1.9}$$

holds for almost all X and Z in their support. By Condition 4, one can rearrange equation (1.9) to

$$X_2(\delta_1 - \delta_{1o}) + X_2^2(\delta_2 - \delta_{2o}) + ... + X_2^p(\delta_p - \delta_{po}) = \ln\left(\frac{\beta_{1o} + X_2\beta_{2o}}{\beta_1 + X_2\beta_2}\right) \tag{1.10}$$

Since X_2 is continuous, taking the $(p+1)^{\text{th}}$ derivative with respect to X_2,

$$0 = (-1)^{p+1}\left[\left(\frac{\beta_{2o}}{\beta_{1o} + X_2\beta_{2o}}\right)^{p+1} - \left(\frac{\beta_2}{\beta_1 + X_2\beta_2}\right)^{p+1}\right]$$

which implies $(\beta_{1o} + X_2\beta_{2o})/(\beta_1 + X_2\beta_2) = X_2\beta_{2o}/X_2\beta_2 = \beta_{2o}/\beta_2$. Plugging into equation (1.10), the right hand side becomes a constant. By Conditions 1 and 3, the equality cannot hold for any non-zero $(\delta - \delta_o)$, thus δ_o is identified. Finally, since Condition 2 is inherently implied by the given specification, the identification of δ_o implies $\beta = \beta_o$. □

Note that one of the most common specifications $X_2 = Z$ is a special case of this result. This result could easily be extended to the cases where X_2 is not univariate and contains discrete random variables.

1.5 Conclusion

It has been widely accepted that a model with multiplicative exponential heteroskedasticity was estimable under Conditions 1 through 4 provided in Section 2. This chapter provides a

proof of identification when the variables are not bijective transformations of one another. But in a more general case, I supply two examples in which those conditions are satisfied but point identification is not obtainable. Consequently, the conditions previously stated in the literature are not sufficient in distinguishing a linear effect from an exponential effect in all cases. To overcome much of the concerns from the lack of a general identification proof, this chapter also provides a proof of identification in a commonly used specification.

In the next chapter, the results here will be utilized to obtain identification for the proposed estimation of an endogenous binary response model. The proposed approach relaxes assumptions that were standard in the literature but I found to be too restrictive in most empirical settings. One of the motivations behind relaxing the assumptions was to allow for potential heteroskedasticity. Obtaining identification in this setting has two challenges (1) relaxing assumptions that were previously used for identification, and (2) identification with multiplicative exponential heteroskedasticity. This chapter provides the foundation for overcoming the second challenge. Therefore the identification strategy in Chapter 2 emphasizes the importance and utility of the results in Chapter 1.

Chapter 2

Relaxing Conditional Independence in an Endogenous Binary Response Model

2.1 Introduction

In recent years, uncovering causal effects has become a cornerstone in economics research. The interest in causality as opposed to mere correlation allows for more plausible policy implications, counter-factual analysis and the disentanglement of causal mechanisms. Endogeneity, correlation between the unobserved heterogeneity and covariates, is prevalent in economic settings and will bias parameter estimates which will ultimately affect the causal interpretations. With more realistic assumptions than those provided in the literature, this chapter proposes a new control function estimator in a binary response setting to address endogeneity.

Binary responses, a 0 or 1 outcome, is a common setting in economics research. For instance, employment, graduating from college, and purchasing decisions, are all be binary outcomes. In order to accurately uncover the true underlying mechanism in a binary response model, many researchers turn to the latent variable set up (sometimes refer to as a hurdle

model) resulting in non-linear estimation. But, treating endogeneity in a non-separable and non-linear setting is not as straight forward as using a "plug-in" instrumental variables estimator in a simple linear regression.

A series of papers (Smith and Blundell (1986), Rivers and Vuong (1988), Blundell and Powell (2004), and Rothe (2009)) have proposed using a control function method in constructing an estimator that appropriately addresses endogeneity. To gain identification, these papers place strong assumptions on the relationships between the latent unobserved errors and the instruments. Essentially they impose an exclusion restriction such that the conditional distribution of the latent error cannot be a function of the instruments. These Control Function assumptions (CF-CI) are equivalent to assuming conditional independence between the latent error and the instrument and are unlikely to hold in an empirical setting.

For instance, in models of labor participation, one may be interested in uncovering the effect of non-wage income on the probability of employment. But there are concerns of endogeneity because one of the main sources of non-wage income is the partner's wages, and their labor force participation decisions are usually simultaneously determined within the household. CF-CI would require that shocks to (non-wage) household income affect labor participation decisions independent of any other included covariates, such as education, age, children in the household, or instruments such as husband's education.

Another example in the field of health economics is evaluating the effect of drug rehabilitation treatment on subsequent substance abuse. There is endogeneity because the covariate of interest, number of visits the client makes during the episode of treatment, is most likely correlated with unobserved characteristics of the client that determine the likelihood of successful treatment. For example, those who are more likely to relapse initially (longer drug use or less community support) are less likely to visit the center during the episode of the

treatment. CF-CI would require that the unobserved characteristics of the client cannot have an interactive effects with other included covariates such as age, income, or marital status.

In a more structural setting, suppose researchers are interested in understanding the welfare loss from government intervention into insurance markets using variation in prices to estimate the demand and marginal cost of insurance. But observed prices are endogenously determined since they are likely correlated with unobserved characteristics. Using exogenous variation in prices, possibly through variation in administrative costs or changes in the competitive environment over markets, endogeneity may be addressed. But, the CF-CI imposes functional form restrictions on the utility function that unobserved characteristics are additively separable from observed market, product or individual characteristics.

This chapter proposes an alternative framework and control function estimator that relaxes this strong assumption. This generalization has been explored in other settings such as the case of endogenous random coefficients for a linear model in Wooldridge (2005) and demand estimation where the unobserved product characteristics does not enter the utility function additively in Gandhi, Kim, and Petrin (2013). More generally, Kim and Petrin (2017) sets up a framework for the "general control function," permitting the unobserved heterogeneity to be a function of the instruments, in the case of additively separable triangular equation models. This chapter extends the general control function approach of Kim and Petrin (2017) to the case of binary response models to propose a new estimator that is valid under the failure of CF-CI.

One of the main contributions of this chapter is to clearly explain why CF-CI would not realistically hold in empirical settings and, given the likely failure of CF-CI, apply the general control function approach of Wooldridge (2005), Gandhi, Kim, and Petrin (2013),

Kim and Petrin (2017) to a binary response setting. A simulation illustrates that given the failure of CF-CI, the general control function approach, as opposed to alternative control function methods of the literature, is needed to accurately recover parameter estimates. Under the more general framework, CF-CI implies testable hypotheses in which standard variable addition or Wald tests can be used. In an empirical application on female labor supply, the CF-CI assumption is easily rejected.

This chapter also adds to the larger literature on control function approaches to triangular simultaneous equations for both additively separable[1] and non-separable models.[2] In the literature, CF-CI has been taken as a required assumption in order to employ a control function approach. In the discussion on identification, this chapter comments on how other control function methods in the literature obtain identification, explains why their approaches to identification can be restrictive, and proposes a simpler alternative. Consequently, this chapter provides an example where, under a reasonable setting, CF-CI with respect to the control variable need not hold to recover structural objects such as the Average Structural Function (ASF) or the Average Partial Effects (APE).

By focusing on the ASF and APE, this chapter contributes to the discussion on interpretation of non-linear models under the presence of endogeneity. Blundell and Powell (2003, 2004) introduced the ASF as a way to interpret binary response models when there is endogeneity. They note that a conditional mean interpretation cannot capture the causal and

[1] Although a latent variable binary response model is non-separable due to the indicator function, separability is imposed inside the indicator function. Consequently results from the additively separable literature may still apply.

[2] Literature on additively separable triangular equation models include Newey, Powell, and Vella (1999), Pinkse (2000), Su and Ullah (2008),Florens, Heckman, Meghir, and Vytlacil (2008), Ai and Chen (2003), Newey and Powell (2003), Newey (2013), Kim and Petrin (2017), and Hoderlein, Holzmann, and Meister (2017). Literature on non-separable triangular equation models include Imbens and Newey (2009), Kasy (2011), Blundell and Matzkin (2014), Chen, Chernozhukov, Lee, and Newey (2014), Kasy (2014), and Hoderlein, Holzmann, Kasy, and Meister (2016).

structural effect that a model incorporating endogeneity should produce. More recently, Lewbel, Dong, and Yang (2012) propose using an Average Index Function (AIF) as a generally easier to identify alternative to the ASF. Lin and Wooldridge (2015) compare the two approaches and conclude the ASF is a more appropriate function for interpretation and is able capture the mechanisms of interest. This chapter further supports the conclusions of Lin and Wooldridge (2015), where it is shown that under the more general framework of this chapter, only the proposed estimation procedure recovers the correct ASF. This chapter also shows that the alternative estimator from Rothe (2009), the Semi-parametric Maximum Likelihood (SML) estimator, actually produces estimates for the AIF, which is shown in simulation to be distinctly and interpretively different from the ASF.

The proposed estimator is presented in a parametric framework but in some empirical contexts, the distributional assumptions may be unrealistic. Therefore this chapter also provides a semi-parametric extension that proposes a new distribution free estimator. Using the observational equivalence results of Khan (2013), the proposed sieve semi-parametric estimator is shown to be consistent under weaker assumptions than those found in the literature. Consequently, this chapter contributes to the literature on semi and non-parametric estimation as a particular application of a semi-parametric two stage sieve estimator. Sieves (as opposed to kernel methods) are suggested in order to impose necessary shape restrictions on the general control function. Asymptotic results are derived using the works of Ai and Chen (2003), Chen, Linton, and Van Keilegom (2003), and Hahn, Liao, and Ridder (2018). A comprehensive simulation study shows that only the proposed estimator can produce accurate parameter and ASF estimates under the failure of CF-CI.

The remainder of this chapter is organized as follows. Section 2 provides motivation for relaxing CF-CI, specifically to the setting of binary response models, and reviews previous

approaches and their potential shortcomings. Section 3 describes the set up of the model and introduces the general control function method of Kim and Petrin (2017) in the binary response setting. Empirical examples are provided to illustrate how CF-CI is unlikely to hold in many economic settings and how the proposed framework captures the potentially complex structure of endogeneity. Section 4 goes into more detail about the operation and interpretation of the general control function approach. Because CF-CI is used to show identification, the generalizations proposed in this chapter put into question whether identification still holds. The Conditional Mean Restriction from Kim and Petrin (2017) that places a shape restriction on the general control function is used to show identification. This section also provides a simulation to illustrate the failure of estimators that require CF-CI when only the weaker CMR assumption holds. Section 5 instructs on the implementation of the proposed estimator and derives the asymptotic properties such as consistency and asymptotic normality. Because the parameters of a binary choice model have no direct economic interpretation, this section also discusses the ASF and APE as structural objects of interest and how to recover them under the proposed framework. Section 6 illustrates the proposed estimator in an empirical application. Using 1991 CPS data, this chapter examines the effect of non-wage income on a married woman's probability of labor force participation. CF-CI implies a testable hypothesis under the proposed framework and a Wald test finds strong statistical evidence that the assumptions of previous estimators are violated. Although the parametric assumptions are likely to hold in the empirical application provided, there are many economic settings where the distributional assumptions are restrictive and unconvincing. The final section extends the framework to a distribution free setting using a semi-parametric estimator. This section provides the asymptotic properties of the semi-parametric estimator as well as a comprehensive simulation study comparing the proposed

approach to other estimators in the literatures.

2.2 Background and Motivation

Consider the latent variable triangular system where y_{1i} is a binary response variable, $\mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$ is a $1 \times (k_1 + k_2)$ vector of "non-endogenous" included and excluded instruments, y_{2i} is a single continuous endogenous regressor, and $\mathbf{x}_i$ is a $1 \times k$ vector where each element is a function of $(\mathbf{z}_{1i}, y_{2i})$ and includes a constant.

$$\begin{aligned} y_{1i} &= \begin{cases} 1 & y_{1i}^* \geq 0 \\ 0 & y_{1i}^* < 0 \end{cases} \\ y_{1i}^* &= \mathbf{x}_i \beta_o + u_{1i} \\ y_{2i} &= m(\mathbf{z}_i)\pi_o + v_{2i} \end{aligned} \tag{2.1}$$

The endogenous variable y_{2i} can be decomposed into its conditional mean and the unobserved endogenous component, v_{2i}. Alternatively one could consider a linear probability model, which has the advantages of being easy to estimate, easy to interpret, and dealing with endogeneity is relatively simple, or at the very least well studied. But linear probability models are restrictive and cannot be representative of the true underlying mechanisms. Their predicted probabilities lie outside the [0,1] bounds which places limitations on the interpretation of the estimates. Therefore this chapter will focus on the latent variable setting.

In this framework, a series of papers, Smith and Blundell (1986), Rivers and Vuong (1988), Blundell and Powell (2004), and Rothe (2009), developed estimators to address endogeneity using a control function approach. The control function approach supposes that there is a particular function (or variable) that when included as an additional covariate,

is able to control for the endogeneity of the other regressors. For example, Rivers and Vuong (1988) shows that if one were to assume that u_{1i} and v_{2i} are bivariate normal and independent of the instruments z_i, then one can derive the following conditional distribution,

$$u_{1i}|v_{2i}, \mathbf{z}_i \sim N\left(\rho\frac{\sigma_1}{\sigma_2}v_{2i}, (1-\rho^2)\sigma_1^2\right) \tag{2.2}$$

where σ_1 and σ_2 are the standard deviations of u_{1i} and v_{2i} respectively and ρ is the correlation coefficient. This conditional distribution provides the foundation for the control function approach in this context. The latent equation can be rewritten as

$$y_{1i}^* = \mathbf{x}_i\beta_o + \gamma_o v_{2i} + \varepsilon_{1i} \tag{2.3}$$

where $\varepsilon_{1i} = u_{1i} - \gamma_o v_{2i}$ and $\gamma_o = \rho\frac{\sigma_1}{\sigma_2}$. Notice that $\varepsilon_{1i}|v_{2i}, \mathbf{z}_i \sim N(0, (1-\rho^2)\sigma_1^2)$ which means there is no endogeneity between the regressors and the new latent error ε_{1i} (i.e.: $E(\varepsilon_{1i}|\mathbf{x}_i, v_{2i}) = 0$). Therefore the reduced form error v_{2i} can be used as a control function; by including v_{2i} as an additional covariate, one can control for the endogeneity in y_{2i} and obtain consistent parameter estimates.

In general, the control function approach constructs a function of the instruments and regressors that can act as a valid proxy for the source of the endogeneity. Of course, in practice, one does not observe v_{2i}, so instead the residuals from a first stage estimation procedure that regresses the endogenous variable on the instruments can be used.

In order to relax the distributional assumptions in Rivers and Vuong (1988), Blundell and Powell (2004) and Rothe (2009) propose alternative semi-parametric estimators. But to obtain non-parametric identification of the distribution of the latent error, they make a rather strong assumption on the relationship between the unobserved errors, u_{1i} and v_{2i}, and the instruments, $\mathbf{z}_i$. This Control Function assumption imposes Conditional Independence

(CF-CI). The CF-CI assumption requires the the instruments $\mathbf{z}_i$ are independent of the latent error u_{1i} after conditioning on the reduced form error v_{2i}:

$$u_{1i}|v_{2i}, \mathbf{z}_i \sim u_{1i}|v_{2i} \tag{2.4}$$

Note that CF-CI is implicit in the set up of Rivers and Vuong (1988). Interpretively, this means that any source of endogeneity must be fully captured through the control variate v_{2i}, or in terms of an exclusion restriction, the conditional CDF $F_{u_1|v_2,\mathbf{z}}(u_{1i}|v_{2i}, \mathbf{z}_i)$ is only a function of v_{2i} (the instruments $\mathbf{z}_i$ are excluded). This exclusion restriction must also hold for all moments which, as will be shown shortly, may be hard to justify.

As a slight relaxation of CF-CI, Rothe (2009) also proposes a Linear Index (CF-LI) sufficiency assumption that, after conditioning on the first stage error and the linear index $x_i\beta_o$, the latent error is independent of the instruments:

$$u_{1i}|v_{2i}, \mathbf{z}_i \sim u_{1i}|v_{2i}, \mathbf{x}_i\beta_o \tag{2.5}$$

Now the instruments can be a part of the conditional distribution but only through the linear index. The linear index restricts the relative direction and magnitudes of the regressors in the conditional distribution. So, although it allows for a more relaxed relationship between the instruments and the unobserved heterogeneity, it is hard to justify in a general setting.

In either case, these assumptions used to obtain identification may be too stringent in many empirical contexts. To give a motivating parametric example, consider a slight variation of the Rivers and Vuong (1988) set up where u_{1i} and v_{2i} are still bivariate normal but are allowed to be heteroskedastic in the instruments; i.e.,

$$\begin{matrix} u_{1i} \\ v_{2i} \end{matrix}\Bigg| \mathbf{z}_i \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} [\sigma_1(\mathbf{z}_i)]^2 & \rho(\mathbf{z}_i)\sigma_1(\mathbf{z}_i)\sigma_2(\mathbf{z}_i) \\ \rho(\mathbf{z}_i)\sigma_1(\mathbf{z}_i)\sigma_2(\mathbf{z}_i) & [\sigma_2(\mathbf{z}_i)]^2 \end{pmatrix}\right) \tag{2.6}$$

Heteroskedasticity is commonly found in empirical data whether it is actually caused by variability in the latent error over the regressors, or by heterogeneity in the slopes as in a random coefficients setting.[3] Even in the linear regression, heteroskedasticity has been accepted as endemic in empirical settings and heteroskedastic robust inference is always employed. Again, by the properties of the bivariate normal distribution, the following conditional distribution is derived.

$$u_{1i}|v_{2i}, \mathbf{z}_i \sim N\left(\rho(\mathbf{z}_i)\frac{\sigma_1(\mathbf{z}_i)}{\sigma_2(\mathbf{z}_i)}v_{2i}, (1-[\rho(\mathbf{z}_i)]^2)[\sigma_1(\mathbf{z}_i)]^2\right) \tag{2.7}$$

This is a fairly small variation to the framework considered in Rivers and Vuong (1988) but ignoring heteroskedasticity can strongly bias parameter estimates. A simple Monte Carlo exercise, detailed in Figure D.1, illustrates the potential bias. Suppose equations (2.1) and (2.6) hold with a single excluded instrument and no included instruments such that $\rho(z_i) = 0.6, \sigma_1(z_i) = \sigma_2(z_i) = \exp(0.25z_i)$. Figure D.1 displays the empirical distribution of the estimators for β where the true value is equal to one. The Rivers and Vuong estimator that ignores heteroskedasticity is substantially biased with estimates of β centered around 1.2. This illustrates that ignoring heteroskedasticity in the context of binary response models produces inconsistent parameter estimates.

Now let us compare the distribution in equation (2.7) to the CF-CI and CF-LI assumptions for the semi-parametric estimators. CF-CI clearly does not hold since the exclusion restriction does not hold: both the conditional mean and conditional variance depend on the instruments. For the CF-LI assumption to hold, the heteroskedastic functions $\sigma_1(\cdot), \sigma_2(\cdot)$,

[3]This is similar to the set up in Kasy (2011) where he provides a counter-example to the control function approach proposed by Imbens and Newey (2009). In Imbens and Newey (2009), they propose using the control variable $V_i = F_{y_1|z}(y_{1i}, z_i)$ which would satisfy CF-CI when the heterogeneity is only one-dimensional, as pointed out in Kasy (2011). In his example, a linear random coefficient model is used to show the failure of CF-CI using the Imbens and Newey control variable. Note that the random coefficient model can be rewritten as a linear model with heteroskedasticity as suggested here.

and $\rho(\cdot)$ could only be functions of the linear index $\mathbf{x}_i\beta_o$. This is quite restrictive and would not generally hold. Therefore the semi-parametric estimators do not apply to this simple parametric setting.

This causes some concern that the control function method may not be valid for this example. But the conditional distribution in equation (2.7) suggests that there should still be a control function approach to address endogeneity. If one were to include $\rho(\mathbf{z}_i)\frac{\psi_1(\mathbf{z}_i)}{\psi_2(\mathbf{z}_i)}v_{2i}$ as an additional covariate, then estimating a heteroskedastic Probit with the following conditional mean will produce consistent parameter estimates.

$$E(y_{1i}|\mathbf{z}_i, v_{2i}) = \Phi\left(\frac{\mathbf{x}_i\beta_o + \rho(\mathbf{z}_i)\frac{\psi_1(\mathbf{z}_i)}{\psi_2(\mathbf{z}_i)}v_{2i}}{\sqrt{(1-[\rho(\mathbf{z}_i)]^2)[\psi_1(\mathbf{z}_i)]^2}}\right) \tag{2.8}$$

Building a control function method from a more general conditional distribution of the latent error is the motivation and starting point for the proposed approach.

This example highlights that the assumptions used to obtain identification in the semi-parametric approaches can be fairly stringent and are not well understood in terms of their consequences. This chapter questions the necessity of the CF-CI assumption and considers its implications in estimation and interpretation. As an alternative to the semi-parametric estimators, I propose an estimator that builds upon the same control function technique but extends the model by relaxing the CF-CI assumption. In the previous heteroskedastic bivariate normal example, this means the instruments can be a part of the conditional variance and the conditional mean.

But the CF-CI and CF-LI assumptions are not imposed superfluously, they are used to obtain identification of the conditional distribution of the latent error. In this chapter, I will first consider a parametric alternative to isolate the necessary conditions for identification. At the end of this chapter, I present a distribution free extension that proposes a semi-parametric

estimator that has strictly weaker identification requirements compared the estimators of Blundell and Powell (2004) and Rothe (2009).

In a related strand of literature on non-parametric triangular simultaneous equation models with additively separable unobserved heterogeneity, Kim and Petrin (2017) question the restrictive control function assumptions in the Non-Parametric Control Function (NP-CF) literature (Newey, Powell, and Vella (1999), Pinkse (2000), Su and Ullah (2008), and Florens, Heckman, Meghir, and Vytlacil (2008)). Because the control function method requires the additional control function assumption for identification, the Non-Parametric Instrumental Variables (NP-IV) approach, as in Ai and Chen (2003), Newey and Powell (2003), Hall, Horowitz, et al. (2005), and Newey (2013), appears to be a superior approach that only requires the weaker Conditional Mean Restriction (CMR). Kim and Petrin (2017) show that a control function approach is still valid under the weaker CMR when a general control function is specified. This chapter extends their results to a binary response model under a latent variable framework and will use the CMR in showing identification.

Alternative estimators that do not require the CF-CI assumption in estimating endogenous binary response models are the special regressor estimator proposed in Lewbel (2000) and Dong and Lewbel (2015) and the maximum score and smoothed maximum score estimator in Hong and Tamer (2003) and Krief (2014). The special regressor estimator requires a regressor, independent of all unobserved heterogeneity, that has large support and without this "special regressor" their procedure is invalid. Alternatively, Hong and Tamer (2003) and Krief (2014) extend the maximum score and smoothed maximum score methods of Manski (1985) and Horowitz (1992) to estimate the structural parameters β_o in the linear index. For identification they require conditional median independence: $\text{Med}(u_{1i}|v_{2i}, \mathbf{z}_i) = \text{Med}(u_{1i}|v_{2i})$. This would allow for general forms of heteroskedasticity

but, as seen in the heteroskedastic bivariate normal example, the conditional median independence assumption is still quite restrictive and would not necessarily hold. Moreover, the conditional median independence assumption does not identify the distribution of the latent error and therefore they cannot recover the ASF and APE.

The proposed framework that allows for relaxation of CF-CI and CF-LI in a parametric setting is introduced in the next section. The proposed general control function estimator directly follows from the conditional distribution of the latent error provided in the framework.

2.3 Model Set Up

Return to the set up described in equation (2.1). The distributional assumptions for u_{1i} and v_{2i} determine the consistent estimation procedure. Although most of the assumptions in the literature are based on a specification of the joint distribution of u_{1i} and v_{2i} (see Rivers and Vuong (1988) and Petrin and Train (2010)), one merely needs to specify the conditional distribution to use the control function approach. For example, if one were to assume $u_{1i}|v_{2i}, \mathbf{z}_i \sim N(0,1)$, so there is no endogeneity and no heteroskedasticity, then a standard Probit maximum likelihood estimation (MLE) procedure yields consistent estimates. On the other hand, if $u_{1i}|v_{2i}, \mathbf{z}_i \sim N(0, \exp(2\mathbf{z}_i\delta))$ such that heteroskedasticity is present, then the standard Probit MLE procedure would be inconsistent but a Het-Probit MLE procedure, included in many statistical packages, would be consistent. If $u_{1i}|v_{2i}, \mathbf{z}_i \sim N(\rho v_{2i}, 1)$, similar to the setting in equation (2.2), then two step CMLE developed by Smith and Blundell (1986) and Rivers and Vuong (1988) would be consistent and other methods that ignore the endogeneity would be inconsistent. More generally, if the CF-CI assumption holds such

that $u_{1i}|v_{2i}, \mathbf{z}_i \sim u_{1i}|v_{2i}$ with some unknown distribution, Blundell and Powell (2004) (for the remainder of the chapter referred to as BP) and Rothe (2009) provide semi-parametric methods that estimates the parameters consistently.

As a first step in relaxing the CF-CI assumption, the following assumption proposes an alternative framework which assumes a more flexible conditional distribution of the latent error.[4]

Assumption 2.3.1. *Consider the set up in equation (2.1), where* $\{y_{1i}, \boldsymbol{z}_i, y_{2i}\}_{i=1}^n$*, is iid. Assume the linear reduced form in the first stage is the true conditional mean*

$$E(y_{2i}|\boldsymbol{z}_i) = m(\boldsymbol{z}_i)\pi_o$$

and the unobserved latent error has the following conditional distribution

$$u_{1i}|\boldsymbol{z}_i, v_{2i}, y_{2i} = u_{1i}|\boldsymbol{z}_i, v_{2i} \sim N\Big(h(v_{2i}, z_i)\gamma_o, \exp(2g(y_{2i}, \boldsymbol{z}_i)\delta_o)\Big)$$

Where $\boldsymbol{z}_i = (\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})$ *and* $m(\boldsymbol{z}_i)$*,* $h(v_{2i}, \boldsymbol{z}_i)$*, and* $g(y_{2i}, \boldsymbol{z}_i)$ *are known vectors and* $h(v_{2i}, \boldsymbol{z}_i)$ *is differentiable in* v_{2i}*.*

The first part of the assumption breaks up the endogenous variable into its conditional mean and what I will refer to as the control variate v_{2i}. Note that by construction, the control variate is mean independent of the instruments. This assumption does not take a stand on the true data generating process of the endogenous variable.[5] In the more general setting of non-separable triangular equation models, Imbens and Newey (2009) consider the

[4]The normality assumption could be easily generalized to just a known distribution with CDF $G(\cdot)$. This allows for a logit specification which is also explored in one of the simulations.

[5]For now this does presume a linear reduced form, but when discussing asymptotic properties of the estimator, if $m(\mathbf{z}_i)$ is a sequence of basis function so that the first stage acts as a non-parametric sieve regression then this will not affect the asymptotic variance estimates.

case of a non-separable first stage

$$y_{2i} = d(\mathbf{z}_i, \eta_i) \tag{2.9}$$

where z_i are the instruments, η_i is unobserved heterogeneity independent of the instruments, and $d(\cdot, \cdot)$ is the unknown and true data generating process in the first stage. In this setting they suggest using the conditional CDF, $e_{2i} = F_{y_2|\mathbf{z}}(y_{2i}, \mathbf{z}_i)$, as the control variable. They show that their proposed control variable satisfies CF-CI and therefore the control function method recovers the parameters of the model. Assumption 2.3.1 does not require full independence between the control variable and the instruments and therefore can use the population residual $v_{2i} = y_{2i} - E(y_{2i}|\mathbf{z}_i)$ as a control variable with the knowledge that it does not satisfy CF-CI. I will discuss the differences between these two approaches after explaining the second part of the assumption.

The second part of Assumption 2.3.1 specifies the conditional distribution that allows for the violation of CF-CI. Both the conditional mean and the conditional variance are functions of the instruments, so the exclusion restriction implied by CF-CI is violated. Under this assumption, the conditional mean of y_{1i} is:

$$E(y_{1i}|\mathbf{z}_i, y_{2i}, v_{2i}) = E(y_{1i}|\mathbf{z}_i, v_{2i}) = \Phi\left(\frac{\mathbf{x}_i\beta_o + h(v_{2i}, \mathbf{z}_i)\gamma_o}{\exp(g(y_{2i}, \mathbf{z}_i)\delta_o)}\right) \tag{2.10}$$

Note that there is a one-to-one mapping between y_{2i} and v_{2i} given the instruments $\mathbf{z}_i$. This implies the mean is preserved regardless of which term is included in the conditioning argument. This result should be unsurprising as the conditional mean appears to be a heteroskedastic Probit model that adjusts for endogeneity using the control function approach, both of which have been discussed extensively in the literature. But in this case, the control function $(h(v_{2i}, \mathbf{z}_i)\gamma_o)$ is a function of both the control variate v_{2i} and the instruments $\mathbf{z}_i$.

This was first introduced in Wooldridge (2005) where he suggests using the following control function,

$$h(v_{2i}, \mathbf{z}_i)\gamma_o = \gamma_{1o} v_{2i} + v_{2i}\mathbf{z}_i\gamma_{2o} \tag{2.11}$$

in a linear regression with random coefficients.[6] Gandhi, Kim, and Petrin (2013) adopted a similar generalization for demand estimation and Kim and Petrin (2017) provides a general control function framework for the case of non-linear but additively separable triangular equation models. As in Kim and Petrin (2017), this generalization will be referred to as the "general control function," as opposed to a more traditional control function that upholds the exclusion restriction (not a function of the instruments) as in Rivers and Vuong (1988) and Petrin and Train (2010).

The proposed framework suggests a simple two step estimator. In the first step, the conditional mean of y_{2i} is estimated to construct the control variate from the residuals ($\hat{v}_{2i}$). In the second step, the residuals are plugged into the conditional mean in equation (2.10) in which parameter estimates are obtained via maximum likelihood estimation. This will be discussed with more detail in Section 5.

How does the proposed approach differ from the setting considered in Imbens and Newey (2009)? In Imbens and Newey (2009), they attempt to flexibly model the true data generating process of the first stage as a possibly non-separable function of instruments and unobserved heterogeneity.[7] They then construct a control variable, $e_{2i} = F_{y_2|\mathbf{z}}(y_{2i}, \mathbf{z}_i)$, that they show satisfies the CF-CI assumption. But they require the instruments to completely independent

[6] He also discusses in this chapter the implementation of the control function approach in a binary response setting such as Probit. But in that example, he does not propose interaction with the instruments and instead only suggests including higher order moments of the reduced form error. So his analysis stops short of what is proposed in this chapter.

[7] The non-separable first stage needs to be monotonic in the unobserved heterogeneity.

of any unobserved heterogeneity and only a single source of unobserved heterogeneity in the first stage. In this chapter, I use a control variate v_{2i} that is always obtainable and must satisfy conditional mean independence, by construction. Then to make up for the relaxation of CF-CI, I flexibly model the relationship between the structural heterogeneity u_{1i}, the control variable v_{2i}, and the instruments $\mathbf{z}_i$ using a general control function.

A major critique to the approach of Imbens and Newey (2009) is the caveat to their framework brought up in Kasy (2011) noting their method only allows for one source of heterogeneity (independent of the instruments) in the first stage. This would prohibit the simple example of random coefficients in the first stage

$$y_{2i} = \eta_{1i} + \eta_{2i}\mathbf{z}_i \tag{2.12}$$

The approach in this chapter allows for this possibility since equation (2.12) can be rewritten in terms of a linear conditional mean with heteroskedasticity in the first stage error.

One may object to the linear in parameters and known distribution specifications in Assumption 2.3.1. If these specifications are not true, this leads to the misspecification of equation (2.10) as the true conditional mean. The general control function and heteroskedastic function, $h(v_{2i}, z_i)\gamma_o$ and $g(y_{2i}, z_i)\delta$, respectively, are assumed to be linear in parameters, which facilitates the identification discussion because it allows for lower level conditions. Alternatively, one can consider any parametric specification, but then lower level conditions for identification would need to be derived to fit the specification. In the extension provided in Section 6, I allow both the general control function and the heteroskedastic function to be non-parametrically specified.

The distribution assumption is particular pertinent in contrast to estimators from BP and Rothe, that have no distributional assumptions. Preserving the distributional assumption

keeps the difficult discussions on identification and interpreting the ASF and APE clear.[8] But to appease any concerns, the extension provided in Section 7 proposes a semi-parametric estimator that is free of any distributional assumptions.

Up till now I have only explained theoretically the consequences of CF-CI in terms of exclusion restrictions on the conditional distribution. But as a researcher with empirical data, how is one to determine why CF-CI may fail to hold? To further motivate the generalization provided in Assumption 2.3.1, the following are two examples taken from applications in the literature where their empirical settings may suggest a violation of CF-CI.

Example 1: Demand for Premium Cable from Petrin and Train (2010)

This example is a simplified version of the application given in Petrin and Train (2010) (hereafter PT), who propose a control function approach for estimating structural models of demand. In their application, they use a multinomial logit in modelling consumer's choice of television reception. To fit a binary response setting, consider the choice of selecting premium cable conditional on already selecting cable as the television reception. Let U_{im} be the marginal utility of individual i choosing premium cable in market m over the utility from not selecting premium cable (so the utility from the outside option is normalized to 0). Violation of the CF-CI can be easily invoked by allowing for a utility that is not additively separable between unobserved utility (u_{1im}) and the observed utility (U_{im}) as in Gandhi, Kim, and Petrin (2013). Suppose the observed utility is

$$U_{im} = \beta_1 p_m + \sum_{g=2}^{5} \beta_{2g} p_m d_{gi} + \mathbf{z}_{11m}\beta_3 + \mathbf{z}_{12i}\beta_4 + (1 + p_m\gamma_1 + \mathbf{z}_{11m}\gamma_2 + \mathbf{z}_{12i}\gamma_3)u_{1im} \quad (2.13)$$

where the variables in $\mathbf{z}_{11m}$ include the market and product characteristics and the variables

[8] An added benefit is the proposed estimator is much easier to implement compared to the semi-parametric approaches. For instance, estimates can be obtained using canned commands in Stata. Hopefully this will persuade empirical economists that implementing generalizations to previous estimators need not be computationally burdensome.

in $\mathbf{z}_{12i}$ include individual characteristics. The variables d_{gi} are dummies of an index of 5 different income levels, this allows price elasticity to be heterogeneous in income. The unobserved utility consists of two components $u_{1im} = \xi_m + \varepsilon_{im}$ where ε_{im} is iid logistic while ξ_m represents unobserved (to the researcher but not to the consumer or producer) attributes of the product. Consequently, ξ_m captures the component of the unobserved utility that is not independent from price. Note that this specification, like Gandhi, Kim, and Petrin (2013), allows for potential interactions between the observable covariates (price, market and product characteristics, and individual characteristics) and the unobserved attributes of the product.

This specification, previously discussed in Gandhi, Kim, and Petrin (2013), can be motivated using the example of unobserved advertisement. For instance, one would expect not only unobserved advertisement to affect utility (though ξ_m) but would also expect a interactive effect with product characteristics. For example, suppose premium cable is marketed with advertisement that emphasizes the number of channels provided. Then advertisement should contribute to utility of consumption interactively with the number of premium cables actually provided.

Even if a researcher were to impose an additively separable form on the utility, it is still unlikely that a simple control function from a reduced form pricing equation may capture the true endogenous structure. Suppose the utility from purchasing premium is

$$U_{im} = \beta_1 p_m + \sum_{g=2}^{5} \beta_{2g} p_m d_{gi} + \mathbf{z}_{11m}\beta_3 + \mathbf{z}_{12i}\beta_4 + u_{1im} \tag{2.14}$$

where the unobserved utility is composed of two components: $u_{1im} = \xi_m + \varepsilon_{im}$, ε_{im} is iid logistic while ξ_m is represents unobserved attributes of the product. The probability

consumer i chooses premium cable in market m is

$$\begin{aligned} P_{im} &= P(U_{im} > 0 | p_m, d_{gi}, z_{11m}, z_{12i}, \xi_m) \\ &= \frac{\exp(\beta_1 p_m + \sum_{g=2}^{5} \beta_{2g} p_m d_{gi} + z_{11m}\beta_3 + z_{12i}\beta_4 + \xi_m)}{1 + \exp(\beta_1 p_m + \sum_{g=2}^{5} \beta_{2g} p_m d_{gi} + z_{11m}\beta_3 + z_{12i}\beta_4 + \xi_m)} \end{aligned} \tag{2.15}$$

and the expected demand from the perspective of the monopolist be $E(P_{im}|p_m, z_{11m}, \xi_m)$. A monopolist will maximize expected profit with respect to price

$$p_m = \arg\max_p (p - MC(z_{2m}, \omega_m)) E(P_{im}|p, z_{11m}, \xi_m) \tag{2.16}$$

From the first order conditions, the optimal price satisfies

$$p_m = \frac{p_m}{|e(z_{11m}, \xi_m)|} + MC(z_{2m}, \omega_m) \tag{2.17}$$

where $e(z_{11m}, \xi_m)$ is the price elasticity of demand. It is evident that prices are not separable in ξ_m and the exogenous characteristics z_{11m}, z_{2m}. If one were to still use the control variable $v_{2m} = p_m - E(p_m|z_{11m}, z_{2m})$ then the CF-CI assumption implies $E(\xi_m|z_{11m}, z_{2m}, v_{2i}) = E(\xi_m|v_{2m})$ and would generally not hold. Therefore the estimators based on the CF-CI assumption would not be valid in this setting. Kim and Petrin (2017) provide a similar example to motivate their general control function in non-linear but additively separable setting.

Example 2: Home-ownership and Income from Rothe (2009)

This example is the application in Rothe (2009) where he considers the effect of income on home-ownership in Germany for low-educated middle age married men. The controls, $\mathbf{z}_{1i}$, included age and an indicator for the presence of children under the age of 16. The instruments, $\mathbf{z}_{2i}$, are wife's education level and an indicator for wife's employment status which should only effect home-ownership through family income. Rothe relaxes CF-CI slightly by proposing the alternative CF-LI assumption. Recall, that the CF-LI assumption requires

the conditional distribution of the latent error to only be a function of the control variable v_{2i} and the linear index $\mathbf{x}_i\beta_o = \mathbf{z}_{1i}\beta_{1o} + y_{2i}\beta_{2o}$. In this example, endogeneity can be explained by omitted variables such as accessibility to loans via credit that are correlated with income. Moreover, as the accessibility to loans via credit lowers (i.e., low credit score), the effect of income and whether or not you have children becomes less important in the decision to purchase a home. So if credit score was observable, one would expect interactive effects between the linear index and credit score. Since v_{2i} acts as a proxy for the omitted variable, the conditional mean of the latent error should include the interactive effect,

$$E(u_{1i}|v_{2i}, \mathbf{x}_i\beta) = v_{2i}\gamma_1 + v_{2i}(\mathbf{x}_i\beta_o)\gamma_2 \tag{2.18}$$

Under this specification, the CF-LI assumption in Rothe is satisfied while CF-CI is violated. However, this places fairly strong restrictions on the coefficients of interactive effects between the omitted variable–credit score– and the included regressors –age, presence of children, and income– such that they must be proportional to the linear index coefficients, β_o. Alternatively, the proposed estimation procedure would recognize the interactive relationship of these effects and could also allow the interactions to have effects not necessarily proportional to the index coefficients.

These two examples provide some economic motivation for the relaxation of the CF-CI assumption. However the CF-CI assumption was used in the literature to gain identification. In equation (2.10), $\mathbf{x}_i$ is a function of $\mathbf{z}_{1i}$ and y_{2i} which both comprise the control function $h(v_{2i}, \mathbf{z}_i)$. Without any restrictions on the control function, the two effects may not be separately identifiable. Wooldridge (2005) notes that the exclusion of $\mathbf{z}_{2i}$ in the structural equation allows for identification of the general control function considered in equation (2.11). I will use the more general CMR from Kim and Petrin (2017) to show identification of the

general control function which also helps to illustrate which general control functions are or are not identified.

2.4 General Control Function

The previous section set up the framework for using a general control function approach in a binary response model. In contrast to other control function methods, the general control function allows for the relaxation of the CF-CI assumption. This section is composed of two parts. The first part explains how identification can still be obtained under CMR and how the CMR relates to the other control function assumptions in the literature. The second part is a short simulation to illustrate how the general control function will aid in estimation when CF-CI does not hold but the true data structure satisfies CMR. In this simulation I emulate the application in Petrin and Train (2010) concerning the demand for cable as empirical context.

2.4.1 Identification

Recently, there has been growing interest in the question of identification for the control function approach in non-parametric non-separable triangular simultaneous equation models.[9] However, the discussion usually starts with independence assumptions between the instruments and the unobservables. Then one searches for a control function that will satisfy strong identification assumptions such as CF-CI in BP or CF-CI and monotonicity in Imbens and Newey (2009). In the setting considered here, I allow for a more flexible relationship between the instruments and the unobserved heterogeneity and then allow for a

[9] Imbens and Newey (2009), Kasy (2011), Hahn and Ridder (2011), Blundell and Matzkin (2014), Chen, Chernozhukov, Lee, and Newey (2014), Torgovitsky (2015), and D'Haultfœuille and Février (2015)

general control function, $h(v_{2i}, \mathbf{z}_i)\gamma_o$, that can address endogeneity in a flexible manner. Of course, since I am only concerned with a binary response model, there are gains to knowing the structure of the non-separability in the outcome equation.

The main concern for identification is separately identifying the mean effect $\mathbf{x}_i\beta_o$ and the control function $h(v_{2i}, \mathbf{z}_i)\gamma_o$. Because both of these terms are perfectly determined by $\mathbf{z}_i$ and v_{2i}, without any additional assumptions on the construction of $h(v_{2i}, \mathbf{z}_i)$, perfect multicollinearity is possible such that the parameters β_o and γ_o are not identified.[10] When linearity of the control function is imposed, as in Assumption 2.3.1, identification requires $E((\mathbf{x}_i, h(v_{2i}, \mathbf{z}_i))'(x_i, h(v_{2i}, \mathbf{z}_i)))$ is non-singular.[11] However that does not place clear restrictions on the composition of the control function. The following assumption provides lower level conditions in which identification is shown.

Assumption 2.4.1. *Let $\pi_o \in \Pi$ and $\beta_o, \gamma_o, \delta_o \in \Theta$ where Π and Θ denote the respective parameter spaces.*

(i) *$E(m(\boldsymbol{z}_i)'m(\boldsymbol{z}_i))$ is non-singular*

(ii) *$E(\boldsymbol{x}_i'\boldsymbol{x}_i)$ is non-singular and the variance-covariance matrix of $E(\boldsymbol{x}_i|\boldsymbol{z}_i)$ has full rank.*

(iii) *$E(h(v_{2i}, \boldsymbol{z}_i)'h(v_{2i}, \boldsymbol{z}_i))$ is non-singular.*

(iv) *(CMR) $E(h(v_{2i}, \boldsymbol{z}_i)|\boldsymbol{z}_i) = 0$*

(v) *$g(y_{2i}, \boldsymbol{z}_i)$ consists of polynomial functions of the elements in $(\boldsymbol{x}_i, h(v_{2i}, \boldsymbol{z}_i))$, does not include a constant, and $E(g(y_{2i}, \boldsymbol{z}_i)'g(y_{2i}, \boldsymbol{z}_i))$ is non-singular.*

(vi) *(β_o', γ_o') is a non-zero vector.*

The first condition insures identification of the first stage parameters. The next three

[10] For example, if $\mathbf{x}_i = (1, \mathbf{z}_{1i}, y_{2i})$ then a general control function of the form $h(v_{2i}, \mathbf{z}_i) = (\mathbf{z}_{1i}, \mathbf{z}_{2i}, v_{2i})$ creates perfect multicollinearity. Even when $\mathbf{z}_{1i}$ is excluded from the general control function (so $\mathbf{x}_i$ and $h(v_{2i}, \mathbf{z}_i)$ do not include the same terms) there is multicollinearity when $y_{2i} = \pi_1 + \mathbf{z}_{2i}\pi_2 + v_{2i}$.

[11] Alternatively if one where to assume that the control function and the heteroskedastic function were non-linear functions then one can verify the rank conditions from Rothenberg (1971) for identification

conditions are used to show $E((\mathbf{x}_i, h(v_{2i}, \mathbf{z}_i))'(\mathbf{x}_i, h(v_{2i}, \mathbf{z}_i)))$ is non-singular. The CMR is the more realistic identification assumption used in Kim and Petrin (2017). The last two conditions help in showing identification in the highly non-linear heteroskedastic Probit model. The following theorem states the identification result.

Theorem 2.4.1. *In the set-up described by equation (2.1) and Assumption 2.3.1, if Assumption 2.4.1 holds then the parameters π_o and $(\beta_o, \gamma_o, \delta_o)$ are identified.*

The proof of Theorem 2.4.1 is provided in the Appendix. The CMR approach to obtain identification using a control function is adopted from Kim and Petrin (2017) where they show non-parametric identification following control function approach in a triangular system with an additively separable error.[12]

The CMR can be interpreted as a way to distinguish between the endogeneity of y_{2i} and the "non-endogeneity" of $\mathbf{z}_i$. By law of iterated expectations,

$$E(u_{1i}|\mathbf{z}_i) = E(E(u_{1i}|\mathbf{z}_i, v_{2i})|\mathbf{z}_i) = E(h(v_{2i}, \mathbf{z}_i)\gamma_o|\mathbf{z}_i) = 0$$

The middle equality holds by the specification provided in Assumption 2.3.1 and the last equality holds by the CMR. As a result, the CMR only implies $\mathbf{z}_i$ is mean independent of u_{1i} and does not require any stronger forms of independence. This is a fairly standard and weak exogeneity assumption on an instrument. In practice, if one is concerned that this restriction is violated then the included instrument, $\mathbf{z}_{1i}$, should be treated as an endogenous variable and the excluded instruments, $\mathbf{z}_{2i}$, should not be used as valid instruments.

[12] Hahn and Ridder (2011) show that a "Conditional Mean Restriction" is insufficient for identifying the ASF in a general non-parametric non-separable model. However I would like to be clear that the CMR they consider is

$$E(y_{1i} - \Psi(\mathbf{x}_i)|\mathbf{z}_i) = 0$$

where $\Psi(\mathbf{x}_i)$ is the unknown ASF. This differs from the CMR consider here which is on the latent error. Although the binary response model is non-separable, since the latent error is additively separable from the mean component $\mathbf{x}_i\beta_o$ within the indicator function, identification follows analogously from Kim and Petrin (2017)

To provide some intuition for the implications, because v_{2i} is mean independent of $\mathbf{z}_i$, the CMR requires each element of $h(v_{2i}, \mathbf{z}_i)$ to include functions of v_{2i} and to be conditionally demeaned. For instance, v_{2i}^2 could not be an element of the control function, but $v_{2i}^2 - E(v_{2i}^2|\mathbf{z}_i)$ could be. In addition, no element can only be a function of $\mathbf{z}_i$ alone – the instruments can only enter as an interaction with functions of v_{2i}. Notice that in the examples provided in the previous section the general control functions satisfy the CMR. This prevents any issues of linear dependence between elements of $\mathbf{x}_i$ and $h(v_{2i}, \mathbf{z}_i)$. Wooldridge (2005) explains that identification holds given exclusion restriction on the instruments $\mathbf{z}_{2i}$ in the structural equation that creates variation in the control variate unexplained by $\mathbf{x}_i$.[13] Consequently, the extra variation in the control variate needs to be used to identify the parameters in the general control function. This can be demonstrated explicitly under the assumptions stated above. Let $(a', b')'$ be a non-random vector such that

$$\begin{pmatrix} \mathbf{x}_i & h(v_{2i}, \mathbf{z}_i) \end{pmatrix} \begin{pmatrix} a \\ \\ b \end{pmatrix} = \mathbf{x}_i a + h(v_{2i}, \mathbf{z}_i) b = 0$$

Taking the conditional expectation with respect to $\mathbf{z}_i$,

$$E(\mathbf{x}_i|\mathbf{z}_i)a + E(h_i|\mathbf{z}_i)b = E(\mathbf{x}_i|\mathbf{z}_i)a = 0$$

Because the variance-covariance matrix of $E(\mathbf{x}_i|\mathbf{z}_i)$ is full rank, a is a zero vector and it follows that b is also a zero vector.

Now how does the CMR compare to the CF-CI? In the heteroskedastic bivariate probit example it is easy to see how CF-CI is violated while the CMR continues to hold. However, the CMR is not strictly weaker in the technical sense where CF-CI implies CMR.[14] However,

[13] An alternative identification strategy is used in Escanciano, Jacho-Chávez, and Lewbel (2016) that does not require an exclusion restriction on the instruments z_{2i}. But in their setting, identification is dependent on non-linearity in the reduced form and they still impose CF-CI as a control function assumption.

[14] I would like to thank David Kaplan for pointing this out.

given the earlier discussion, the CMR is more in line with what our prior beliefs on what endogeneity is. Consider the following example in which CF-CI holds but CMR does not. Let $E(u_{1i}|\mathbf{z}_i, v_{2i}) = v_{2i} + v_{2i}^2 - \sigma_v^2$ where $\sigma_v^2 = Var(v_{2i})$ such that CF-CI holds. But suppose there is heteroskedasticity in the first stage such that $E(v_{2i}^2|\mathbf{z}_i) \neq \sigma_v^2$ which implies $E(v_{2i} + v_{2i}^2 - \sigma_v^2|\mathbf{z}_i) = E(v_{2i}^2|\mathbf{z}_i) - \sigma_v^2 \neq 0$ and the CMR does not hold. Interpretively what does this mean? It means that the specification of endogeneity is quadratic in the first stage residual while the quadratic term is deviations from the unconditional variance even when there is heteroskedasticity in the first stage. So for the CMR to fail, the endogeneity must depend on $v_{2i}^2 - \sigma_v^2$ instead of $v_{2i}^2 - E(v_{2i}^2|\mathbf{z}_i)$, deviations for the conditional variance. Therefore I would argue that although the CMR is not strictly weaker than CF-CI, it better reflects how we perceive endogeneity and is much more plausible to hold in empirical settings.

By now, it seems that the relaxation of CF-CI, especially compared to the parametric approach of Rivers and Vuong (1988), is fairly straightforward. Putting aside heteroskedasticity for a moment, the difference is only including the control variate $\hat{v}_{2i}$ as an addition covariate, as suggested in Rivers and Vuong (1988), or including terms such as $(\hat{v}_{2i}, \hat{v}_{2i}^2 - \hat{E}(v_{2i}^2|\mathbf{z}_i), \mathbf{z}_i\hat{v}_{2i})$, as the general control function approach proposed in this chapter. One may wonder whether the relaxation of CF-CI to allow for a general control function is really necessary and whether it would have an impact empirically. The following simulation aims to show the importance of allowing for a general control function when it is called for. The results of the simulation suggest that there is a high cost to not specifying a general control function method when CF-CI fails, but there is very little cost in allowing for a more flexible specification of the general control function when it is not truly present. The detrimental impacts of presuming CF-CI when it does not hold is seen not only in the parameter estimates but in economic objects of interest such as the estimated choice probabilities and

price elasticities.

2.4.2 Simulation: General Control Function in the Demand for Premium Cable

The data generating process will emulate the setting described in Example 1 above, which is a simplification of the application given in PT. Recall that in this example I wish to estimate the demand for premium cable (conditional on already selecting cable as the television provider) but am concerned that price is endogenous and correlated with unobserved attributes. The latent utility function given in equation (2.13) is a function of product characteristics, such as the number of channels (z_{11m}), and individual characteristics of the consumers ($\mathbf{z}_{12i}$), including income, single family household indicator, rent indicator, age, and age squared. Building on the example of advertisement and marketing (part of the unobserved product attributes), I interact ξ_m with product characteristics (number of channels) and individual characteristics (age).

For simplicity, it is assumed that there is an exogenous cost shifter that acts as a valid instrument. As in PT, price will be interacted with 5 income level dummies to allow the price elasticity of premium cable to differ by income levels.

A discussion on the construction of the variables as well as a table of summary statistics is provided in Appendix C. As mentioned previously, it is important to note that the data generating process specifies a general control function that satisfies the CMR but does not satisfy CF-CI.

Table E.2 provides the parameter estimates for the different Logit specifications. As found in PT, without addressing any endogeneity (column (1)), there is actually a positive effect

of price for the higher income groups (in this simulation only the highest income group) and a negative effect for number of cable channels offered. Addressing endogeneity by including just the control variable (column (2)), as in PT, significantly strengthens the coefficient estimate on price and significantly alters to the coefficient estimate on number of channels to be positive. This is because price is strongly correlated with the number of channels offered and therefore addressing the endogeneity in price will also affect the estimated coefficient on number of channels. But allowing for the general control function in columns (3) and (4), the parameter estimates are much closer to their true value. For instance, the number of channels becomes much less impactful once the unobserved attributes, such as advertisement, of the premium channels is control for. In addition, the income effects are slightly higher than they are in column (2).

The difference between columns (3) and (4) is that column (3) the general control function is correctly specified by including interactions between the control variate and the relevant instruments (number of channels and age) while in column (4) the general control function includes more terms that are not actually relevant such as interactions between the control variate and household size and income. This is to explore the realistic situation that researchers would not typically have prior knowledge as to which terms to include in the general control function. The simulation results illustrate that there is very little loss to precision of the parameter estimates when one over specifies the general control function (column (4)). This shows there is very little cost to allowing the flexibility in estimation even when the true form may be more simplistic.

But these parameter estimates provide little interpretative value. Usually of more interest are the choice probabilities which in this binary context corresponds to the ASF (the derivation of the ASF will be discussed in more detail in Section 4). Figure D.2 illustrates

how the ASF varies over price for an additional 5 channels of premium cable, assuming the individual is 35 years old in a family of 3 with income equal to $85,000. Estimates from a linear probability model, OLS and 2SLS (in orange), are also included as a comparison. OLS and Logit (dotted lines) which do not address endogeneity result in upward sloping ASF while the remaining estimators more realistically provide downward sloping ASF. The correctly-specified Logit (GCF) and over-specified Logit (Over) both follow the true ASF quite closely. Although Logit (CV) performs better compared to Logit or the linear specifications, there is still some cost in not allowing for a flexible control function.

The price elasticity of demand for premium cable is calculated as,

$$Elasticity = E\left(\frac{\partial E(y_{1i}|\mathbf{z}_{1i}, p)}{\partial p} \times \frac{p}{E(y_{1i}|\mathbf{z}_{1i}, p)}\right) \tag{2.19}$$

The linear probability models estimated by OLS or 2SLS produces a conditional mean, $E(y_{1i}|\mathbf{z}_{1i}, p)$, not strictly positive nor bounded below 1, this will result in imprecise and extreme elasticity estimates. Table E.3 presents the estimated price elasticities for the different estimation procedures. OLS and 2SLS unsurprisingly provide poor estimates and Logit which does not address endogeneity greatly underestimates the price elasticity as inelastic. The Logit CV estimate is in a similar range to that produced in PT but the specifications that allow for more flexibility, Logit (GCF) and Logit (Over), are much closer to the true value. Again, as seen in the parameter estimates, there is very little cost in terms of efficiency, to including more terms in Logit (Over) when a simpler control function is the true specification.

Now that the general control function is shown to be consequential and the complications concerning identification have been addressed, consistency of the estimation procedure is insured along with standard regulatory conditions. The next section discusses the estimation

procedure in more detail and derives consistent estimates of the asymptotic variance. Since the parameters are usually of little interest in a latent variable model, the next section also discusses the formulation, identification, and estimation of the ASF and APE using the proposed estimation procedure.

2.5 Estimation and Interpretation

The estimation procedure proposed in this chapter for the parametric model is a standard two step estimator. In the first stage, the conditional mean function $E(y_{2i}|\mathbf{z}_i) = m(\mathbf{z}_i)\pi_o$ is estimated using standard LS regression techniques.[15] The control variable is constructed from the reduced form residuals, $\hat{v}_{2i} = y_{2i} - m(\mathbf{z}_i)\hat{\pi}$, and used in the second step. In the second stage, one would maximize the following likelihood

$$\begin{aligned}\mathcal{L}(y_{1i}, \mathbf{x}_i, \mathbf{z}_i; \hat{\pi}, \beta, \gamma, \delta) = \sum_{i=1}^{n} y_{1i} \log \left[\Phi \left(\frac{\mathbf{x}_i\beta + h(\hat{v}_{2i}, \mathbf{z}_i)\gamma}{\exp(g(y_{2i}, \mathbf{z}_i)\delta)} \right) \right] \\ + (1 - y_{1i}) \log \left[1 - \Phi \left(\frac{\mathbf{x}_i\beta + h(\hat{v}_{2i}, \mathbf{z}_i)\gamma}{\exp(g(y_{2i}, \mathbf{z}_i)\delta)} \right) \right]\end{aligned} \tag{2.20}$$

with respect to β, γ and δ to obtain estimates of the parameters. In addition to relaxing assumptions in the literature, the proposed estimation procedure is quite simple to implement using commands from standard statistical packages.[16] However, the estimated standard errors need to be adjusted to account for the variation from using the residual from the first stage as an approximation for the control variate. Asymptotic variance formulas that account for the multi-step approach are given in the next section, although a common alternative

[15] Alternatively, one may consider a non-parametric first stage regression to obtain estimates of a conditional mean function. Using sieve, asymptotic results would follow directly from Newey (1994) which differs from the asymptotic theory presented in this chapter. However, Ackerberg, Chen, and Hahn (2012) explain that the asymptotic variance estimator under the framework of the semi-parametric plug-in two step estimator is numerically equivalent to the asymptotic variance estimator in the parametric framework as long as the parametric specification is flexible enough.

[16] For example, the parameter estimates can be obtained using `reg` and `hetprobit` commands in Stata.

would be to bootstrap the standard errors.

As for consistency and asymptotic normality, this is a simple application of MLE to a heteroskedastic Probit model with a generated regressor in which asymptotics are well-established. The next subsection provides the asymptotic properties and the asymptotic variance derivation under a two step M-estimation framework.

2.5.1 Asymptotic Properties

The two-step estimator can be written in a GMM framework by stacking the moment conditions,

$$E\big(M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \beta_o, \gamma_o, \delta_o)\big) = E\begin{pmatrix} (y_{2i} - m(\mathbf{z}_i)\pi_o)m(\mathbf{z}_i)' \\ S_i(\pi_o, \beta_o, \gamma_o, \delta_o) \end{pmatrix} = 0 \tag{2.21}$$

where $S_i(\pi, \beta, \gamma, \delta) = \partial\mathcal{L}(y_{1i}, \mathbf{x}_i, \mathbf{z}_i; \pi, \beta, \gamma, \delta)/\partial\theta$ denotes the score,

$$\begin{aligned} S_i(\pi, \theta) =& \frac{(y_{1i} - \Phi_i(\pi, \theta))\phi_i(\pi, \theta)}{\Phi_i(\pi, \theta)(1 - \Phi_i(\pi, \theta))\exp(g(y_{2i} - m(\mathbf{z}_i)\pi, \mathbf{z}_i)\delta)} \\ & \times \begin{pmatrix} \mathbf{x}_i' \\ h(y_{2i} - m(\mathbf{z}_i)\pi, \mathbf{z}_i)' \\ -(\mathbf{x}_i\beta + h(y_{2i} - m(\mathbf{z}_i)\pi, \mathbf{z}_i))g(y_{2i} - m(\mathbf{z}_i)\pi, \mathbf{z}_i)' \end{pmatrix} \end{aligned}$$

$\Phi_i(\cdot)$ and $\phi_i(\cdot)$ are shorthand for the conditional CDF and PDF evaluated at the linear index $x_i\beta$. Note that estimation using the stacked moment conditions is equivalent to the two step approach previously described. Although using the GMM framework is useful for deriving the asymptotic variance of the estimator, it is suggested to use the two step approach in implementation to avoid issues of slow convergence.

Let $\theta' = (\beta', \gamma', \delta')$ and let Π and Θ denote the parameter spaces of π and θ respectively. Consistency follows from Theorem 2.6 of Newey and McFadden (1994).

Theorem 2.5.1. *In the set-up described by equation (2.1) where assumptions 2.3.1 and 2.4.1 hold, if* $\pi_o \in \Pi$ *and* $\theta_o \in \Theta$*, both of which are compact, then the GMM estimators that solve:*

$$(\hat{\pi}, \hat{\theta}) = \underset{(\pi,\theta)\in\Pi\times\Theta}{\arg\min} \left[\frac{1}{n}\sum_{i=1}^{n} M(y_{1i}, y_{2i}, \boldsymbol{z}_i; \pi, \theta)\right]' \left[\frac{1}{n}\sum_{i=1}^{n} M(y_{1i}, y_{2i}, \boldsymbol{z}_i; \pi, \theta)\right] + o_p(1) \quad (2.22)$$

where $M(y_{1i}, y_{2i}, \boldsymbol{z}_i; \pi, \theta)$ *are the stacked moment conditions in equation (2.21), are consistent,* $\hat{\pi} - \pi_o = o_p(1)$ *and* $\hat{\theta} - \theta_o = o_p(1)$*.*

Proof is provided in the appendix. Showing asymptotic normality follows from Theorem 6.1 in Newey and McFadden (1994).

Theorem 2.5.2. *In the set-up described by equation (2.1) where assumptions 2.3.1 and 2.4.1 hold, if* $\pi_o \in \text{int}(\Pi)$ *and* $\theta_o \in \text{int}(\Theta)$*, both of which are compact, then for* $(\hat{\pi}, \hat{\theta})$ *that solves equation (2.22),* $\sqrt{n}(\hat{\theta} - \theta_o) \xrightarrow{d} N(0, V)$ *where*

$$V = G_{2\theta}^{-1} E\left(\Xi_i(\pi_o, \theta_o)\Xi_i(\pi_o, \theta_o)'\right) G_{2\theta}^{-1\prime} \quad (2.23)$$

where $\Xi_i(\pi_o, \theta_o) = S_i(\pi_o, \theta_o) + G_{2\pi}G_{1\pi}^{-1}(y_{2i} - m(\boldsymbol{z}_i)\pi_o)m(\boldsymbol{z}_i)'$ *and*
$\begin{pmatrix} G_{1\pi} & G_{1\theta} \\ G_{2\pi} & G_{2\theta} \end{pmatrix} = E(\bigtriangledown_{(\pi,\theta)} M(y_{1i}, y_{2i}, \boldsymbol{z}_i; \pi_o, \theta_o))$ *is defined in detail in the appendix.*

The proof is provided in the appendix. Note that the asymptotic variance takes into account the variation introduced from the first stage. A consistent estimator for the asymptotic variance would be the method of moments estimator that replaces all the unknown parameters with their consistent estimates and then use sample averages in place of expectations. Although this section provides consistency and $\sqrt{n}$-asymptotic normality for the second stage parameter estimates, the parameters themselves bear very little interpretative value. The next two subsections discuss the derivation of the ASF and the APE and their importance for economic interpretation. These structural objects are magnitudes of effects that

empirical researchers can use to discuss the effectiveness of a particular policy or the average probability of a successful outcome for an individual with a particular set of characteristics.

2.5.2 Average Structural Function

Researchers are often interested in using the data and model estimates to infer the average predicted probability of success at particular point of the observed data. When there is no endogeneity, this quantity can be easily described by the conditional mean, which in the case of binary response, is equivalent to the propensity score. As explained in BP, when endogeneity is present, the conditional mean is unable to capture the structural relationship between the endogenous variable and the outcome. In particular, most studies wish to uncover the effect of a structural intervention over the endogenous variable on the outcome, while the conditional mean can only capture a reduced form effect over changes in the instruments.

For clarification, let us consider a simple linear structural equation.

$$y_i = \mathbf{x}_i\beta_o + u_i \tag{2.24}$$

Without endogeneity, $E(u_i|x_i) = 0$ and the interpretation of the average outcome for a given observation $\mathbf{x}^o$ is simply the conditional mean: $\mathbf{x}^o\beta_o$. The corresponding partial effect would be the slope parameter β_o. But when endogeneity is introduced, $E(u_i|\mathbf{x}_i) \neq 0$, the conditional mean is composed of two parts.

$$E(y_i|\mathbf{x}_i = \mathbf{x}^o) = \mathbf{x}^o\beta + E(u_i|\mathbf{x}_i = \mathbf{x}^o) \tag{2.25}$$

The first component is the structural direct effect of $\mathbf{x}_i$ while the second component is the endogenous indirect effect of $\mathbf{x}_i$ due to the presence of endogeneity. For instance, consider

the ubiquitous example of returns to education where education is endogenous due to unobserved ability. Then the structural direct effect is the average wage for particular education level (independent of ability) and the endogenous indirect effect is the contribution of average ability for that given education level on wages. But BP argues that one should only be interested in the structural direct effect because if one were to consider a policy intervention on the level of education (ie: mandatory schooling) there would be no changes in the distribution of ability and therefore one would only want to capture the structural direct effect.

To derive the ASF, BP instruct that one should integrate over the unconditional distribution of the unobserved heterogeneity in the structural equation. If the structural equation (2.24) includes an intercept then $E(u_i) = 0$ and the ASF is $\mathbf{x}^o\beta_o$, not equal to the conditional mean but still the same as the case of no endogeneity.

Next is to extend the analysis to the binary response model.

$$y_i = 1\{\mathbf{x}_i\beta_o + u_i > 0\} \tag{2.26}$$

When there is independence between the latent error u_i and the regressors $\mathbf{x}_i$, the conditional mean – equivalent to the propensity score – is

$$E(y_i|\mathbf{x}_i = \mathbf{x}^o) = F_{-u}(\mathbf{x}^o\beta_o) \tag{2.27}$$

which calculates the probability of success for an individual with characteristics $\mathbf{x}^o$. Now consider the case when there is no longer independence between the latent error and the regressors so the unconditional CDF is not equal to the conditional CDF; i.e., $F_{-u}(-u) \neq F_{-u|\mathbf{x}}(-u;\mathbf{x})$ where $F_{-u|\mathbf{x}}(\cdot;\cdot)$ is the conditional CDF in which the first argument is the point of evaluation and the second argument is the conditioning argument. One can understand

the violation of independence either through the standard interpretation of endogeneity, $E(u_i|\mathbf{x}_i) \neq 0$, or possible due to endogeneity at higher moments such as heteroskedasticity, $Var(u_i|\mathbf{x}_i) \neq Var(u_i)$. Then the propensity score is

$$E(y_i|\mathbf{x}_i = \mathbf{x}^o) = F_{-u|\mathbf{x}}(\mathbf{x}^o\beta_o; \mathbf{x}^o) \tag{2.28}$$

in which the first argument in $F_{-u|\mathbf{x}}(\mathbf{x}^o\beta_o; \mathbf{x}^o)$ is the point of evaluation which, corresponds to the structural direct effect, and the second argument is the conditioning argument, which corresponds to the endogenous indirect effect. As in the linear case, the conditional mean does not capture a structural interpretation. Therefore, to obtain the ASF, one can integrate over the unconditional distribution of the unobserved heterogeneity to obtain: $F_{-u}(\mathbf{x}^o\beta_o)$. Now the ASF only captures the structural direct effect of $\mathbf{x}_i$ and is not clouded by the influence of endogeneity. However, in calculating the ASF, the unconditional distribution of u_i is usually unknown or at least not specified when estimating the structural parameters β_o.

Wooldridge (2005) studies the ASF in more depth and provides a more rigorous investigation of the derivation of the ASF. Using the same notation as above, the structural model of interest is $E(y_i|\mathbf{x}_i, u_i) = \mu_1(\mathbf{x}_i, u_i)$, where $\mathbf{x}_i$ is observed covariates and u_i is unobserved heterogeneity. Then the ASF is defined as $ASF(\mathbf{x}^o) = E_u(\mu_1(\mathbf{x}^o, u_i))$ where the subscript of u is meant to emphasize that the expectation is taken with respect to the unconditional distribution of u_i. Using Lemma 2.1 from Wooldridge (2005), which is essentially an application of law of iterated expectations, the ASF can also be calculated from

$$\begin{aligned} ASF(\mathbf{x}^o) &= E_w(\mu_2(\mathbf{x}^o, w_i)) \\ \mu_2(\mathbf{x}^o, w_i) &= \int_{\mathcal{U}} \mu_1(\mathbf{x}^o, u) f_{u|w}(u; w_i)\eta(\mathrm{d}u) \end{aligned}$$

where $\mathcal{U}$ is the support of u_i and $f_{u|w}(\cdot;\cdot)$ is the conditional density of the unobserved heterogeneity u_i given w_i with respect to a σ-finite measure $\eta(\cdot)$. Essentially, one can use a conditioning argument w_i to help identify the ASF. In many instances the conditioning argument w_i will include components of the covariates $\mathbf{x}_i$, but it is important to note that the evaluation of the ASF requires the ability to distinguish between the point of evaluation $\mathbf{x}^o$ and the conditioning argument w_i.[17] This will be important when I discuss the implications of the CF-LI assumption on the derivation and estimation of the ASF.

To apply Lemma 2.1 from Wooldridge (2005), the following conditions must hold

(i) *(Ignorability)* $E(y_i|\mathbf{x}_i, u_i, w_i) = E(y_i|\mathbf{x}_i, u_i)$

(ii) *(Conditional Independence)* $D(u_i|\mathbf{x}_i, w_i) = E(u_i|w_i)$

Notice that conditional independence in this context is with respect to the conditioning argument, w_i which has yet to be specified in our context. When the conditioning argument is simply the control variate, v_{2i}, then it is in fact the same conditional independence assumption of BP. Therefore the CF-CI assumption of BP is also used to obtain identification of the ASF. But, as seen when showing identification of the parameters, is CF-CI really necessary?

Consider the conditioning argument as both the control variate v_{2i} and the instruments $\mathbf{z}_i$. This easily satisfies the ignorability assumption $E(y_{1i}|\mathbf{x}_i, u_{1i}, v_{2i}, z_i) = E(y_{1i}|\mathbf{x}_i, u_{1i})$

[17] Wooldridge (2005) considers the example of the heteroskedastic Probit model where in equation (2.26), it is assumed u_i is normally distributed with $Var(u_i|\mathbf{x}_i) = \exp(2\mathbf{x}_i\delta)$. Then the covariates $\mathbf{x}_i$ are used as the conditioning argument (ie: $w_i = \mathbf{x}_i$) such that

$$\begin{aligned} ASF(\mathbf{x}^o) &= E_{\mathbf{x}_i}\left(\int_{\Re} 1\{\mathbf{x}^o\beta_o + u > 0\} f_{u|\mathbf{x}}(u;\mathbf{x}_i)\mathrm{d}u\right) \\ &= E_{\mathbf{x}_i}\left(\Phi\left(\frac{\mathbf{x}^o\beta}{\exp(\mathbf{x}_i\delta)}\right)\right) \end{aligned}$$

where the expectation is taken with respect to the x_i in the heteroskedastic function (part of the conditioning argument) and not with respect to the structural direct effect of x^o. Therefore, even when the conditioning argument is the same as the covariates in the structural equation, it is necessary to be able to distinguish between the two when composing the ASF.

given ignorability of the excluded instruments $\mathbf{z}_{2i}$ and automatically satisfies this version of conditional independence $D(u_{1i}|\mathbf{x}_i, v_{2i}, \mathbf{z}_i) = D(u_{1i}|v_{2i}, \mathbf{z}_i)$ since $\mathbf{x}_i$ is composed of v_{2i} and $\mathbf{z}_i$. Then under Assumption 2.3.1,

$$\begin{aligned}\mu_2(x^o, (v_{2i}, \mathbf{z}_i)) &= \int_{\Re} 1\{\mathbf{x}^o\beta_o + u > 0\} f_{u|v,\mathbf{z}}(u; v_{2i}, \mathbf{z}_i)\mathrm{d}u \\ &= E_u(1\{\mathbf{x}^o\beta_o + u > 0\}|v_{2i}, \mathbf{z}_i) \\ &= \Phi\left(\frac{\mathbf{x}^o\beta_o + h(v_{2i}, \mathbf{z}_i)\gamma_o}{\exp(g(y_{2i}, \mathbf{z}_i)\delta_o)}\right)\end{aligned}$$

and the ASF is

$$ASF(\mathbf{x}^o) = E_{v_2,z}(\mu_2(\mathbf{x}^o, (v_{2i}, \mathbf{z}_i))) = E_{v_2,\mathbf{z}}\left(\Phi\left(\frac{\mathbf{x}^o\beta_o + h(v_{2i}, \mathbf{z}_i)\gamma_o}{\exp(g(y_{2i}, \mathbf{z}_i)\delta_o)}\right)\right) \tag{2.29}$$

where the expectation is taken with respect to the unconditional distribution of v_{2i} and $\mathbf{z}_i$. A consistent method of moments estimator would replace the unknown parameter values with their consistent estimates, $(\hat{\pi}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$, and in place of the expectation, take sample averages. Therefore identification of the ASF is still possible without the CF-CI assumption of BP.

Next is to examine the derivation of the ASF in the BP framework where CF-CI is assumed. Let $G_{CF-CI}(\cdot; v_{2i})$ be the CDF of $-u_{1i}|v_{2i}, \mathbf{z}_i$ that will be estimated non-parametrically and recall that CF-CI implies that $\mathbf{z}_i$ is excluded from the conditional distribution function. Then the ASF is easily calculated as,

$$ASF_{CF-CI}(\mathbf{x}^o) = E_{v_2}(G_{-u_1|v_2}(\mathbf{x}^o\beta_o; v_{2i})) \tag{2.30}$$

where the expectation is taken with respect to v_{2i}. Comparing equations (2.29) and (2.30), highlights the impact of CF-CI on interpretation. Since there is endogeneity, the effect of $\mathbf{x}^o$ on the predicted probability of success can be broken down between the structural direct effect and an endogenous indirect effect. The allure of the CF-CI assumption is it

immediately distinguishes between the two effects in the conditional distribution function $G_{u_1|v_2}(\mathbf{x}^o\beta_o; v_{2i})$ where the first argument captures the structural direct effect and the second argument should entirely control for endogenous indirect effect. But when CF-CI fails, and this structure of the conditional CDF is still presumed, the lines between the structural direct effect and an endogenous indirect effect become blurred. Consequently, in estimation, the ASF calculated when incorrectly imposing CF-CI will not be able to correctly average out the endogenous indirect effect.

In a more flexible framework, Rothe assumes the CF-LI which slightly relaxes the CF-CI by allowing the conditional distribution to be a function of the instruments through the linear index $\mathbf{x}_i\beta_o$. Recall, that CF-LI means $D(u_{1i}|v_{2i}, \mathbf{z}_i) = D(u_{1i}|v_{2i}, \mathbf{x}_i\beta_o)$. Using results from Manski (1988), identification of β_o and $G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i}) = F_{u_1|v_2,\mathbf{x}\beta_o}(\mathbf{x}_i\beta_o; v_{2i}, \mathbf{x}_i\beta_o)$ which is the conditional CDF of u_{1i} evaluated at $\mathbf{x}_i\beta_o$ can be obtained. As mentioned before, the CF-LI assumption is still a fairly strong restriction on the conditional distribution of $u_i|v_{2i}, \mathbf{z}_i$. Compared to the specification in Assumption 2.3.1, this would require the control function and the heteroskedastic function to be constructed with the linear index and not as more flexible functions of the instruments. But for now, consider the most optimistic case where the CF-LI assumption holds, then how does one calculate and estimate the ASF?

Again applying the framework provided in Wooldridge (2005), where now the conditioning argument includes the control variate and the linear index, $w_i = (v_{2i}, \mathbf{x}_i\beta_o)$.

$$\begin{aligned}
\mu_2(\mathbf{x}^o, w_i) &= \int_{\Re} 1\{\mathbf{x}^o\beta_o + u > 0\} f_{u|v,\mathbf{x}\beta}(u; v_{2i}, \mathbf{x}_i\beta_o)\mathrm{d}u \\
&= E_u(1\{\mathbf{x}^o\beta_o + u > 0\}|v_{2i}, \mathbf{x}_i\beta_o) \\
&= F_{u_1|v_2,\mathbf{x}\beta_o}(\mathbf{x}^o\beta_o; v_{2i}, \mathbf{x}_i\beta_o)
\end{aligned}$$

notice that the linear index appears twice as arguments: first at the point of evaluation for the

conditional CDF $\mathbf{x}^o\beta_o$ (the structural direct effect) and as part of the conditioning argument $\mathbf{x}_i\beta_o$ (the endogenous indirect effect). Applying Lemma 2.1 from Wooldridge (2005), the ASF when the true data generating process satisfies the CF-LI assumption is

$$ASF_{CF-LI}(\mathbf{x}^o) = E_{v_2,\mathbf{x}\beta_o}(F_{-u_1|v_2,\mathbf{x}\beta_o}(\mathbf{x}^o\beta_o; v_{2i}, \mathbf{x}_i\beta_o)) \tag{2.31}$$

where the expectation is taken with respect to the joint distribution of the conditioning arguments $(v_{2i}, \mathbf{x}_i\beta)$. The immediate issue is that the ASF cannot be written in terms of the identified function $G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i})$ that is estimated using the proposed SML estimator in Rothe. The identified function is the conditional CDF evaluated at and conditioned on the same linear index. Therefore one cannot distinguish between the direct structural effect and indirect endogenous effect of the linear index. This reiterates the importance of being able to separately identify the conditioning argument from the point of evaluation for estimation.

Rothe suggests using $E_{v_2}(G_{CF-LI}(\mathbf{x}^o\beta_o, v_{2i}))$ as the ASF but this only averages out the part of the endogenous indirect component due to v_{2i}, and does not average out any the effect due to the linear index. Therefore, the ASF proposed by Rothe is equal to the true ASF only when CF-CI assumption of BP holds. So although it may be tempting to consider the CF-LI assumption as a compromise to allow for flexibility in terms of the relationship between the unobserved heterogeneity and the instruments, the true ASF is not identified under the CF-LI assumption.

In fact, the ASF proposed by Rothe is estimating the AIF of Lewbel, Dong, and Yang (2012) who suggest using it as an alternative to the ASF since it is generally much easier to identify. They define the AIF as

$$\begin{aligned} AIF(\mathbf{x}^o) &= E(1\{\mathbf{x}_i\beta_o + u_{i1} > 0\}|\mathbf{x}_i\beta_o = \mathbf{x}^o\beta_o) \\ &= E_{v_2}(G_{CF-LI}(\mathbf{x}^o\beta_o, v_{2i})) \end{aligned} \tag{2.32}$$

Given the choices of Propensity Score, ASF, and AIF, as possible ways to interpret the estimates of the model, how should one proceed? Lin and Wooldridge (2015) address this issue by comparing these functions, proposed in the context of binary response, in linear regression case (as in equation (2.24)) to see if they uncover the direct structural component, $\mathbf{x}^o\beta_o$. Earlier in this section, the propensity score is shown to not be reflective of the mechanisms researchers are interested in when endogeneity is present. Lin and Wooldridge (2015) also show this for the AIF, explaining that "the AIF suffers from essentially the same shortcomings as the propensity score because it is affected by correlation between the unobservables and the observed [endogenous explanatory variables]."

The next section discusses the derivation of the APEs. Again, the APEs should isolate the structural impact of varying a particular covariate and therefore should be derived from the ASF. Consequently presuming CF-CI or CF-LI affects the derivations and interpretations of the APEs.

2.5.3 Average Partial Effects

Similar to the interpretation of β_o in a linear regression (as in equation (2.24)), the APE should capture the causal (structural direct) effect of a regressors on the outcome variable. In the binary response framework, the parameters are scale invariant and therefore provide very little for interpretation and are generally not comparable across different specifications. Alternatively, the APE provide a comparable statistic that can be used for interpretation. Let x_j^o and β_{jo} denote the j^{th} elements of $\mathbf{x}^o$ and β_o respectively. Then the partial effect of the j^{th} element of $\mathbf{x}_i$ is defined as the partial derivative of the ASF with respect to x_j^o averaged in the population. Under the setting consider in Assumption 2.3.1, the partial

effect is,

$$\begin{aligned}\partial ASF(\mathbf{x}^o)\Big/\partial x_j^o &= \partial E_{v_2,\mathbf{z}}\left(\Phi\left(\frac{\mathbf{x}^o\beta_o + h(v_{2i},\mathbf{z}_i)\gamma_o}{\exp(g(y_{2i},\mathbf{z}_i)\delta_o)}\right)\right)\Big/\partial x_j^o \\ &= E_{v_2,\mathbf{z}}\left(\partial\Phi\left(\frac{\mathbf{x}^o\beta_o + h(v_{2i},\mathbf{z}_i)\gamma_o}{\exp(g(y_{2i},\mathbf{z}_i)\delta_o)}\right)\Big/\partial x_j^o\right) \\ &= E_{v_2,\mathbf{z}}\left(\phi\left(\frac{\mathbf{x}^o\beta_o + h(v_{2i},\mathbf{z}_i)\gamma_o}{\exp(g(y_{2i},\mathbf{z}_i)\delta_o)}\right)\frac{\beta_{jo}}{\exp(g(y_{2i},\mathbf{z}_i)\delta_o)}\right)\end{aligned}$$

To obtain the APE, one plugs in $\mathbf{x}_i$ for $\mathbf{x}^o$ and averages over the joint distribution of $(\mathbf{x}_i, v_{2i}, \mathbf{z}_i)$,

$$APE = E\left(\phi\left(\frac{\mathbf{x}_i\beta_o + h(v_{2i},\mathbf{z}_i)\gamma_o}{\exp(g(y_{2i},\mathbf{z}_i)\delta_o)}\right)\frac{\beta_{jo}}{\exp(g(y_{2i},\mathbf{z}_i)\delta_o)}\right) \tag{2.33}$$

Use sample averages in place of expectations and consistent estimates of the parameters to estimate. Notice that the derivative is only taken with respect to the structural direct component, the argument in the ASF, but after the derivative is taken, one will average over the joint distribution of $\mathbf{x}_i$, v_{2i}, and $\mathbf{z}_i$ in both the structural direct effect and the endogenous indirect effect together.

How does using the AIF instead of the ASF affect the APE derivation under the CF-LI assumption? First, the correct APE under the CF-LI assumption

$$APE_{CF-LI} = \partial ASF(\mathbf{x}_i)\Big/\partial x_{ji} = E\left(f_{-u|v_2,\mathbf{x}\beta_o}(\mathbf{x}_i\beta_o; v_{2i}, \mathbf{x}_i\beta_o)\beta_{jo}\right) \tag{2.34}$$

where $f_{-u|v_2,\mathbf{x}\beta_o}(\cdot; v_{2i}, \mathbf{x}_i\beta_o)$ is the conditional PDF. Since I am averaging over the point of evaluation and the conditioning argument, one may be hastily optimistic in thinking this is identified from the conditional CDF, $G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i}) = F_{u_1|v_2,x\beta_o}(\mathbf{x}_i\beta_o; v_{2i}, \mathbf{x}_i\beta_o)$. However, the correct PDF cannot be derived the from this function since

$$\partial G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i})/\partial[\mathbf{x}_i\beta_o] \neq f_{-u|v_2,\mathbf{x}\beta_o}(x_i\beta_o; v_{2i}, \mathbf{x}_i\beta_o)$$

Consequently, the APE in Rothe are also incorrectly calculated from the AIF,

$$APE_{AIF} = E\left((\partial G_{CF-LI}(\mathbf{x}_i\beta_o, v_{2i})/\partial[\mathbf{x}_i\beta_o])\beta_{jo}\right) \tag{2.35}$$

This discussion has provided a theoretical argument for the differences between the AIF and ASF and why one should prefer the ASF. But one may wonder whether all of this matters in practice. Once you start averaging over components, minute differences in calculations may be diminished in their impact. Perhaps the AIF used by Rothe may do a "good enough" job in approximating the true ASF. This is investigated in a simulation study in the next section where the CF-LI assumption holds true in the underlying data generating process and the ASF from the proposed estimator and AIF from the SML estimator proposed by Rothe are calculated and compared.[18] The simulation results suggest that when the CF-LI assumption holds, both the proposed method and the SML estimator from Rothe perform quite well in terms of parameter estimates. However, there is a stark difference in ASF estimates, consistent with the previous analysis. In the simulation, the poor ASF estimates using the SML procedure can be entirely attributed to the fact that under CF-LI, the SML can only recover the AIF which can be starkly different from the ASF.

2.5.4 Simulation: ASF Estimates for the Effect of Income on Home-ownership

This simulation models the home-ownership and income application in Rothe (2009) as a contextual setting. Rothe uses a sample of '981 married men aged 30 to 50 that are working full time and have completed at most the lowest secondary school track of the German

[18]Rothe (2009) also considers the case that CF-LI assumption holds but CF-CI does not hold in the simulation study. The second design introduces heteroskedasticity as a function of the linear index $\mathbf{x}_i\beta_o$, in the unobserved latent error, u_{1i}. However, only results on coefficient estimates, and not the ASF or APE estimates, are reported.

education system' from a 2004 wave of the German Socio-economic Panel. The outcome y_{1i} is an indicator that takes on the value 1 if an individual owns their home and 0 if they are renting. The included instruments $\mathbf{z}_{1i}$ are individual's age (z_{11i}) and an indicator of the presence of children younger than 16 (z_{12i}). The endogenous variable of interest, y_{2i} is household income and there are two excluded instruments: indicators for the wife's education level (intermediate z_{21i} and advanced z_{22i}) and an indicator for her employment status (z_{23i}). A more detailed discussion of the data generating process and table of summary statistics are presented in the Appendix.

In this simulation the CF-LI assumption holds in the underlying data generating process such that the distribution of u_{1i} conditional on v_{2i} and the exogenous regressors ($\mathbf{z}_i$) is,

$$u_{1i}|\mathbf{z}_i, v_{2i} \sim N\Big(y_{2i}\gamma_{1o} + v_{2i}(\mathbf{x}_i\beta_o)\gamma_{1o}, \exp(2 \times (\mathbf{x}_i\beta_o)\delta_o)\Big)$$

Motivated by the explanation in example 2, by allowing for an interactive effect in the conditional mean function, I am allowing for interactions between the omitted variable, credit score, and the linear index. Heteroskedasticity is also introduced which allows for variability in the variance of unobservables conditional on observables. Since $u_{1i}|v_{2i}, \mathbf{z}_i \sim u_{1i}|v_{2i}, \mathbf{x}_i\beta$, the CF-CI assumption is violated but the CF-LI assumption holds. This simulation will examine in more detail the CF-LI assumption as a relaxation of the CF-CI assumption, investigating whether the discussion on the ASF, AIF, and APE holds true in practice. Given the analysis of the previous section, the Rothe SML estimator should be able to estimate the parameters β_o well but unable to correctly calculate the ASF and APE because it cannot distinguish between the two effects (structural direct and endogenous indirect) of the linear index, $x_i\beta_o$. Implementation of the SML and the proposed estimator are explained in more detail in the appendix.

Table E.5 reports the coefficient estimates for the simulated data as well as the estimates in the Rothe application as a comparison.[19] All the second stage coefficient estimates are normalized such that the coefficient on Children in the Household is one. The simulated data is not an exact replica, but the estimates are in the same range and the change in the coefficient estimates as one starts to control for the endogeneity all move in the same direction. As expected, the estimates for Het-Probit (GCF) – the proposed estimator – and SML, columns (8) and (9), are quite similar and close to the true values in column (10).

Figures D.3 and D.4 show the ASF estimates for a 40 year old with children under the age of 16 in the household as it varies over the endogenous regressor, log(total income). In Figure D.3, the OLS and Probit estimators perform poorly since they do not address endogeneity at all. The 2SLS and Probit (CV) estimates are much closer to the true ASF but predict a slightly flatter ASF.

Recall from the earlier discussion, even if the SML is producing consistent parameter estimates, one would incorrectly estimate the ASF and consequently the APE. This is because the CF-LI assumption does not correctly average out the distribution of the unobserved heterogeneity. Figure D.4 reports the true ASF, the AIF (not a structural object of interest) and the estimated ASFs for the proposed Het-Probit (GCF) and the semi-parametric SML. The true ASF correctly averages over the distribution of the unobserved heterogeneity while the wrong ASF only averages out the v_{2i} components of the unobserved heterogeneity.

As expected, the proposed estimator does a good job estimating the true ASF while the SML estimator does a good job estimating the AIF. In this simulation I find that there can be a fairly stark difference between the ASF and AIF which means the differences in the estimators are consequential. For instance, the AIF would predict the average probability for

[19] SML and Het Probit are estimated using a Nelder-Mead Simplex Method in Matlab.

home-ownership for an individual with a log total income of 7.65 to be 0.595 while the ASF would predict an average probability of 0.461, a substantial difference. This further reiterates the discussion in Section 4; i.e., even under the CF-LI assumption, the SML estimator is not capturing the true ASF.

The simulated distribution of the APE are reported in Table E.6. The true APE is 0.6448 so the proposed Het-Probit CF estimator has the closest mean whereas the mean of the SML estimates is the third closest following the mean of the 2SLS estimates. But the difference between these estimators in interpretation is minimal: a 10% increase in total income results in either a 0.0626 increase in probability of home ownership according to the Het-Probit (GCF) or a 0.0699 increase according to Rothe's SML. The estimators that suffer the most are the ones that do not address the issue of endogeneity at all, OLS and Probit, and are distinctly biased downwards.

Therefore I find in this simulation study that under the CF-LI assumption parameter estimates are similar across the two estimators. But when looking at the ASF, the estimates diverge significantly. I show that this can be entirely accounted by the fact that the SML estimator is actually estimating the AIF which is not equal to the ASF (and should not be interpreted in the same way).

2.6 Empirical Example

To showcase the estimator in an empirical example, I examine married women's labor force participation using 1991 CPS data.[20] All tables and figures referenced in this section can

[20]Data is part of the supplementary material provided with the textbook "Econometric Analysis of Cross Section and Panel Data" by Jeffrey Wooldridge. Data can be downloaded at https://mitpress.mit.edu/books/econometric-analysis-cross-section-and-panel-data

be found in Appendix D. Table E.7 provides some summary statistics for the data set. The dependent variable is Employed (=1 when the individual is in the labor force) where approximately 58% of married women in the sample participate in the labor force. The last two column divide the sample over the binary outcome and reports the summary statistics for the other observable characteristics.

The structural outcome equation is,

$$\begin{aligned} Employed_i = 1\{&\beta_1 + nwifinc_i\beta_2 + educ_i\beta_3 + exper_i\beta_4 + kidslt6\beta_5 + kidsge6\beta_6 \\ &+ nwifinc_i \times kidslt6\beta_7 + nwifinc_i \times kidsge6\beta_8 + u_{1i} > 0\} \end{aligned} \tag{2.36}$$

where the economic interest is in estimating the effect of non-wife income on the probability of being in the labor force. Since there is a trade-off between work and leisure, by relaxing the budget constraint such that an individual has other sources of income, one would expect the individual to be less likely to work. From the summary statistics, those not working tend to have higher non-wife income. But this can not be interpreted as a causal effect since there is concern that other sources of income would be endogenously determined with the wife's labor force participation. In particular, husband's employment, which partly determines the non-wife income, would probably be decided simultaneously with wife's employment.

Utilizing husbands education level as an instrument, the causal effect of non-wife income on wife's labor force participation can be parse out. Since education and the probability of working are generally correlated, the instrument is easily argued to be relevant. In fact, the F-statistic of significance for the first stage is quite large as seen in Table E.8.[21] Excludability of the instrument follows from the argument that husband's education level should not directly effect the wife's choice of labor force participation except through the channels of how it

[21] The standard benchmark of 10 from Stock, Wright, and Yogo (2002) only applies to the relative bias in 2SLS with homoskedasticity. It is an open area of research to determine benchmarks for non-standard cases.

effects the non-wife income. The other controls considered in this example are the wife's education level, experience, and dummy variables for whether or not they have kids younger than 6 and kids 6 and older.

Table E.8 reports the reduced form coefficient estimates, the second stage parameter estimates for several different specifications of a Probit model, and the SML estimator proposed by Rothe. For the second stage estimates the coefficient on Education is normalized 1, since the model is only identified to scale. This allows for comparisons across the different estimators. The second column specifies a standard Probit model which assume no endogeneity and homoskedasticity in the latent error. This is slightly relaxed in column (3) where heteroskedasticity is allowed. The specifications in columns (4)-(8) all address endogeneity in one form or another. The fourth column corresponds to the setting of Rivers and Vuong (1988) where they address endogeneity by only including the control variable as an additional covariate, maintaining the CF-CI assumption. The next three columns are variations on the proposed estimator all of which relax the CF-CI assumption by either allowing for heteroskedasticity in the latent error and/or allowing for a general control function. The final column presents results using the SML estimator (from Rothe) which impose no distributional assumptions but require either CF-CI or CF-LI assumptions.

I find that addressing endogeneity with only a control variable reduces the effect of non-wife income (in columns (4) and (5)). But then allowing for a general control function, where the control variate interacts with the children dummies, raises the effect of non-wife income and also switches the signs for the interactions with the children (although not statistically significant). When a general control function is used without allowing for heteroskedasticity the bootstrapped standard errors increase substantially. This is due to very small (almost 0) coefficient estimates for education which blows up the scaled parameter estimates. When

heteroskedasticity is allowed, then the coefficient estimate on education becomes statistical different from 0 which results in lower standard errors of the scaled parameter estimates. Finally, the SML estimates are found to be quite similar to the proposed estimator results in column (6). This would suggest that the CF-LI assumption may in fact hold in this setting.

When looking at the control function parameters I see particularly large effects when interacting the reduced form error $\hat{v}_{2i}$ with the children dummies (in specification (7)). Intuitively this makes sense since one would imagine the endogenous decision making process of who in the household should work (either husband, wife or both) depends a lot on the presence of children in the household. For instance, if there are very young children in the household then the trade-off is not just between work and leisure but must also consider the cost of childcare if both parents enter the workforce. Therefore it would make sense that there is a negative interactive effect such that when one partner is working, the other is less likely to in order to provide childcare.

Since this chapter proposes a more flexible specification, Table E.9 provides Wald test results on different specifications. The first 4 columns test the null hypothesis that non-wife income is in fact exogenous. One of the benefits of the control variable approach is the variable addition test it supplies. One can test the null hypothesis of no endogeneity by testing whether all the coefficients in the control function are 0. Under all combination of modelling assumptions (such as homoskedasticity/heteroskedasticity and control variable/general control function) I find strong evidence of endogeneity. However this is conditional on the instrument being exogenous and, since the model is just identified, there is no way to test for exogeneity of the instrument.

The remainder of the table tests the different components of the CF-CI assumption in alternative specifications. The middle two columns test the null hypothesis that the control

variable is sufficient in capturing the full impact of the endogenous part of y_{2i}. In other words, testing the significance of the coefficients on the additional terms in the general control function. The null is rejected at the 10% level under homoskedasticity and rejected at the 5% level under heteroskedasticity. This gives statistical evidence of the violation of CF-CI, through the general control function, in this empirical applications. Finally the last three columns test the null hypothesis of homoskedasticity (i.e., all of the coefficients in heteroskedastic function are 0). There is strong statistical evidence of the violation of CF-CI through the presence of heteroskedasticity, easily rejecting homoskedasticity at the 5% level. Given these results, the preferred specification should be the Het-Probit (CF), I reject the possibility of homoskedasticity and reject the inclusion of only the control variable in favor of a general control function.

To understand the consequences of the different specification on the interpretation of the results, Table E.10 provides estimates of the APEs and their bootstrapped standard errors with respect to the endogenous variable, non-wife income. The most significant change in the estimates is when one starts to address the issue of endogeneity. In the linear models, the 2SLS APE estimates shrink to about 3/4 of the OLS estimates but are still statistically significant. In the non-linear models, a similar reduction in APE estimates is observed when controlling for endogeneity (about 3/5). But in the models that relax CF-CI completely (Het Probit (GCF)), the APE is no longer statistically significant and even switches its sign.

Putting the APE estimates into interpretive setting, if the non-wife income increases by \$10,000 – a fairly substantial increase–, according to the preferred specification, the likelihood of the wife working decreases by around 1.16 percentage points, a fairly negligible effect. As suggested in the discussion on coefficient estimates, this small effect is most likely driven by a heterogeneity over the presence of children in the household. Therefore examining the ASF

for different combination of ages of children present in the household can be informative. Since the APEs average over the distribution of all the covariates, these differing effects, tend to be washed out in the single statistic.

Figures D.5-D.8 show the ASF with respect to non-wife income for a married woman with high school education, 20 years of experience, and different combination of ages of children present in the household. The Probit models have a fairly linear ASF which explains why 2SLS gives fairly similar estimates of the APE.

The first thing to note is that the ASF using Probit estimates is much more negatively sloped in all the figures. This is because without addressing the issue of endogeneity, (i.e., if the husband works then the wife is less likely to work and visa versa, but these decisions are made simultaneously) I would expect to see this sort of substitution effect. Once endogeneity is controlled for in Probit (CV), the slope lessens. The much more striking revelation is with the proposed Het-Probit (GCF) and SML estimators, I see a positive slope when there are both children in the household and fairly flat ASF when there are only older children. When there are no children or just very young children in the household then the ASF is much more negatively sloped. This goes to show that relaxing CF-CI and allowing for a much more flexible specification in the conditional distribution of the latent error makes a interpretative impact.

In this setting, the normality assumption seems likely to hold, especially since the semi-parametric estimator (SML) produces fairly linear ASF estimates. But in other empirical applications, the normality assumption may be too restrictive and unconvincing. Consequently, neither the SML estimator or the proposed Het-Probit (GCF) estimator are strictly weaker in their assumptions. The SML estimator imposes CF-CI (or CF-LI) by not allowing for general heteroskedasticity and a flexible general control function. But on the other hand,

the proposed approach imposes distributional assumptions. This divergence between the proposed parametric estimator and the semi-parametric estimators offered in the literature leads to the following distribution free extension.

2.7 Extension: Semi-Parametric Distribution Free Estimator

In some empirical settings, imposing normality on the latent error may not be a reasonable assumption. Therefore this section offers an alternative semi-parametric estimator that does not depend on distributional assumptions. The main result of this section is that by allowing the heteroskedastic function and general control function to be non-parametrically specified, the semi-parametric variation of the proposed Het-Probit (GCF) is actually a distribution free estimator. This section will go into detail as to why this is true, how to obtain non-parametric identification and what the asymptotic properties of the semi-parametric estimator are. But for an applied researcher, the results of this section imply that as long as the heteroskedastic function and general control function are flexibly specified (i.e., sieve basis functions) the normality assumption that appears to be used in Assumption 2.3.1 is non-binding.

How is this distribution free estimator possible? A recent paper, Khan (2013), has noted an observational equivalence result concerning binary response models: a heteroskedastic Probit model with a non-parametric heteroskedastic function is observationally equivalent to a "distribution free" model with only a conditional median restriction. The utility of this result, is one may use simple estimation procedures (such as a semi-parametric Het-Probit MLE) while not making any strong distributional assumptions to obtain structural param-

eter estimates and possibly even identify and estimate choice probabilities and marginal effects. This section extends the result to the case of endogeneity in a flexible manner that allows for the relaxation of CF-CI. By introducing a general control function into the conditional median restriction, the observational equivalence holds under endogeneity and a simple estimation procedure is obtainable. Consequently, this section proposes a semi-parametric estimator based on assumptions that are more realistic than any other control function methods in the literature for endogenous binary response models.

Section 7.1 reviews the observational equivalence result in Khan (2013) and extends it to the case of endogeneity. Since this framework considers non-parametric functions for both the general control function and the heteroskedastic function, identification will be shown under this more general scenario. Section 7.2 derives the asymptotic properties of the semi-parametric estimator. Using the results in Song (2016), proofs of consistency and the rate of convergence only need to be slightly altered to allow for the semi-parametric general control function. Finally, this extension ends with a comprehensive simulation study. Over a variety of conditional distributions (some satisfying CF-CI and some not), the performance of the proposed semi-parametric Het-Probit (GCF) estimator is compared to the parametric Rivers and Vuong (1988) estimator and the SML of Rothe (2009). The simulation results suggest the proposed approach can handle a variety of alternative distributions while still allow for the violation of CF-CI.

2.7.1 Observational Equivalence and Identification

Consider the following binary response setting without endogeneity,

$$y_i = 1\{\mathbf{x}_i\beta_o + u_i \geq 0\} \tag{2.37}$$

where $\mathbf{x}_i$ is a vector of covariates, and u_i is the unobserved heterogeneity. The following two assumptions restates the setting of the two observationally equivalent models in Khan (2013).[22]

Assumption 2.7.1 (Conditional Median Restriction)**.** *In the set up described by equation (2.37)*

(i) $\boldsymbol{x}_i \in \Re^k$ *is assumed to have density with respect to a Lebesgue measure, which is positive on the set* $\mathcal{X} \subseteq \Re^k$.

(ii) *Let* $p_o(t, \boldsymbol{x}_i)$ *denote* $P(-u_i < t|\boldsymbol{x}_i)$*, and assume*

(a) $p_o(\cdot, \cdot)$ *is continuous on* $\Re \times \mathcal{X}$.

(b) $p_o'(t, \boldsymbol{x}_i) = \partial p_o(t, \boldsymbol{x}_i)/\partial t$ *exists and is continuous and positive on all* $\Re$ *for all* $\boldsymbol{x}_i \in \mathcal{X}$.

(c) $p_o(0, \boldsymbol{x}_i) = 1/2$.

(d) $\lim_{t\to-\infty} p_o(t, \boldsymbol{x}_i) = 0$, $\lim_{t\to\infty} p_o(t, \boldsymbol{x}_i) = 1$.

Assumption 2.7.2 (Heteroskedastic Probit)**.** *In the set up described by equation (2.37)*

(i) $\boldsymbol{x}_i \in \Re^k$ *is assumed to have density with respect to a Lebesgue measure, which is positive on the set* $\mathcal{X} \subseteq \Re^k$.

(ii) $u_i = \sigma_o(\boldsymbol{x}_i)e_i$ *where* $\sigma_o(\cdot)$ *is continuous and positive on* $\mathcal{X}$ *a.s, and* e_i *is independent of* $\boldsymbol{x}_i$ *with any known (e.g. logistic, normal) distribution with median 0 and has a density function which is positive and continuous on the real line.*

Theorem 2.1 of Khan (2013) states that under the above assumptions, the two models are observationally equivalent. The equivalence between the two models is in terms of the

[22] Assumption 2.1 correspond to CM1 and CM2 and Assumption 2.7.2 corresponds to HP1 and HP2 in Khan (2013)

choice probabilities: $P(y_{1i}|\mathbf{x}_i)$. In other words, both models will generate the same choice probability functions and therefore cannot be distinguished from one another on that basis. This means that a researcher is able to use estimators developed under Assumptions 2.7.2 such as a semi-parametric heteroskedastic Probit while only imposing the weaker distributional assumptions under Assumption 2.7.1. This allows for easy estimation using canned commands in popular statistical programs such as Stata, Matlab or R, but still preserving the "distribution free" interpretation.

Previously, endogeneity was understood as a non-zero conditional mean, but to fit the framework in Khan (2013), endogeneity will be determined by a non-zero conditional median: $\text{Med}(-u_i|\mathbf{x}_i) \neq 0$. This would violate assumptions Assumptions 2.7.1(ii) part (c) and 2.7.2(ii). Therefore the provided observational equivalence result from Khan (2013) is no longer applicable.

Now consider the set up in equation (2.1) and suppose I define the non-zero conditional median as,

$$h_o(v_{2i}, \mathbf{z}_i) \equiv \text{Med}(-u_{1i}|\mathbf{z}_i, v_{2i}) = \text{Med}(-u_{1i}|\mathbf{z}_{1i}, y_{2i}) \tag{2.38}$$

where the second equality holds because y_{2i} is merely a function of $\mathbf{z}_i$ and v_{2i}. This function should look familiar as it would be the non-parametric version of the general control function introduced in Assumption 2.3.1. This function captures the part of the unobserved latent error that is correlated with the endogenous variable. Again, I allow for the violation of CF-CI since the conditional median is a function of all the condition arguments including the instruments. Thus far, I have made no assumptions on what the function $h_o(\cdot, \cdot)$ should be and therefore the control function is completely general.

Now I can incorporate the general control function into the model assumptions to allow for

arbitrary endogeneity. The following assumptions include slight adjustments to Assumptions 2.7.1 and 2.7.2 to incorporate endogeneity.

Assumption 2.7.3 (General Conditional Median Restriction)**.** *In the set up described by equation (2.1)*

(i) $(v_{2i}, \boldsymbol{z}_i) \in \Re^{1+k_1+k_2}$ *is assumed to have density with respect to a Lebesgue measure, which is positive on the set* $(\mathcal{V} \times \mathcal{Z}) \subseteq \Re^{1+k_1+k_2}$.

(ii) *Let* $p_o(t, v_{2i}, \boldsymbol{z}_i)$ *denote* $P(-u_{1i} < t | v_{2i}, \boldsymbol{z}_i)$*, and assume*

(a) $p_o(\cdot, \cdot, \cdot)$ *is continuous on* $\Re \times (\mathcal{V} \times \mathcal{Z})$.

(b) $p_o'(t, v_{2i}, \boldsymbol{z}_i) = \partial p_o(t, v_{2i}, \boldsymbol{z}_i)/\partial t$ *exists and is continuous and positive on all* $\Re$ *for all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$.

(c) $p_o(h_o(v_{2i}, \boldsymbol{z}_i), v_{2i}, \boldsymbol{z}_i) = 1/2$ *where* $h_o(v_{2i}, \boldsymbol{z}_i)$ *is continuous on all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$.

(d) $\lim_{t \to -\infty} p_o(t, v_{2i}, \boldsymbol{z}_i) = 0$, $\lim_{t \to \infty} p_o(t, v_{2i}, \boldsymbol{z}_i) = 1$.

Assumption 2.7.4 (Endogenous Heteroskedastic Probit)**.** *In the set up described by equation (2.1)*

(i) $(\boldsymbol{z}_i, v_{2i}) \in \Re^{1+k_1+k_2}$ *is assumed to have density with respect to a Lebesgue measure, which is positive on the set* $(\mathcal{V} \times \mathcal{Z}) \subseteq \Re^{1+k_1+k_2}$.

(ii) $u_{1i} = \sigma_o(v_{2i}, \boldsymbol{z}_i) e_{1i} + h_o(v_{2i}, \boldsymbol{z}_i)$ *where* $\sigma_o(v_{2i}, \boldsymbol{z}_i)$ *is continuous and positive on* $(\mathcal{V} \times \mathcal{Z})$, $h_o(v_{2i}, \boldsymbol{z}_i)$ *that is continuous on all* $(v_{2i}, \boldsymbol{z}_i) \in (\mathcal{V} \times \mathcal{Z})$*, and* e_i *is independent of* $(v_{2i}, \boldsymbol{z}_i)$ *with any known (e.g. logistic, normal) distribution with median 0 and has a density function which is positive and continuous on the real line.*

Modifying the observational equivalence result to this setting is almost trivial. Instead of focusing the model on a zero median restriction, it is acknowledged that the median is

non-zero but a general conditional median function, $h_o(v_{2i}, z_i)$, is specified. Theorem 2.7.1 states this result with the proof provided in the appendix.

Theorem 2.7.1 (Observational Equivalence)**.** *Under Assumptions 2.7.3 and 2.7.4, the two models are observationally equivalent.*

Extending the results in Khan (2013) to the case of endogenous regressors is not novel. In a working paper, Song (2016) uses a more traditional control function method to address endogeneity. He imposes an exclusion restriction on the conditional median function, same as the conditional median independence assumption proposed in Krief (2014): $Med(u_{1i}|v_{2i}, \mathbf{z}_i) = f(v_{2i})$. Although these assumptions are weaker than CF-CI, the exclusion restrictions that are imposed are constrictive and unnecessary.

Utilizing Theorem 2.7.1, the following assumption is an alternative to Assumption 2.3.1 that allows for a non-parametric general control function and a non-parametric heteroskedastic function.

Assumption 2.7.5. *Consider the set up in equation (2.1), where* $\{y_{1i}, \boldsymbol{z}_i, y_{2i}\}_{i=1}^n$*, is iid. In the first stage, the true conditional mean is*

$$E(y_{2i}|\boldsymbol{z}_i) = m_o(\boldsymbol{z}_i)$$

and the unobserved latent error has the following conditional distribution

$$u_{1i}|\boldsymbol{z}_i, v_{2i}, y_{2i} = u_{1i}|\boldsymbol{z}_i, v_{2i} \sim N\Big(h_o(v_{2i}, \boldsymbol{z}_i), \exp(2 \times g_o(y_{2i}, \boldsymbol{z}_i))\Big)$$

Where $\boldsymbol{z}_i = (\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})$ *and* $m_o(\boldsymbol{z}_i)$*,* $h_o(v_{2i}, \boldsymbol{z}_i)$*, and* $g_o(y_{2i}, \boldsymbol{z}_i)$ *are unknown function.*

Since the normal distribution is used in this framework and is symmetric, the conditional median is equal to the conditional mean. Therefore the remaining discussion will return

to the conditional mean interpretation of endogeneity. With this slight variation from the parametric framework, Assumption 2.7.5 suggests a distribution free estimator comparable to the other semi-parametric estimators of the literature. To reiterate, in implementation the normal distribution is used according to Assumption 2.7.5, but in interpretation, there are no distributional strings attached because of the observational equivalence result.

However, this generalization further complicates identification. First, the model is only identified to scale, which can be solved with a normalization that assumes the last coefficient in a linear index $\mathbf{x}_i\beta_o$ is equal to 1: $\beta_{ko} = 1$. Similar to the previous literature, identification of the non-parametric heteroskedastic function is obtained by assuming that the last regressor, x_{ki}, conditional on all other random variables in the numerator has a density function with respect to the Lebesgue measure that is positive on $\Re$ and all other terms in the numerator have bounded support.[23] Second, as in the parametric model, introducing a general control function without any restrictions will not be identified relative to the linear index, $\mathbf{x}_i\beta$, because they rely on the same sources of variation. Using an analogous CMR, a shape restriction on the general control function insures that there is variation unexplained by the linear index.

[23]This is essentially assumption RC2(i) in Khan (2013). As he notes in Remark 3.2, the bounded support condition can be relaxed to the finite fourth moments. To illustrate how this condition is used in identification, consider the non-endogenous case where one would like to show identification β_o and σ_o from the choice probability $\Phi\left(\frac{\mathbf{x}_i\beta_o}{exp(\sigma_o(\mathbf{x}_i))}\right)$. First, suppose not, suppose there exists a $\beta \neq \beta_o$ and $\sigma \neq \sigma_o$ such that

$$\frac{\mathbf{x}_i\beta}{exp(\sigma(\mathbf{x}_i))} = \frac{\mathbf{x}_i\beta_o}{exp(\sigma_o(\mathbf{x}_i))}$$

for all $\mathbf{x}_i \in \mathcal{X}$. But because $\mathbf{x}_{ki}$, conditional on x_{-ki} has a density function with respect to the Lebesgue measure that is positive on $\Re$ and $\mathbf{x}_{-ki}$ is bounded, for any realization of $\mathbf{x}^*_{-k}$ in the support, there exists a x^*_k also in the support such that

$$\frac{\mathbf{x}^*_{-k}\beta_{-k} + x^*_k}{exp(\sigma(\mathbf{x}^*))} > 0 \text{ and } \frac{\mathbf{x}^*_{-k}\beta_{-ko} + x^*_k}{exp(\sigma_o(\mathbf{x}^*))} < 0$$

which is a contradiction.

Assumption 2.7.6. *Let $m_o(\cdot) \in \mathcal{M}$, $h_o(\cdot) \in \mathcal{H}$, and $g_o(\cdot) \in \mathcal{G}$ denote the function spaces and $\beta_{ko} \in \mathcal{B}$ denote the parameter space.*

(i) *$E(\boldsymbol{x}_i'\boldsymbol{x}_i)$ is non-singular, $E(\boldsymbol{x}_i|\boldsymbol{z}_i)$ is full rank.*

(ii) *(CMR) $E(h_o(v_{2i}, \boldsymbol{z}_i)|\boldsymbol{z}_i) = 0$*

(iii) *the last component of $\boldsymbol{x}_i$, x_{ki}, is an included instrument whose coefficient is normalized to 1 such that,*

$$\boldsymbol{x}_i\beta_o = \boldsymbol{x}_{-ki}\beta_{-ko} + x_{ki}$$

and x_{ki} conditional on $(\boldsymbol{x}_{-ki}, h_o(v_{2i}, \boldsymbol{z}_i))$ has a density function with respect to the Lebesgue measure that is positive on $\Re$ and $(\boldsymbol{x}_{-ki}, h_o(v_{2i}, \boldsymbol{z}_i))$ has bounded support.

The first two parts are taken from assumption 2.4.1. The last part imposes the scale normalization and is crucial in identifying the heteroskedastic function. There is no consensus in the literature on how to choose which regressor should have the scaled coefficient. Song (2016) uses the endogenous regressor since it will be continuously distributed and more likely to satisfy the support requirements. But then no inference can be made on the structural parameter whose value must be assumed to be non-zero. Therefore I suggest scaling on an instrument whose relevancy is not in question and has sufficient support.

A quick remark on parts (ii) and (iii): in some simple scenarios, the CMR in assumption (ii) is sufficient for the second part of assumption (iii), as long as x_{ki} conditional on $\mathbf{x}_{-ki}$ has a density function with respect to the Lebesgue measure that is positive on $\Re$. For instance, consider the linear case where the general control function is of the form

$$h_o(v_{2i}, \mathbf{z}_i) = \sum_{p \in \mathcal{P}} b_p(\mathbf{z}_i)(v_{2i}^p - E(v_{2i}^p|\mathbf{z}_i)) \tag{2.39}$$

where $b_p : \mathcal{Z} \to \Re$ and the set $\mathcal{P}$ consists of unique elements from the real line. Supposing

$\mathcal{P}$ does not include 0, then the CMR is satisfied. Also, for ease of understanding, consider the common case that $\mathbf{x}_{-ki} = y_{2i}$ so there is only one included instrument that is acting at the normalized covariate $x_{ki} = z_{1i}$. Then, conditional on any realization $h = h_o(v_{2i}, z_i)$ and $y_2 = y_{2i}$, one cannot precisely determine the corresponding z_{1i}. Therefore z_{1i} conditional on $(y_{2i}, h_o(v_{2i}, \mathbf{z}_i))$ has a density function with respect to the Lebesgue measure that is positive on $\Re$.

The following Theorem states the general identification result.

Theorem 2.7.2. *In the set-up described by equation (2.1) and Assumption 2.7.5, if Assumption 2.7.6 holds then $(m_o(\cdot), \beta_o, h_o(\cdot), g_o(\cdot))$ are identified.*

Proof is given in the appendix. Alternatively, if one is concerned that Assumption 2.7.6(iii) is unlikely to hold then the researcher may always turn to exclusion restrictions as a sufficient condition. If x_{ki} is assumed to be excluded in the control function, then, given the proper bounded and unbounded supports, part (iii) of Assumption 2.7.6 is easily satisfied.

With identification, the proposed estimation procedure is quite simple and similar to the parametric version in Section 4. In the first stage, the conditional mean function $E(y_{2i}|\mathbf{z}_i) = m_o(\mathbf{z}_i)$ is estimated using standard non-parametric regression techniques such as sieves or kernels. The control variable is constructed from the residuals $\hat{v}_{2i} = y_{2i} - \hat{m}(\mathbf{z}_i)$ and plugged into the second step. In the second stage, one would use sieves to estimate the non-parametric components. In this case sieves are preferred over other non-parametric methods since it will be easier to impose the CMR on the general control function. Let $\{b_l(v_{2i}, \mathbf{z}_i), l = 1, 2, ..., L_{hn}\}$ and $\{c_l(v_{2i}, \mathbf{z}_i), l = 1, 2, ..., L_{gn}\}$ be sequences of basis function of $(v_{2i}, \mathbf{z}_i)$ such

that $b_l(v_{2i}, \mathbf{z}_i)$ satisfy the CMR for all $l = 1, 2, ..., L_{hn}$. So the sieve spaces are defined as

$$\begin{aligned}\mathcal{H}_n = \{h : (\mathcal{V} \times \mathcal{Z}) \to \Re,\, h(v_{2i}, \mathbf{z}_i) = \sum_{l=1}^{L_{hn}} b_l(v_{2i}, \mathbf{z}_i)\gamma_l \\ : E(b_l(v_{2i}, \mathbf{z}_i)|\mathbf{z}_i) = 0 \text{ and } \gamma_1, ..., \gamma_{L_{hn}} \in \Re\}\end{aligned} \tag{2.40}$$

$$\mathcal{G}_n = \{g : (\mathcal{V} \times \mathcal{Z}) \to \Re, g(y_{2i}, \mathbf{z}_i) = \sum_{l=1}^{L_{gn}} c_l(y_{2i}, \mathbf{z}_i)\delta_l : \delta_1, ..., \delta_{L_{gn}} \in \Re\} \tag{2.41}$$

Finally, one would maximize the following likelihood

$$\begin{aligned}\mathcal{L}(y_{1i}, \mathbf{x}_i, \mathbf{z}_i; \hat{m}, \beta, \gamma, \delta) = \sum_{i=1}^{n} y_{1i} \log\left[\Phi\left(\frac{\mathbf{x}_i\beta + h(\hat{v}_{2i}, \mathbf{z}_i)}{\exp(g(y_{2i}, \mathbf{z}_i))}\right)\right] \\ + (1 - y_{1i}) \log\left[1 - \Phi\left(\frac{\mathbf{x}_i\beta + h(\hat{v}_{2i}, \mathbf{z}_i)}{\exp(g(y_{2i}, \mathbf{z}_i))}\right)\right]\end{aligned} \tag{2.42}$$

with respect to β, $h(\cdot) \in \mathcal{H}_n$, and $g(\cdot) \in \mathcal{G}_n$. Same as in the parametric version, to implement, this is as simple as running the `hetprobit` command in Stata for the second stage. However, inference should reflect both the two step estimation process and the non-parametric specification. The next section provides the asymptotic results for the proposed distribution free estimator.

2.7.2 Asymptotic Properties

Song (2016) derives consistency and convergence rates of the semi-parametric Het Probit (CF) estimator when the control function satisfies CF-CI and imposes the exclusion restriction (i.e.: $h_o(v_{2i}, \mathbf{z}_i) = h_o(v_{2i})$). Therefore the asymptotic results only need to be slightly augmented to allow for the general control function. Explanations of the notation is left to the appendix. The following assumption collects the remaining low level regulatory conditions needed for consistency of the second stage parameters.

Assumption 2.7.7.

(i) *Let* $(\beta_{-ko}, h_o(\cdot), g_o(\cdot)) \in (\mathcal{B} \times \mathcal{H} \times \mathcal{G}) = \Theta$ *denote the joint parameter space. For any* $(\beta_{-k}, h(\cdot), g(\cdot)) \in \Theta$,

(a) $\beta_{-k} \in \mathcal{B} \subset \Re^{k-1}$ *where* $\mathcal{B}$ *is compact,*

(b) $\Phi\left(\frac{\boldsymbol{x}_i\beta + h(v_{2i}, \boldsymbol{z}_i)}{\exp(g(y_{2i}, \boldsymbol{z}_i))}\right) \in \Lambda_c^s((\mathcal{V} \times \mathcal{Z}), w_1)$ *for* $s > 0$ *and* $w_1 \geq 0$,

(c) $h(v, \boldsymbol{z})$ *is continuously differentiable with respect to its first component such that*

$$\sup_{\boldsymbol{z} \in \mathcal{Z}} \sup_{v \in \mathcal{V}} \frac{\partial h(v_2, \boldsymbol{z})}{\partial v_2} < C < \infty$$

(ii) $\int (1 + ||(v, \boldsymbol{z})||^2)^{w_2} f_{v,\boldsymbol{z}}(v, \boldsymbol{z}) d(v, \boldsymbol{z}) < \infty$ *where* $f_{v,\boldsymbol{z}}(\cdot)$ *denotes the joint density function and* $w_2 > w_1 > 0$.

(iii) *For*

$$b^{L_{hn}}(v_{2i}, \boldsymbol{z}_i) = (b_1(v_{2i}, \boldsymbol{z}_i), ..., b_{L_{hn}}(v_{2i}, \boldsymbol{z}_i))$$

$$c^{L_{gn}}(y_{2i}, \boldsymbol{z}_i) = (c_1(y_{2i}, \boldsymbol{z}_i), ..., c_{L_{gn}}(y_{2i}, \boldsymbol{z}_i))$$

$E(b^{L_{hn}}(v_{2i}, \boldsymbol{z}_i)' b^{L_{hn}}(v_{2i}, \boldsymbol{z}_i))$ *and* $E(c^{L_{gn}}(y_{2i}, \boldsymbol{z}_i)' c^{L_{gn}}(y_{2i}, \boldsymbol{z}_i))$ *are non-singular for all* n.

Part (i) collects conditions on the parameter and functional space. The second component constraints the predicted probability function to be in a weighted Holder ball with radius c, smoothness s and weight function $(1 + ||\cdot||^2)^{w_1/2}$ as defined in equation (B.12) in the appendix. Since the parameter space is a weighted Holder ball, there exists a projection mapping from a standard sieve spaces constructed from power series, Fourier series, splines, or wavelets to the parameter space as $n \to \infty$. The last component allows for a Taylor expansion around the control variate v_{2i} since it is estimated in the first stage. Part (ii) replaces any compactness conditions on $(v_{2i}, \mathbf{z}_i)$ and the part (iii) insures point identification of the sieve coefficients.

The following theorem provides consistency of the proposed estimator. Proof is omitted since the arguments are identical to those provided in Song (2016).

Theorem 2.7.3. *In the set-up described by equation (2.1) where Assumptions 2.7.5, 2.7.6, and 2.7.7 hold, if the first stage estimator $\hat{v}_{2i}$ satisfies*

$$\sup_{(\boldsymbol{x}_i,\boldsymbol{z}_i)\in\mathbf{X}\times\mathcal{Z}} |\hat{v}_{2i} - v_{2i}| = O_p(\tau_v)$$

where $\tau_v = o_p(1)$, then the estimators that maximize the log likelihood in equation (2.20) are consistent,

$$||\hat{\beta}_{-k} - \beta_{-ko}|| = o_p(1)$$

$$\left|\left|\Phi\left(\frac{\boldsymbol{x}_i\hat{\beta} + \hat{h}(\hat{v}_{2i},\boldsymbol{z}_i)}{\exp(\hat{g}(y_{2i},\boldsymbol{z}_i))}\right) - \Phi\left(\frac{\boldsymbol{x}_i\beta_o + h_o(v_{2i},\boldsymbol{z}_i)}{\exp(g_o(y_{2i},\boldsymbol{z}_i))}\right)\right|\right|_{\infty,w_1} = o_p(1)$$

Note that consistency of the first stage non-parametric estimator can be obtained using Proposition 3.6 of Chen (2007) under fairly standard and relaxed conditions.[24] This theorem provides consistency of both the parametric component of the second stage estimator as well as the predicted probability function. This will be used in providing consistency of APE estimates in the following corollary.

Corollary 2.7.1. *Under the conditions of Theorem 2.7.3, the APE estimator with respect to component j of $\boldsymbol{x}_i$ is consistent:*

$$\left|\left| n^{-1}\sum_{i=1}^{n} \phi\left(\frac{\boldsymbol{x}_i\hat{\beta} + \hat{h}(\hat{v}_{2i},\boldsymbol{z}_i)}{\exp(\hat{g}(y_{2i},\boldsymbol{z}_i))}\right)\frac{\hat{\beta}_j}{\exp(\hat{g}(y_{2i},\boldsymbol{z}_i))} - E\left(\phi\left(\frac{\boldsymbol{x}_i\beta_o + h_o(v_{2i},\boldsymbol{z}_i)}{\exp(g_o(y_{2i},\boldsymbol{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i},\boldsymbol{z}_i))}\right)\right|\right| = o_p(1)$$

[24] The conditions for consistency of the first stage estimator include $\mathcal{Z} \in \Re^{k_1+k_2}$ is a Cartesian product of compact intervals, $E(y_{2i}|\mathbf{z}_i)$ is bounded, and $m_o(\mathbf{z}_i) = E(y_{2i}|\mathbf{z}_i) \in \Lambda^s$ where $s > (k_1 + k_2)/2$

The next assumption collects the remaining low level conditions needed to derive the convergence rate of the parametric component of the second stage estimator.

Assumption 2.7.8.

(i) *Any* $h(\cdot) \in \mathcal{H}$, $h(v_{2i}, \boldsymbol{z}_i) \in \Lambda_c^s((\mathcal{V} \times \mathcal{Z}), w_1)$ *for some* $s > 0$ *and* $w_1 \geq 0$.

(ii) *Any* $g(\cdot) \in \mathcal{G}$, $g(y_{2i}, \boldsymbol{z}_i) \in \Lambda_c^s((\mathcal{V} \times \mathcal{Z}), w_1)$ *for some* $s > 0$ *and* $w_1 \geq 0$.

(iii) *The smoothness exponent of the Holder space satisfies* $2s \geq 1 + k_1 + k_2$.

(iv) *In Assumption 2.7.7(ii),* $w_2 > w_1 + s$

This assumption places stronger smoothness assumptions as well as further restricts the tail behavior of the covariates. The convergence rates are derived for the one-step estimator where the control variate v_{2i} is assumed to be known and not estimated in a first stage. Therefore the estimator is defined as,

$$\tilde{\theta} \equiv (\tilde{\beta}_{-k}, \tilde{h}, \tilde{g}) = \underset{(\beta_{-k}, h, g) \in \mathcal{B} \times \mathcal{H}_n \times \mathcal{G}_n}{\arg\max} \sum_{i=1}^{n} y_{1i} \log \left[\Phi \left(\frac{\mathbf{x}_i \beta + h(v_{2i}, \mathbf{z}_i)}{\exp(g(y_{2i}, \mathbf{z}_i))} \right) \right] + (1 - y_{1i}) \log \left[1 - \Phi \left(\frac{\mathbf{x}_i \beta + h(v_{2i}, \mathbf{z}_i)}{\exp(g(y_{2i}, \mathbf{z}_i))} \right) \right] \tag{2.43}$$

The following theorem states the rate results, again the proof is omitted since the arguments are almost identical to those given in Song (2016).[25]

Theorem 2.7.4. *In the set-up described by equation (2.1) under Assumptions 2.7.5, 2.7.6, 2.7.7, and 2.7.8, then the estimator described in equation (2.43) satisfies*

$$||\tilde{\beta}_{-k} - \beta_{-ko}|| = O_p \left(\sqrt{\frac{\max(L_{hn}, L_{gn})}{n}} + L_{hn}^{-s/(1+k_1+k_2)} + L_{gn}^{-s/(1+k_1+k_2)} \right)$$

[25] The only difference is that now the control function is a function of both v_{2i} and the instruments z_i. Derivation of the convergence rates still follow from Theorem 3.2 of Chen (2007), but now the approximation rate of the control function converges at a slightly different rate: $||h_o - \pi_n h_o|| = L_{hn}^{-s/(1+k_1+k_2)}$.

The proposed estimator converges much slower than the parametric rate. This will effect the performance of the estimator as seen in the simulation study given in the following section.

2.7.3 Simulation

This simulation study will be a broad examination of the proposed Semi-Parametric Heteroskedastic Probit with a General Control Function (SP Het Probit (GCF)) in a variety of settings. There is one included and one excluded instrument drawn from the following joint distribution,

$$\begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \right) \tag{2.44}$$

The common data generating process is

$$\begin{aligned} y_{1i} &= \begin{cases} 1 & y_{1i}^* \geq 0 \\ 0 & y_{1i}^* < 0 \end{cases} \\ y_{1i}^* &= y_{2i}\beta_o + z_{1i} + u_{1i} \\ y_{2i} &= \pi_{1o} + \pi_{2o} z_{1i} + \pi_{3o} z_{2i} + v_{2i}; \end{aligned} \tag{2.45}$$

where $\beta_o = 1$ and $\pi_o = \left(-1/\sqrt{2}, -1/\sqrt{6}, 1/\sqrt{2}\right)'$. The control variate v_{2i} is drawn from a $N(0,1)$. This means that there is a strong first stage with an R^2 of approximately 0.50. The unobserved heterogeneity u_{1i} will be decomposed into the general control function and a mean zero random variable e_{1i} that determines the conditional distribution of the latent error,

$$u_{1i} = h_o(v_{2i}, \mathbf{z}_i) + \sqrt{2} e_{1i} \tag{2.46}$$

such that e_{1i} and $h_o(v_{2i}, \mathbf{z}_i)$ are standardized to have variance equal to one. This means $Var(u_{1i}) \approx 3$ and $Var(y_{2i}\beta_o + z_{1i}) \approx 2.45$. The simulation experiment considers three different control functions that satisfy the CMR:

$$
\begin{aligned}
h_o^1(v_{2i}, \mathbf{z}_i) &= v_{2i} \\
h_o^2(v_{2i}, \mathbf{z}_i) &= \left(z_{1i}/3 + z_{2i}/\sqrt{3}\right) v_{2i} \\
h_o^3(v_{2i}, \mathbf{z}_i) &= \left(1 + \frac{z_{1i}}{1 + (z_{1i}/3 - 2z_{2i}/3)^2}\right) v_{2i}/\sqrt{2.5}
\end{aligned}
$$

The coefficients in the linear control function, $h_o^2(v_{2i}, \mathbf{z}_i)$, are chosen so a projection of $h_o^2(v_{2i}, \mathbf{z}_i)$ on just the control variate (v_{2i}) only explains about 35% of the variation in $h_o^2(v_{2i}, \mathbf{z}_i)$. This means relaxing CF-CI should have meaningful consequences. The functional form and of the non-linear control function, $h_o^3(v_{2i}, \mathbf{z}_i)$, is chosen so a projection of $h_o^3(v_{2i}, \mathbf{z}_i)$ on $(v_{2i}, z_{1i}v_{2i}, z_{2i}v_{2i})$ explains 90% of the variation and therefore a linear approximation is very reasonable. Moreover the decomposition of variance explained is split 50-50 between just the control variate (v_{2i}) and the terms that are interacted with instruments ($z_{1i}v_{2i}$ and $z_{2i}v_{2i}$). Again, this means relaxing CF-CI should have meaningful consequences.

This simulation also considers four different conditional distributions for the latent error:

$$
\begin{aligned}
e_{1i}^1 &\sim Logistic(0, 3/\pi^2) \\
e_{1i}^2 &\sim Uniform(-\sqrt{12}/2, \sqrt{12}/2) \\
e_{1i}^3 &\sim \mathcal{T}(0, 3)/\sqrt{3} \\
e_{1i}^4 &\sim 0.5N\left(-0.8, (1 - 0.8^2)\right) + 0.5N\left(0.8, (1 - 0.8^2)\right) \\
e_{1i}^5 &\sim Logistic\left(0, (z_{1i}^2/2 + 3y_{2i}^2/4)/\pi^2\right)
\end{aligned}
$$

Notice that a combination of the first control function with any of the first three conditional distributions does not violate CF-CI. Only when the conditional distribution of the latent er-

ror is a function of the instruments, either through the control function or heteroskedasticity (as in e_{1i}^5), is CF-CI violated.

The simulation results will be presented in three segments. In the first segment, the data generating process satisfies CF-CI. This only allows for the first control function $h_o^1(v_{2i}, \mathbf{z}_i)$ and the first four conditional distributions $(e_{1i}^1, e_{1i}^2, e_{1i}^3, e_{1i}^4)$. The SML estimator is expected to perform as well, if not better, than the proposed method on all accounts (parameter estimates, ASF estimates, and APE estimates). The second segment considers the remaining two general control functions that do not satisfy CF-CI. Notice that $h_o^2(v_{2i}, \mathbf{z}_i)$ is linear in parameters and therefore can be estimated parametrically but $h_o^3(v_{2i}, \mathbf{z}_i)$ is a non-linear function and whose functional form will be treated as unknown and will be estimated non-parametrically. For the final segment heteroskedasticity is introduced so the non-parametric heteroskedastic function in the proposed estimator must capture both the misspecified distribution and the heteroskedasticity in the latent error.

In addition to the proposed semi-parametric estimator, the simulation experiment will employ the two step control function Probit estimator of Rivers and Vuong (1988) (Probit (CF)) and the SML estimator of Rothe (2009) as a comparison. The SML estimator is implemented with a Gaussian kernel of order 1. Although asymptotically the SML estimator requires higher order kernels, Rothe finds that lower order kernels perform better in small samples. As suggested in Rothe (2009), bandwidths for the SML estimator were treated as additional parameters to be optimized over. All three estimators use the same first stage estimates for v_{2i}: the residual from regressing y_{2i} on $\mathbf{z}_i$.

Two issues with the proposed method arose during implementation. First the proposed estimator is fairly sensitive to different starting values. But using 15 randomized starting values helps to avoid local maxima. Second, since the estimator incorporates two non-

parametric functions that need to be approximated via sieves, the number of parameters increases quite quickly. To reduce the number of parameters, a reasonable restriction on the general control function such that $h_o(v_{2i}, \mathbf{z}_i) = h_o(\mathbf{z}_i)v_{2i}$ is used. For sample size $n = 250$, both polynomial series approximating $h_o(\mathbf{z}_i)$ and $g_o(y_{2i}, \mathbf{z}_i)$ only include first order terms. For sample size $n = 500$ the polynomial series approximating $h_o(\mathbf{z}_i)$ only includes first order terms while the polynomial series approximating $g_o(y_{2i}, \mathbf{z}_i)$ includes both first and second order terms. For sample size $n = 1,000$, both polynomial series are up to order 2. Alternatively, one can consider a penalization method to restrict the number of non-zero covariates. This extension is left to future research.

All tables and figures referenced in this section can be found in the Appendix. They report the bias, standard deviation (Std. Dev.), root mean squared error (RMSE), and the 25^{th}, 50^{th}, and 75^{th} sample quantiles of the parameter estimates for 1,000 repetitions of the simulation. The following summarizes the results for the three cases: CF-CI holds, CF-CI is violated by a general control function, and CF-CI is violated by heteroskedasticity.

CF-CI holds

Tables E.11-E.14 report the simulation results for the estimates of β_o when CF-CI holds. The results show all three estimators perform fairly well in terms of bias even though the Probit (CF) estimator imposes a misspecified distribution (assumes normality when it does not hold). The only exceptions are the proposed estimator under a Uniform and Gaussian Mixture distribution for the latent error. This bias is stronger when $n = 500$ and $n = 1,000$ which is when there are a larger number of higher order terms for the non-parametric functions. This would suggest that there may be gains to adding a penalized approach to control for potentially large number of irrelevant terms.

Also the proposed estimator is much less efficient than the alternate methods. The

efficiency gain for using the SML estimator can be substantial. In the cases of the Uniform Distribution and the Gaussian Mixture, the standard deviation can be reduced by 1/2 when using the SML estimator instead of the proposed approach. Nevertheless, I see all the estimators perform well in terms of ASF estimates in Figure D.9-D.12.

Violation of CF-CI with General Control Function

Now, the data generating process includes general control functions that violates CF-CI by including the instruments. Recall that there is two possible general control functions: one that is linear in parameters (h_o^2), so that it is consequently parametrically specified, and another that is non-linear and must be estimated non-parametrically.

Tables E.15-E.18 report the β_o estimate results when the control function is linear. The simulation results suggest SML estimator has a fairly strong negative bias compared to the other two estimators where in some cases the 75^{th} sample quantile falls below the true parameter value of 1. Surprisingly, the parametric Probit (CF) estimator does not have as strong of a bias as the SML estimator, even though it is also implicit imposing CF-CI. But the proposed estimation approach still incurs fairly large standard deviation due to the numerous parameters needed in estimation. Therefore both of the other approaches, Probit (CF) and SML, fair better in terms of RMSE.

This places some doubt onto the proposed approach and whether any realist gains on previous approaches is even possible. But examining the ASF estimates clarifies the matter. The figures show that the proposed SP Het Probit (GCF) substantially outperforms the other estimators by better estimating true ASF across all the different distributions. Tables E.19-E.22 and Figures D.17-D.20 report the results when the control function is unknown and estimated non-parametrically. The conclusions stay the same although there tends to be a larger bias (compared to the cases of linear general control function) for the proposed

estimator. But this is because the form of general control function is unknown and can only be approximated.

Comparing these results to the previous segment, distributional misspecification appears to have a lighter effect on the parameter estimates than violations of CF-CI. The Probit (CF) estimator performs quite well under distributional misspecification while the Probit (CF) and SML estimators display stronger bias and poor ASF estimates when CF-CI does not hold. Moreover, the proposed estimator always has a larger RMSE compared to the other approaches which suggests that there is consequential trade-off in terms of efficiency of the SML estimator and smaller bias of the proposed approach.

Violation of CF-CI with Heteroskedasticity

The final segment only looks at the Logistic distribution for the latent error but introduce heteroskedasticity as a further violation of CF-CI. Simulation results are reported in Tables E.23-E.25. When the sample size is 250, only the first order polynomials are included in approximating the heteroskedasticity. Consequently, the proposed estimator performs poorly and on par (in terms of bias) with the SML estimator. But when I allow for higher order terms when the sample size increases, the bias of the proposed estimator diminishes significantly compared to the alternative methods. Examining the ASF estimates in Figures D.21-D.23, only the SP Het Probit (GCF) estimator follows the true ASF closely which the other two estimators suffer even under the simplest control function $h_o^1(v_{2i}, \mathbf{z}_i) = v_{2i}$.

Overall the proposed estimator correctly adapts to the scenario in which CF-CI is violated while the alternative estimation methods are restricted by imposing the assumption. An additional benefit to the proposed method is much simpler to implement and can be done using canned commands in common statistical packages. So for an applied researcher the proposed method is more general than alternative estimators and is easier to implement.

However this simulation also brings to light some weaknesses of the SP Het Probit (GCF) estimator. First, this estimator is quite inefficient due to the large number of parameters it needs to estimate. Therefore the proposed procedure could benefit by a dimension reduction. Second, the proposed estimator is quite sensitive to starting values and without prior knowledge of what the true parameter value should be, this may pose some challenge to implementation. Randomizing around scaled parameter estimates from the Probit (CF) estimator for starting values is a promising possibility, as this simulation study shows. The parametric estimator tends to perform fairly well even under violation of CF-CI and with a misspecified distribution.

2.8 Conclusion

This chapter presents a new control function approach to endogeneity in a binary response model that does not impose CF-CI. Applying a similar framework as Kim and Petrin (2017), this chapter uses a general control function method that allows the instruments to be a part of the conditional distribution of the unobserved heterogeneity. The proposed estimator is consistent and asymptotically normal. In simulations, it is shown that the general control function method is necessary to obtain accurate parameter estimates under the weaker CMR setting. Moreover, structural objects of interest such as the ASF and APE can be recovered in the general framework presented in the chapter. Without CF-CI, other estimators of the literature are unable to correctly estimate the ASF and APE resulting in inaccurate economic interpretations. In the empirical application, a Wald test uncovers strong statistical evidence for the violation of CF-CI, although there are fairly minimal difference in the economic interpretations produced by the different estimators.

The proposed estimator is introduced in a parametric framework which may be unrealistic in some economic settings. Therefore a semi-parametric extension is provided that places no distributional assumptions on the unobserved heterogeneity. Simulations show that when CF-CI is violated and the distribution of the latent error is misspecified, the proposed semi-parametric estimator consistently estimates the parameters and the ASF. But, the simulations also uncover some drawbacks to the proposed semi-parametric approach. Due to the fairly large dimension of the parameter space, the proposed approach is quite inefficient relative to other estimators in the literature. An interesting avenue for further research is to develop a more efficient semi-parametric estimator that still allows for the relaxation of CF-CI.

The motivation for this chapter was to propose an estimation procedure built upon a model and assumptions that are much more reflective of what we would expect in empirical data. By creating a model that is much more flexible and realistic as well as an estimation procedure that is easy to implement, the proposed approach will be a useful addition to an economists tool-kit of estimators. The next chapter approaches a different setting but with a similar purpose. In Chapter 3, a joint work with Jeffrey Wooldridge and Ying Zhu, we consider a panel binary response (large N, small T) in which the standard joint maximum likelihood approaches have simple and restrictive specifications for the individual heterogeneity and do not allow for serial correlation. Empirical data calls for more flexibility so we propose an approach that can capture individual persistence through several mechanisms. First, we introduce individual heterogeneity in the levels and the slopes that are allowed to be potentially correlated with the covariates. Second, we allow for serial correlation in the latent error. The resulting estimator is a pooled correlated random effects heteroskedastic Probit in which identification will again rely on the results provided in the first chapter.

Both of the proposed approaches in these two chapters will find their utility in empirical work as they push the frontiers of the literature on how to incorporate flexibility driven by the demands of data.

Chapter 3

Behavior of Pooled and Joint Estimators in Probit Model with Random Coefficients and Serial Correlation[1]

3.1 Introduction

Multilevel data analysis is among the long standing statistical tools that leverages heterogeneity in the data. One of the most frequent occurrences in application is panel data where the first level is time and the second level is individuals. Given the broad framework provided by a multi-level setting, there is an absence of times series analysis in the multilevel literature that appears in panel data settings. In particular, when observations are recorded over time we expect the data is display a strong amount of persistence. This persistence can arise with individual-specific heterogeneity or with serially correlated errors.

Economic theory can provide motivation as to why would expect to see persistence in the data. In modelling demand, purchasing behavior can be traced back to a utility max-

[1]This is joint work with Jeffrey Wooldridge and Ying Zhu

imization problem where if one allows for heterogeneous agents – in preferences or income effects – the estimating equation should allow for individual heterogeneity. In Wooldridge (2010), the individual-specific heterogeneity in a program evaluation framework is motivated by "the usual omitted ability story." The individual-specific heterogeneity controls for any individual characteristic – such as ability or motivation – that may be correlated with program participation. In the application of these examples, the individual heterogeneity is an unobserved random variable.

Even after allowing for individual-specific heterogeneity, one would expect a strong presence of serial correlation in the errors. In the field of labor economics, outcomes such as employment, wages, and health outcomes are strongly persistent and exhibit clear signs of auto-correlation. Bertrand, Duflo, and Mullainathan (2004) survey the empirical literature on evaluating treatment effect that apply the Difference in Difference technique and found that out of 69 studies, only 5 explicitly address serial correlation. They also show that the consequences of not correcting for serial correlation can be severe for inference. By evaluating placebo interventions, ignoring serial correlation can result in concluding a "effect" at the 5 percent level for up to 45 percent of the placebo interventions.

This chapter intends to further explore the effects of persistence – individual-specific heterogeneity and serial correlation – in popular estimation procedures in a binary response setting. We are interested in any robustness properties these estimation procedures may provide either theoretically or in simulations. The most common formulation of a model for a panel binary response, $y_{it} \in \{0, 1\}$, is derived from the latent variable set up that allows for level individual heterogeneity.

$$y_{it} = 1\{a_i + x_{it}\beta + \varepsilon_{it} > 0\} \tag{3.1}$$

where x_{it} is a vector of observed random variables – the covariates– and a_i is an unobserved random variable – the individual heterogeneity. To begin, we will assume a_i, x_{it}, and ε_{it} are all independent from one another. If we make the following "random effects" assumption,

$$a_i|x_{i1}, ..., x_{iT} \sim N(\alpha, \sigma_1^2) \tag{3.2}$$

then a Joint Maximum Likelihood Estimation (JMLE) procedure that integrates the random effect a_i is consistent. If we assume the idiosyncratic error ε_{it} is normally distributed this results in the random effect Probit estimator. Alternatively, if we assume a logistic distribution then random effects Logit estimator is used. Estimation can be computational more difficult given a logistic distribution since it does not mix well with any other distribution. Consequently, this chapter will more heavily examine the Probit case, but most of the analysis can be extended to the Logit case as well.

However the conditional independence assumption that is implicit in the random effects assumption in equation 3.2 can be quite stringent. In the linear panel data model literature in econometrics, there are two popular modelling approaches in the literature to relax the conditional independence of a_i and the x's. One is the Fixed Effect (FE) approach and the other is the Mundlak device. The FE approach runs a pooled OLS regression of the form $(y_{it} - \bar{y}_i)$ on $(x_{it} - \bar{x}_i)$. This allows for arbitrary correlation between the individual heterogeneity a_i and the x's. Alternatively, one can model the correlated random effects using a Mundlack device. The Mundlack device proposes allowing a_i to be correlated with the x's through time constant functions of the data. A common implementation is to use the time averages $\bar{x}_i$. Wooldridge (2018), Proposition 2.1, shows that running a pooled OLS regression of the form y_{it} on x_{it}, $\bar{x}_i$ yields the same estimates for as the FE approach. However, the FE approach does not allow us to estimate functions that involve the conditional mean of

the heterogeneity, which might be of interest in certain applications (as we will see soon).

Extending the discussion to a non-linear setting, using the FE approach results in an incidental parameters problem and consequently serious biases in the coefficient estimates. Greene (2004) and Fernández-Val (2009) provide a more complete discussion of the FE approach for Probit. To avoid these issues, we propose applying the Mundlack device to the setting of equation (3.1) by assuming

$$a_i = \alpha + \bar{x}_i \xi_a + u_{1i} \quad \text{where} \quad u_{1i} | x_{i1}, ..., x_{iT} \sim N(0, \sigma_1^2). \tag{3.3}$$

We could use a JMLE procedure that integrates out the random effect u_{1i} or, as in the linear case, we could consider a simpler Pooled Maximum Likelihood (PMLE) approach. The PMLE approach makes no assumptions on the joint distribution over the time observations but pools the likelihood over i and t.

So far, we have only introduced individual heterogeneity into the level but there is little reason as to why the individual heterogeneity should be restricted to a level effect. While introducing random slopes is much more common in the linear regression literature (see Hall, Horowitz, et al. (2005), Swamy (1970), and Swamy and Tavlas (1995)), there has been fewer papers that attempt to account for the unobserved heterogeneity in slope parameters in a nonlinear model like Probit.[2] One of the reasons has to do with the fact that joint estimation methods are so far the dominant approach. A JMLE approach, which requires obtaining the joint distribution of $(y_{i1}, ..., y_{iT})$ conditional on $(x_{i1}, ..., x_{iT})$, can be computationally difficult, and we may not even have enough assumptions to obtain the joint distribution. In any case, the JMLE will generally require more assumptions to consistently estimate the parameters. The benefit from the additional assumptions and computational burden is

[2]Hausman and Wise (1978) and Akin, Guilkey, and Sickles (1979) introduce random coefficients in the multinomial and ordered Probit models.

greater asymptotic efficiency.

Extending the specification in equation (3.1) we will assume,

$$y_{it} = 1\{a_i + x_{it}b_i + \varepsilon_{it} > 0\} \tag{3.4}$$

where now both a_i and b_i are unobserved random variables capturing the level and slope individual heterogeneity.

Before we dive into the more complicated joint methods, perhaps it would be wise for us to take a step back and ask the following question: What features of the model should we focus on? In evaluating policy interventions, the ultimate interest usually concerns the treatment effect. While the average treatment effect coincides with the slope coefficient in a linear model with only additive heterogeneity, in a nonlinear model like the one above, the average treatment effect is much more complex in its derivation.

The concepts of the Average Structural Function (ASF) are simultaneously proposed by Blundell and Powell (2004) and Wooldridge (2005), in which the average treatment effect should be derived from the ASF. Using the notation in Wooldridge (2005), the conditional mean of y is defined as, $E(y|x,q) = \mu_1(x,q)$ where x are observed covariates – x_{it} in the setup above – and q is unobserved heterogeneity – a_i and b_i in the setup above. Assuming the standard distributional assumptions in the Probit case ($\varepsilon \sim N(0,1)$ and independent of all other random variables), applied to our model of interest:

$$\mu_1(x_{it}, (a_i, b_i)) = \Phi(a_i + x_{it}b_i). \tag{3.5}$$

Then the ASF averages the above equation over the distribution of unobserved heterogeneity. The treatment effect is the difference of the ASF over the treated and not treated, but this can vary over the observed covariates. Unlike the linear model, the complexity of the non-linear

model allows the treatment effect to be heterogeneous over the distribution of the covariates. Then averaging over the distribution of the observed covariates renders the average treatment effect.

It is useful to begin with a framework that unifies the discussion of treatment effects in models with unobserved heterogeneity. Average treatment effect is usually reserved for cases of binary treatment and average partial effect for the continuous analogue. In the continuous case, in lieu of taking difference, the partial derivative of the ASF produces the partial effect. We will refer to the Average Partial Effect (APE) synonymously for the binary and continuous case.

If our focus is on the APEs, then adopting the Mundlak device to model the unobserved heterogeneity in slope parameters would seem a sensible approach. Combined with a pooled estimation method, the Mundlak approach treats the data as if it is one long cross section and computation is typically straightforward.

Motivated by Mundlack, we return to setting of equation (3.4) and assume

$$a_i = \alpha + \bar{x}_i\xi_a + u_{1i} \quad \text{where} \quad u_{1i}|x_{i1}, ..., x_{iT} \sim N(0, \sigma_1^2) \tag{3.6}$$

$$b_i = \alpha + \bar{x}_i\xi_b + u_{2i} \quad \text{where} \quad u_{2i}|x_{i1}, ..., x_{iT} \sim N(0, \sigma_1^2). \tag{3.7}$$

As previously mentioned will focus on two different estimation routes: JMLE and PMLE. The Joint MLE procedure derives joint distribution of $(y_{i1}, ..., y_{iT})$ conditional on $(\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$ by integrating out the random effects (u_{1i}, u_{2i}). This integral is not solved in closed form and in estimation is approximated using numerical methods. This can cause more computational issues including failures due to non-convergence and long estimation times. Because it is a full MLE method, the JMLE produces the efficient estimates of the parameters $\alpha, \beta, \xi_a, \xi_b, \sigma_1^2$, and σ_2^2. However it does assume ε_{it} is iid over i and t and (u_{1i}, u_{2i}) are bivariate normal

independent of ε_{it} and $(\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$. These assumptions could be relaxed theoretically, but trying to implement the more flexible models turn out computationally costly so that the current state of statistical software makes these assumptions in implementation.

In the alternative pooled framework, it is computationally easy for us to relax the assumption that ε_{it} is independent over i and t. Since we are considering a panel setting, it is natural to expect serial correlation in the latent error. Although it is not as efficient as a joint procedure, it is robust to serial correlation. The drawback to this approach is that it cannot separately identify the coefficients $(\alpha, \beta, \xi_a, \xi_b)$ from the scaling factor $1/\sqrt{1+\sigma_1^2}$ but it is consistent in estimating the scaled parameters, $\theta_\sigma = \theta/\sqrt{1+\sigma_1^2}$ where θ represents any of the coefficients.

This leads to an interesting trade-off between the two estimation procedures. The JMLE can separately identify and estimates the variance of the random effects and should be more efficient, but is not robust to serial correlation and may be computationally more demanding. On the other hand, the PMLE is robust to serial correlation but is less efficient and can only estimate the scaled coefficients. However, if we focus on the APEs, then the lack of identification would not pose any issue in the interpretations of the results. In this case, it is possible that precise estimates of individual coefficients may have a much smaller impact on the estimates of the APEs.

We conduct extensive simulation experiments for the Probit model comparing the JMLE and PMLE. We look at both the continuous – with and without strong dependence – and binary treatment cases. The pooled approach performs as we expect: less efficient but consistent over different levels of serial correlation in the latent error. We do find some surprising trends in the coefficient estimates using the JMLE procedure. We find that even under no serial correlation, the coefficients estimates have a serious negative bias. The

driving factor appears to be poor estimation of the variance components, σ_1 and σ_2, which tend to also have significant negative bias. But these biases seem to cancel each other out when examining the estimates of the scaled coefficients, even under the presence of serial correlation. Consequently, the APE calculated from the JMLE estimates appear to have robustness properties with respect to serial correlation.

The remainder of this chapter is organized as follows. Section 2 presents the model set up and assumptions. Section 3 goes into more detail in deriving the two estimation procedures. In particular, we discuss how the JMLE procedure fits into the Generalized Linear Mixed Effects Model literature and how the pooled approach results in a heteroskedastic Probit estimator. Section 4 derives the average structural function and the corresponding APE. We provide a more detailed discussion on how the heterogeneity is incorporated into the APEs. Section 5 present the specifications for the simulation study and discusses the results. Section 6 uses the two estimators in an application to investigate if our simulation results hold with empirical findings. Section 7 presents a short discussion on extending this analysis to the Logit case. There is no easy implementation of a pooled approach since no distribution mixes well with the logistic distribution and the JMLE approach does not provide consistent estimates of coefficients which leads to the question of what is a more robust statistic: the APEs or the log-odds. Finally we conclude with a summary of our results and their implications.

3.2 Model Set Up

We are considering a binary response model in a panel setting with small T and large N allowing for correlated random effects in the intercept as well as the coefficient of interest,

$$\begin{aligned} y_{it} &= 1\{a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2 + \varepsilon_{it} > 0\} \\ a_i &= \alpha + g_1(\bar{x_i})\xi_a + u_{1i} \\ b_i &= \beta_1 + g_2(\bar{x_i})\xi_b + u_{2i}. \end{aligned} \tag{3.8}$$

Motivated by the Mundlack device, we will assume that the elements of $g_1(\cdot)$ and $g_2(\cdot)$ are known functions of the time averages $\bar{\mathbf{x}}_i$. This could of course be generalized to know functions of all the time observations. The random intercept and coefficients are allowed to be correlated with all time observations of the x's, $(\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})$, through the linear functions $g_1(\bar{\mathbf{x}}_i)\xi_a$ and $g_2(\bar{\mathbf{x}}_i)\xi_b$. We will assume the following independence and distributional assumptions hold:

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \Big| x_{i1}, ..., x_{iT} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right). \tag{3.9}$$

This means that the random effects a_i and b_i have a known distribution and are independent of the x's conditional on the time averages, $\bar{\mathbf{x}}_i$. Generally, u_{1i} and u_{2i} are allowed to drawn from a multivariate normal distribution with possible correlation, however we found that, in the simulation, allowing for a general variance covariance structure was quite straining to compute. In the remainder of this chapter we will assume $\sigma_{12} = 0$ but the analysis can be easily extended to allow for the correlated case. Finally, the idiosyncratic error ε_{it} is assumed to be independent of all other random variables in the model and independent over i. In order to allow persistence in the outcome that is unrelated to the covariates, ε_{it} is

serially correlated over t following an AR(1) process,

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + e_{\varepsilon it} \tag{3.10}$$

An AR(1) process does a fair job at modeling the persistence in outcomes that we see in empirical data. However, this could be extended even further by allowing for an AR(p) process or even a ARMA(p,q).

3.3 Estimation Methods

Given the set-up above, we will consider two different estimation procedures: JMLE and PMLE. Section 15.8 of Wooldridge (2010) reviews these two estimation methods (as well as others) and their accompanying assumptions and implications with only additive individual heterogeneity. We extend this by introducing slope individual heterogeneity and focus on the implications of serial correlation.

3.3.1 Mixed Effects Probit

We will first look at a JMLE, referred to as the Mixed Effects (ME) Probit, which can be derived through two similar but different framework: one can be viewed as an extension to the Random Effects Probit (described in Wooldridge (2010) chapter 15.8) to allow for a random coefficient and the other as a Generalized linear Mixed Model (GLMM) with a Bernoulli distribution and a Probit link function.

Under the first framework, we may consider the set up described in equation (3.8), so the marginal density of y_{it} conditional on the contemporaneous regressors and random

coefficients is,

$$f(y_{it}|\mathbf{x}_{it}, a_i, b_i) = \Phi\left(a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2\right)^{y_{it}} \left(1 - \Phi\left(a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2\right)\right)^{(1-y_{it})}. \quad (3.11)$$

If one were to assume independence across t, the joint density (over time) will be a product of the marginal densities. By allowing for correlated random effects and integrating out the random effects (u_{1i}, u_{2i}) we obtain,

$$\begin{aligned} f(y_{i1}, ..., y_{iT}|\mathbf{x}_{i1}, ..., \mathbf{x}_{iT}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\prod_{t=1}^{T} &\Phi\big(\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a \\ &+ x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b + u_1 + x_{1it}u_2\big)^{y_{it}} \times \Big(1 - \Phi\big(\alpha + x_{1it}\beta_1 \\ &+ \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b + u_1 + x_{1it}u_2\big)\Big)^{(1-y_{it})} \\ &\times \frac{1}{\sigma_1\sigma_2}\phi\left(\frac{u_1}{\sigma_1}\right)\phi\left(\frac{u_2}{\sigma_2}\right)\mathrm{d}u_1\mathrm{d}u_2 \end{aligned} \quad (3.12)$$

where σ_1 and σ_2 are the standard deviations of u_{1i} and u_{2i} respectively. The integral is not solved in closed form but can estimated using numerical methods.[3] Taking the log of equation (3.12) gives the conditional log likelihood for each i. Maximizing the sum, over i, of the log likelihood with respect to the parameters $\alpha, \beta_1, \beta_2, \sigma_1$, and σ_2 produces the JMLE estimator.

As for the second framework, let us consider the definition of GLMM in chapter 4 of McCulloch and Neuhaus (2001) (with notations altered slightly),

$$\begin{aligned} Y_i(= (y_{i1}, ..., y_{iT}))|\mathbf{u} &\sim \text{ independent } f_{Y_i|\mathbf{u}}(y_i|\mathbf{u}) \\ h(E(Y_i|\mathbf{u})) &= \mathbf{X}_i B + \mathbf{Z}_i\mathbf{u} \\ \mathbf{u} &\sim f_U(u) \end{aligned} \quad (3.13)$$

where $\mathbf{X}$ and $\mathbf{Z}$ are considered fixed design matrices and $\mathbf{u}$ is the only random effect such

[3]The simulation uses a mean-variance adaptive Gauss–Hermite quadrature, but other procedures such as a Laplacian approximation could be used.

that $\mathbf{X}B$ is the fixed component and $\mathbf{Z}u$ is the random component. Note that in our set up, $h(\cdot)$ is the inverse of the Standard Normal CDF which is why this estimator will be referred to as the Mixed Effects (ME) Probit estimator. Then defining the components in equation (3.13) to match the set up described by equation (3.8) yields,

$$\begin{aligned}
\mathbf{X}_i &= \begin{pmatrix} 1_T, & \mathbf{x}_{1i}, & \mathbf{x}_{2i}, & 1_T \times g_1(\bar{\mathbf{x}}_i), & 1_T \times g_2(\bar{\mathbf{x}}_i) \end{pmatrix} \\
B' &= \begin{pmatrix} \alpha & \beta_1 & \beta_2' & \gamma_a' & \gamma_b' \end{pmatrix} \\
\mathbf{Z}_i &= \begin{pmatrix} 1_T, & \mathbf{x}_{1i} \end{pmatrix} \\
\mathbf{u} &= \begin{pmatrix} u_{1i} \\ \\ u_{2i} \end{pmatrix}
\end{aligned}$$

where 1_j is a $j \times 1$ vector ones and $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ are the stacked time observations for individual i. Then the standard GLMM estimator computes the log likelihood under the assumption of independence across t and then integrating out the random effects $\mathbf{u} = (u_{1i}, u_{2i})'$.

There are several concerns that should be addressed with this estimator. First, in practice, the second level equations (defining a_i and b_i) are often not given a flexible specification that allows for correlation with the regressors $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$. It is perhaps the assumption of "fixed design matrices" in the GLMM literature that leads to a general lack of concern for correlation between the random effect $\mathbf{u}$ and the design matrices $\mathbf{X}$ and $\mathbf{Z}$. As our simulation study will show, not allowing for correlated random effects will result in heavily biased parameter estimates.

Second, the JMLE can be quite computationally demanding. The discussion for the results of the simulation study will provide more detail, but to summarize, the ME Probit estimator is more likely to fail to converge and if it does converge, takes much longer than the alternate estimator. The failure of converges is more frequent when the true data generating

process does not have any random effects and therefore the parameters are at the boundary of the identified set ($\sigma_1^2 = \sigma_2^2 = 0$). The slower speeds are because the ME Probit estimator must numerically approximate several integrals.

Third, note that this estimator depend on independence across t. Since we are introducing serial correlation through an AR(1) process, we would expect the estimator to be inconsistent. In particular, let us re-examine the joint distribution of $(y_{i1}, ..., y_{iT})$ conditional on the x's assuming correlation over t. Suppose,

$$\begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix} \sim N(0, \Sigma_\varepsilon)$$

where in an AR(1) process with correlation coefficient ρ, Σ_ε will have the form,

$$\Sigma_\varepsilon = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & \dots & \rho & 1 \end{pmatrix}. \tag{3.14}$$

By the properties of conditional distributions,

$$\begin{aligned} f(y_{i1}, ..., y_{iT}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, a_i, b_i) =& f(y_{iT}|y_{i1}, ..., y_{iT-1}, \mathbf{x}_{1i}, \mathbf{x}_{2i}, a_i, b_i) \\ & \times f(y_{iT-1}|y_{i1}, ..., y_{iT-2}, \mathbf{x}_{1i}, \mathbf{x}_{2i}, a_i, b_i) \\ & \times \cdots \times f(y_{i2}|y_{i1}, \mathbf{x}_{1i}, \mathbf{x}_{2i}, a_i, b_i) \\ & \times f(y_{i1}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, a_i, b_i), \end{aligned} \tag{3.15}$$

solving for the conditional means of y_{it} for $t \geq 2$ yields

$$
\begin{aligned}
&E(y_{it}|y_{i1},...,y_{it-1},\mathbf{x}_{1i},\mathbf{x}_{2i},a_i,b_i) \\
&\quad = E(E(y_{it}|u_{it-1},y_{i1},...,y_{it-1},\mathbf{x}_{1i},\mathbf{x}_{2i},a_i,b_i)|y_{i1},...,y_{it-1},\mathbf{x}_{1i},\mathbf{x}_{2i},a_i,b_i) \\
&\quad = E\left(\Phi\left(\frac{a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2 - \rho u_{it-1}}{\sqrt{1-\rho^2}}\right)\middle| y_{i1},...,y_{it-1},\mathbf{x}_{1i},\mathbf{x}_{2i},a_i,b_i\right) \\
&\quad = \left[\int_{a_i+x_{1it-1}b_{1i}+\mathbf{x}_{2it-1}\beta_2}^{\infty} \Phi\left(\frac{a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2 - \rho u}{\sqrt{1-\rho^2}}\right)\phi(u)\ \mathrm{d}u\right]^{y_{it-1}} \\
&\qquad \times \left[\int_{-\infty}^{a_i+x_{1it-1}b_{1i}+\mathbf{x}_{2it-1}\beta_2} \Phi\left(\frac{a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2 - \rho u}{\sqrt{1-\rho^2}}\right)\phi(u)\ \mathrm{d}u\right]^{(1-y_{it-1})} \\
&\quad \equiv E(y_{it}|y_{it-1},x_{1it},x_{1it-1},\mathbf{x}_{2it},\mathbf{x}_{2it-1},a_i,b_i).
\end{aligned}
$$

In words, the mean of y_{it} conditional on all past observations and the random effects a_i and b_i is only a function of the data from the last time period and the random effects. However, due to the nonlinearity of the Probit model, the conditional mean relies on the past data in a complicated manner. It is the difference between $E(y_{it}|y_{it-1},x_{1it},x_{1it-1},\mathbf{x}_{2it},\mathbf{x}_{2it-1},a_i,b_i)$ and $E(y_{it}|x_{1it},\mathbf{x}_{2it},a_i,b_i)$ that would suggest the ME Probit estimator is inconsistent under an AR(1) process.

Of course one could consider estimating using a joint likelihood based on a AR(1) model rather than assume independence. However this would require correctly specify the dependence structure as AR(1), where in empirical data, a simple AR(1) process may not be able to truly capture the complex time dependencies. Moreover, Keane (1994) discusses the difficulty of doing so directly and instead proposes a simulated variation of a Method of Moments estimator. In this chapter we do not consider a simulated version of the JMLE method since this is done rarely in practice.

3.3.2 Pooled Heteroskedastic Probit

The PMLE method is an alternative to a JMLE method and requires fewer assumptions. Unlike the ME Probit, the Pooled Heteroskedastic Probit does not depend on correct specification of the joint likelihood and therefore is consistent under the presence of serial correlation. Note that the pooled method will produce a conditional mean similar to a heteroskedastic probit. The heteroskedasticity in the pooled method is due to the heterogeneous slope coefficient, rather than the traditional interpretation of heteroskedasticity in the latent error. In deriving the Pooled Heteroskedastic Probit estimator, we apply the assumptions stated in section 2, $(u_{1i} + x_{1it}u_{2i} + \varepsilon_{it})|\mathbf{x}_{1i}, \mathbf{x}_{2i} \sim N(0, 1 + \sigma_1^2 + x_{1it}^2\sigma_2^2)$ and

$$E(y_{it}|\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Phi\left(\frac{\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1it}^2\sigma_2^2}}\right). \tag{3.16}$$

Plugging the conditional mean into the Bernoulli density, taking logs, and pooling over i and t, yields the following log likelihood

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} \ln\left(\Phi\left(\frac{\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1it}^2\sigma_2^2}}\right)\right) \\ \times (1 - y_{it}) \ln\left(1 - \Phi\left(\frac{\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1it}^2\sigma_2^2}}\right)\right)\end{aligned} \tag{3.17}$$

A standard practice in statistical packages is to assume an exponential function variance function which insures a strictly positive variance in estimation. We can then approximate equation (3.16) with,

$$E(y_{it}|\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Phi\left(\frac{\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b}{\exp(1/2\ln(1 + \sigma_1^2 + x_{1it}^2\sigma_2^2))}\right)$$

$$= \Phi\left(\frac{\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1+\sigma_1^2}\exp(1/2\ln\left(1+\frac{\sigma_2^2}{1+\sigma_1^2}x_{1it}^2\right)}\right)$$
$$= \Phi\left(\frac{\alpha_\sigma + x_{1it}\beta_{1\sigma} + \mathbf{x}_{2it}\beta_{2\sigma} + g_1(\bar{\mathbf{x}}_i)\xi_{a\sigma} + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_{b\sigma}}{\exp(v(x_{1it}))}\right) \tag{3.18}$$

where $\theta_\sigma = \theta/(\sqrt{1+\sigma_1^2})$ is the scaled coefficient for $\theta = \alpha, \beta_1, \beta_2, \xi_a, \xi_b$. As a result, using a Pooled heteroskedastic Probit approach does not allow us to separately identify $(\alpha, \beta_1, \beta_2, \xi_a, \xi_b)$ and $\sqrt{1+\sigma_1}$. If we focus on the APEs (to be defined formally in the subsequent section), we only need the estimates of the scaled coefficient. The function $v(x_{1it})$ does not include a constant as a necessary requirement for identification in heteroskedastic Probit. Any constant in $v(x_{1it})$ would be incorporated into the scaling factor. We can approximate $v(x_{1it})$ using a polynomial expansion[4]:

$$\begin{aligned} v(x_{1it}) &= \frac{1}{2}\ln\left(1+\frac{\sigma_2^2}{1+\sigma_1^2}x_{1it}^2\right) \\ &\approx \frac{1}{2}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\left(\frac{\sigma_2^2}{1+\sigma_1^2}x_{1it}^2\right)^n \end{aligned} \tag{3.19}$$

where we have used a Taylor expansion in the second line.

Compared to the ME Probit estimator, the Pooled Heteroskedastic Probit is computationally simple, identified when there are no random effects, and consistent under serial correlation. The drawbacks are having to approximate the function $v(x_{1it})$ while using a preprogrammed command, a loss of efficiency comparative to the JMLE, and not being able to separately identify the scaled parameters.

[4]One could maximize the log-likelihood in equation (3.17) directly without approximating the heteroskedastic function $v(x_{1it})$. However, to allow for easy implementation by using preprogrammed commands such as `hetprobit` in STATA, $v(x_{1it})$ can be well approximated by $x_{1it}\delta_1 + x_{it}2\delta_2 + x_{1it}^3\delta_3 + x_{1it}^4\delta_4$.

3.4 Average Partial Effects

As discussed in our introduction, a more meaningful statistic in our model of interest are the Average Partial Effects (APE).This section discusses the identification and formulation of the ASF and APEs using the results of Wooldridge (2005). Identification is shown using Lemma 2.2 and then the ASF is calculated using the results of Lemma 2.1. We will then discuss the derivation and interpretation of the APEs and contrast it to the Partial Effects at the Average that is commonly computed in lieu of the APEs.

The set up explained in Section 2 can be seen as an extension to the Probit example given in Wooldridge (2005) allowing for a random coefficient. A consequence of the Mundlack device, the observable random variables (w in his notation) that help identify the unobserved heterogeneity (q in his notation) are the time averages of the covariates, $\bar{\mathbf{x}}_i$. We start from the Average Structural Function (ASF) defined in Blundell and Powell (2004). The ASF defines the structural relationship between the expected outcome and the covariates, averaging out all the unobserved heterogeneity. To obtain identification of the ASF using Lemma 2.2, the following ignorability assumptions must be satisfied. Applied to the notation of our model, the first is an excludeability assumption that requires,

$$E(y_{it}|x_{it}, a_i, b_i, \bar{x_i}) = E(y_{it}|x_{it}, a_i, b_i), \tag{3.20}$$

and the second is a selection on observables assumption that requires,

$$D(a_i, b_i|\bar{x_i}, x_{it}) = D(a_i, b_i|\bar{x_i}), \tag{3.21}$$

where $D(\cdot)$ denotes the distribution. Equation (3.9) satisfies the ignorability assumptions and therefore following Lemma 2.2, the ASF will be identified from $\mu_2(x_{it}, \bar{x}_i) = E(y_{it}|x_{it}, \bar{x_i})$

where,

$$\begin{aligned}
\mu_2(x_{it}, \bar{x}_i) =& E(y_{it}|x_{it}, \bar{x}_i) \\
=& E(1\{a_i + x_{1it}b_{1i} + \mathbf{x}_{2it}\beta_2 + \varepsilon_{it}\} > 0|x_{it}, \bar{x}_i) \\
=& E(1\{\alpha + x_{1o}\beta_1 + \mathbf{x}_{2o}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_b \\
& + u_{1i} + x_{1o}u_{2i} + \varepsilon_{it} > 0\}|x_{it}, \bar{x}_i) \\
=& \Phi\left(\frac{\alpha + x_{1it}\beta_1 + \mathbf{x}_{2it}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1it}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1it}^2\sigma_2^2}}\right). \qquad (3.22)
\end{aligned}$$

The ASF, $E_{(a_i,b_i)}(\mu_1(x_o, (a_i, b_i)))$, can be calculated as,

$$\begin{aligned}
E_{(a_i,b_i)}(\mu_1(x_o, (a_i, b_i))) &= E_{\bar{\mathbf{x}}_i}(\mu_2(x_o, \bar{\mathbf{x}}_i)) \\
&= E_{\bar{\mathbf{x}}_i}\left(\Phi\left(\frac{\alpha + x_{1o}\beta_1 + \mathbf{x}_{2o}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1o}^2\sigma_2^2}}\right)\right)
\end{aligned}$$

using Lemma 2.1. Next we take the partial derivative of the ASF with respect to x_1^o (the variable of interest). Under typical regularity conditions that allow the derivative to pass through the integration, the partial effect with respect to x_{1o} evaluated at the values x_{1o} and x_{2o} takes on the form:

$$\begin{aligned}
PE(x_{1o}, \mathbf{x}_{2o}) =& \frac{\partial E_{\bar{\mathbf{x}}_i}(\mu_2(x_o, \bar{\mathbf{x}}_i))}{\partial x_{1o}} \\
=& \partial E\left(\Phi\left(\frac{\alpha + x_{1o}\beta_1 + \mathbf{x}_{2o}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1o}^2\sigma_2^2}}\right)\right) \Big/ \partial x_{1o} \\
=& E\Bigg(\phi\left(\frac{\alpha + x_{1o}\beta_1 + \mathbf{x}_{2o}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1o}^2\sigma_2^2}}\right) \times \Bigg[\frac{\beta_1 + g_2(\bar{\mathbf{x}}_i)\xi_b}{\sqrt{1 + \sigma_1^2 + x_{1o}^2\sigma_2^2}} \\
& - (x_1^o\sigma_2^2)\frac{(\alpha + x_{1o}\beta_1 + \mathbf{x}_{2o}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_b)}{(1 + \sigma_1^2 + x_{1o}^2\sigma_2^2)^{3/2}}\Bigg]\Bigg) \qquad (3.23)
\end{aligned}$$

Then we define the $APE = E\left(PE(x_{1it}, x_{2it})\right)$, where the inner expectation is with respect to

$\bar{\mathbf{x}}_i$ and then an outer expectation with respect to x_{1it} and x_{2it}. Since we will have estimates for σ_1^2 and σ_2^2 using the ME Probit estimator, the APE can be estimated by plugging in the parameter estimates and then replacing the expectations with the sample averages. Alternatively, following the Pooled Heteroskedastic Probit estimator, in which we are using an exponential function for the heteroskedasticity function and only obtain estimates of the scaled parameters, the partial effect can also be written as,

$$\begin{aligned} PE(x_{1o}, \mathbf{x}_{2o}) =& \frac{\partial E_{\bar{\mathbf{x}}_i}(\mu_2(x_o, \bar{\mathbf{x}}_i))}{\partial x_{1o}} \\ =& \partial E_{\bar{\mathbf{x}}_i}\left(\Phi\left(\frac{\alpha_\sigma + x_{1o}\beta_{1\sigma} + \mathbf{x}_{2o}\beta_{2\sigma} + g_1(\bar{\mathbf{x}}_i)\xi_{a\sigma} + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_{b\sigma}}{\exp(v(x_{1o}))}\right)\right) \Big/ \partial x_{1o} \\ =& E_{\bar{\mathbf{x}}_i}\Bigg(\phi\left(\frac{\alpha_\sigma + x_{1o}\beta_{1\sigma} + \mathbf{x}_{2o}\beta_{2\sigma} + g_1(\bar{\mathbf{x}}_i)\xi_{a\sigma} + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_{b\sigma}}{\exp(v(x_{1o}))}\right) \\ & \times \exp(-v(x_{1o}))\Big(\beta_{1\sigma} + g_2(\bar{\mathbf{x}}_i)\xi_{b\sigma} - (\partial v(x_{1o})/\partial x_{1o}) \\ & \times (\alpha_\sigma + x_{1o}\beta_{1\sigma} + \mathbf{x}_{2o}\beta_{2\sigma} + g_1(\bar{\mathbf{x}}_i)\xi_{a\sigma} + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_{b\sigma})\Big)\Bigg). \end{aligned} \tag{3.24}$$

To estimate the above quantity, we replace the scaled coefficients and the heteroskedastic function $v(x_{1o})$ with the Pooled Heteroskedastic Probit estimates.

In this chapter, we advocate the APE calculated from the ASF as the statistic that most appropriately captures the effect of interest. However, the literature also places value on what we will refer to as the Partial Effect at the Average (PEA). In the linear case, the APE and PEA are equivalent whereas in the nonlinear case they can be quite different. The source of their difference follows from the basic principle that expectations cannot pass through nonlinear functions. In our model, the PEA is simply the partial derivative of $E_{\bar{\mathbf{x}}_i}(\mu_1(x_o, (E(a_i|\bar{\mathbf{x}}_i), E(b_i|\bar{\mathbf{x}}_i))))$ where the unobserved heterogeneity are evaluated at their conditional means:

$$PEA(x_{1o}, \mathbf{x}_{2o}) = \frac{\partial E_{\bar{\mathbf{x}}_i}(\mu_1(x_o, (E(a_i|\bar{\mathbf{x}}_i), E(b_i|\bar{\mathbf{x}}_i))))}{\partial x_{1o}}$$

$$
\begin{aligned}
=&E_{\bar{\mathbf{x}}_i}\Big(\phi\left(\alpha + x_{1o}\beta_1 + \mathbf{x}_{2o}\beta_2 + g_1(\bar{\mathbf{x}}_i)\xi_a + x_{1o}g_2(\bar{\mathbf{x}}_i)\xi_b\right) \\
&\times\left(\beta_1 + g_2(\bar{\mathbf{x}}_i)\xi_b\right)\Big).
\end{aligned} \tag{3.25}
$$

Note that the PEA only incorporates the part of the heterogeneity that is correlated with the observables. In fact, there is no distinction between the PEA in our model that allows for heterogeneity in the level and slope and the APE in a model that assumes constant effects but time averages enter the structural function as additional covariates. We argue that using the APEs calculated from equations (3.23) and (3.24) truly capture the heterogeneous effect while the PEA mutes the genuine impact of heterogeneity.

3.5 Simulation

In this section, we investigate the behaviour of the two estimators with simulated data. In particular we are interested in the trade off in the robustness properties of the Pooled Heteroskedastic Probit estimator under serial correlation and the ability of the ME Probit estimator to separately identify the variance components. We consider several variations on the same model described in equation (1).

1. The covariates are iid over i and t and drawn from the following multivariate normal distribution

$$
\begin{pmatrix} x_{1it} \\ x_{2it} \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right). \tag{3.26}
$$

   and the random coefficients are generated as,

$$
\begin{aligned}
a_i &= -0.25 - 0.5\bar{x}_{1i} - 0.25\bar{x}_{1i}^2 - 0.1\bar{x}_{1i}\bar{x}_{2i} + u_{1i} \\
b_i &= 1.25 - 0.5\bar{x}_{1i} - 0.25\bar{x}_{1i}^2 - 0.1\bar{x}_{1i}\bar{x}_{2i} + u_{2i}
\end{aligned} \tag{3.27}
$$

where the random effects u_{1i}, u_{2i} are generated from the following multivariate normal distribution,

$$\begin{pmatrix} u_{1i} \\ \\ u_{2i} \end{pmatrix} \sim N\left(0, \begin{pmatrix} 0.5 & 0 \\ \\ 0 & 0.25 \end{pmatrix}\right). \tag{3.28}$$

2. The variable of interest is iid over i but correlated over t through an AR(1) process

$$\begin{aligned} x_{1i1} &= a_{xi} + e_{1i1} \\ x_{1it} &= 0.5a_{xi} + 0.5x_{1it-1} + e_{1it}, \qquad t = 2, 3, ..., T \end{aligned} \tag{3.29}$$

where $a_{xi} \sim iid\ N(1, 0.2)$ is the persistent individual effect, $e_{1i1} \sim iid\ N(0, 0.2)$ and $e_{1i1} \sim iid\ N(0, 1 - 0.5^2 - 0.2(0.5^2))$ are additional noise terms. Although x_{1it} is not independent over t, it is identically distributed. To induce correlation between the regressors, we let $x_{2i} = 0.7 + 0.3x_{1it} + e_{2it}$, $e_{2it} \sim iid\ N(0, 1 - 0.3^2)$. The correlated random coefficients are generated in the same was as the first DGP described in equations (3.27) - (3.28).

3. As in DGP 1, the covariates are generated iid over i and t and distributed as in equation (3.26). In this DGP, we are interested if using a simple specification for the correlated random coefficients when the random coefficients are generated in a more flexible manner would still result in the correlated random effects specifications performing better than just the random effects specifications that do not allow for any correlation between the random coefficients and the covariates. The random coefficients will be generated with the following equations,

$$\begin{aligned} a_i &= -0.25 - 0.25\bar{x}_{1i}^3 - 0.15\bar{x}_{2i}^4 + u_{1i} \\ b_i &= 1.25 - 0.25\bar{x}_{1i}^3 - 0.15\bar{x}_{2i}^4 + u_{2i} \end{aligned} \tag{3.30}$$

but in estimation, we will only include the polynomial functions of the time averages

up to order 2.

Finally, we vary the cases over the number of time periods observed ($T = 5, 10, 20$) and the level of serial correlation in the unobserved heterogeneity ε_{it} ($\rho = 0, 0.4, 0.8$ in equation (3.10)). We expect the ME Probit estimator to be inconsistent under serial correlation while the Pooled Heteroskedastic Probit estimator to be consistent under serial correlation or independence. In addition, we would expect the ME Probit estimator to be more efficient than the Pooled estimator under serial independence.

We will estimate two specifications for both the ME Probit and Pooled Heteroskedastic Probit estimators. The first specification incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while the second specification assumes that a_i and b_i are random effects that are correlated with the x's through their time averages.

3.5.1 Computational Results

In addition to providing results on estimation consistency, it seems prudent to report results on the computational ease of implementation. All estimation was performed in STATA[5] using the commands `meprobit` and `hetprobit`. Since the JMLE requires numerically integrating out the random effects, we expect the ME Probit estimator to take longer.

Tables G.1–G.3 present the average length of time of the two estimators with several notable features. Although the estimation times may seem short (5 seconds at most), this is reflective of the simple specification of only two covariates and fairly small sample sizes of 300 individuals. As the specifications become more complex and the sample size increases, the time to compute will lengthen. Therefore we will focus on the relative speed between the

[5]STATA is among the most popular software used by researchers in social science. Other software such as Matlab and R also have built-in commands that perform similar functions as those in STATA.

two estimators. As expected, the ME Probit estimator always takes much longer than the Pooled Heteroskedastic Probit estimator almost 6 times as long. In addition, the distribution of the ME Probit times are much more variable, with standard deviations as large as 9.03 (in DGP 1). So it appears that there may be some outliers skewing the distribution to the right.

This confirms our initial expectations that the ME Probit estimator will suffer computationally compared to the pooled method.

3.5.2 Parameter Estimates

The estimates from the ME Probit estimator and Pooled Heteroskedastic Probit estimator are not comparable "as is" because the Pooled Heteroskedastic Probit estimator is only able to estimate the scaled coefficients. Therefore we will first look at the "de-scaled" ME Probit coefficient estimates and then compare the two estimation procedures with the "scaled" coefficient estimates. Recall that we will need to calculate the scaled ME Probit estimates by dividing the coefficient estimates by the estimate of the scale value ($\sqrt{1+\hat{\sigma}_1}$).

Tables G.4 – G.6 present the bias and standard deviation (given in parenthesis) of the de-scaled ME Probit estimates. The first thing to note is that specification (1) performs quite poorly at any given level of serial correlation. Since specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the covariates, this emphasizes the importance of allowing for correlation between the random coefficients and the covariates. In this particular setting, not doing so would result in an interpretation that the regressors x_{1it} has no strong predictive power for the outcome y_{1it}.

Turning to the correct specification that allows for correlated random coefficients, specification (2) in DGP 1 and 2, under no serial correlation ($\rho = 0$), the ME Probit appears

to perform well with very little bias. But in DGP 3 where specification (2) acts as an approximation to a more flexible correlated random coefficients, there appears to be significant amount of bias. But the bias is at a lesser degree than not attempting to control for correlated random coefficients at all. As the level of serial correlation increases, we see an increasingly positive bias on the de-scaled coefficient estimates. This confirms our earlier discussions where we would expect the JMLE procedure to be sensitive to correlation over the time dimension.

We do see a quite loss to efficiency when using specification (2) relative to specification (1). This is because a correlated random coefficient approach requires including several more terms, polynomials of the time averages, that may be strongly collinear with one another. We would expect that as the distribution of the time averages becomes more precise (ie: number of time observations increase), the terms in the correlated random coefficient become more collinear. So even though the standard deviations of specification (1) decreases as the number of time observations increases, we see the standard deviations for specification (2) increase. Finally, as the level of serial correlation increases, there is also an increase in the standard deviation. Therefore using the JMLE procedure, introducing serial correlation results in increasing bias and increasing variance in the de-scaled parameter estimates.

But to compare the two estimation procedures, we will need to look at the scaled parameter estimates. Both the scaled Pooled Heteroskedastic Probit and ME Probit estimates are presented in Tables G.7 – G.9. Again, and with both estimators, there is strong bias with specification (1) that does not allow the random coefficients to be correlated with the x's.

Moreover, as we expected, the bias of the Pooled Heteroskedastic Probit estimator is unaffected by the level of serial correlation. But more surprising is that the bias from the presence of serial correlation of the scaled ME Probit estimates is diminished. In fact the

bias of the scaled ME Probit estimates appears to be approaching the level of bias observed for the Pooled Heteroskedastic Probit estimates.

We also see that the ME Probit scaled coefficient estimates are slightly more efficient even under serial correlation. Of course one should expect a JMLE procedure to be more efficient than a PMLE procedure when the distribution is correctly specified. However, this need not be true when the joint likelihood is miss-specified and the pooled likelihood is still correctly specified. The efficiency of the ME Probit estimator in this simulation should not be mistaken as a general result, however, since it appears to be a fairly stable result across all three DGP, it may be worth investigating a theoretical result.

Therefore there seems to be a bias efficiency trade-off in terms of the scaled parameters estimates. If one were to compare on the basis of Root Mean Squared Error, as in Table G.10, the ME Probit estimator appears superior to the Pooled Heteroskedastic Probit estimator under all of the different sampling scenarios.

Why are the ME Probit estimates of the scaled coefficients performing so well when the de-scaled coefficients are quite poor? The answer lies in the estimation of the scaling factor in which the ME Probit is able to identify and estimate σ_1^2. Figures F.1-F.9 present the empirical distribution of $\hat{\sigma}_1^2$. Recall the true variance is 0.5, so under no serial correlation, the ME Probit estimator does a fair job of estimating the variance component. But as serial correlation increased, the distribution of the variance estimates move towards the right suggesting an upward bias. Since we also see an upward bias in the de-scaled coefficient estimates, the biases cancel each other when calculating the scaled coefficients.

Since ME Probit assumes no serial correlation, it is interpreting the persistence in the latent error as part of the individual heterogeneity. Returning to the latent variable set up

in equation (3.8), the unobserved error that the estimator is trying to parse is,

$$(\text{unobserved error})_{it} = u_{1i} + x_{1it}u_{2i} + \varepsilon_{it} \tag{3.31}$$

The serial correlation in ε_{it} appears a lot like the persistence induced by the additive heterogeneity u_{1i}. Consequently, the estimate of the variance component would be biased up and less precise compared to the case of no serial correlation.

An advantage of the JMLE procedure is that it is able to identify both of the variance components σ_1^2 and σ_2^2. However given this analysis, one should be wary of the validity of the σ_1^2 estimates if there is concern for serial correlation. As for σ_2^2, the results reported in Table G.11 mirror the other coefficient estimates. The de-scaled estimates display increasing bias over the level serial correlation which cancels with the scaling such that the scaled estimates appear unbiased. Incorrectly assuming that there is no serial correlation would lead a researcher to incorrectly conclude that the distribution of the random coefficient is much larger than it truly is. But nevertheless, the bias seems to work in our favor if one is more interested in APE, which we typically are.

As mentioned previously, the scaled coefficients are really what is used to determine APE estimates. The results from the simulation thus far would suggest that both estimation procedures will perform reasonable well given that they both estimate the scaled coefficients with fairly small bias (when they allow for correlated random effects). The question remains if any efficiency gains will be observed when using a miss-specified JMLE over a PMLE. Moreover, we saw that not allowing for correlated random effects (specification (1)) will bias the parameter estimates. Some may hope that the simple specification will still be able to capture an average effect. But from theory, we know that this is unlikely in a non-linear model such as probit because the average does not pass through non-linear functions.

Finally, we saw that poor parameter estimates for DGP 3, where the correlated random coefficient structure was only approximated. If we are only interested in APE, can only an approximation for the random coefficient be sufficient in capturing the correlation structure with the covariates.

3.5.3 Average Partial Effect Estimates

Estimates of the Average Partial Effects with respect to x_{1it} are presented in Tables G.12 – G.14. In line with the results on coefficient estimates, not allowing the random effects to be correlated with the xs (specification (1)) results in significant bias across all three DGPs. When we allow the random effects to be correlated with the xs (specification (2)), the bias in the APE estimates shrink considerably, for both estimation procedure.

In DGP 1 and 2, the ME Probit estimates appear to have smaller bias are more efficient than the Pooled Heteroskedastic Probit estimates. Therefore the bias in the de-scaled coefficient and the bias in the variance component of the scaling factor neutralize each other, resulting in very little bias in the ME Probit estimates for APE over any level of serial correlation. In DGP 3, the Pooled Heteroskedastic Probit estimator tends to have slightly smaller bias but is less efficient than the ME Probit estimator. This suggests that there might be some bias efficiency trade-off, but after a quick examination of the RMSE, the ME probit estimator is preferred.

But what does this mean for a researcher working with empirical data? At first glance, one should trust the Pooled Heteroskedastic Probit estimates over the ME Probit estimates since the specified likelihood is robust to arbitrary correlation over the time dimension. But the simulation results suggest that under a simple correlation structure, such as AR(1), there may be robustness in the scaled parameter estimates and APE using a JMLE procedure

where the joint likelihood is misspecified.

However, it should also be emphasized that similar scaled coefficient and APE estimates between the ME Probit and Pooled Heteroskedastic Probit procedures should not trick a researcher into thinking that there is statistical evidence that the assumptions underlying the JMLE necessarily hold. Consequently, any interpretations based on variance estimates, such as the amount of variation explained by the "random" and "fixed" components should be taken with a hearty amount of skepticism.

Finally, to emphasize the importance of correctly understanding the unobserved heterogeneity and how to incorporate it into the descriptive statistics, we provide calculations of the true PEA as a comparison to the true APE over a single sample. Many researchers, turn to the PEA as a simpler to calculate approximation to the APE. As explained earlier, the PEA plugs in the averages for the unobserved heterogeneity rather than integrating it out. Although quicker to compute, this does not truly reflect the data structure we believe is present.

Table G.15 present the results over all 3 DGPs and increasing time observations. Note that the true APE and PEA should not vary by any serial correlation in the latent error. The PEA systematically over-estimates the effect in comparison to the APE. This can be easily explained by comparing the Partial Effect equations (3.23) and (3.25).

First, the PEA does not incorporate the scaling factor $1/\sqrt{1+\sigma_1^2+x_{1o}^2\sigma_2^2}$. Because of the chain rule, the APE and PEA can be broken down into two similar terms. We refer to the first as the "Probit scaling" which consists of the Standard Normal CDF ($\phi(\cdot)$) evaluated at some point in relation to the covariates x_{1o} and x_{2o}. This term insures that the partial effect diminishes as the covariates x_{1o} and x_{2o} get large in absolute value. It is a consequence of the Probit functional form that bounds the average structural function

between 0 and 1. The second term multiplied by the Probit scaling is the "latent" effect. This is the effect of the random variable of interest to the latent index. In a standard Probit with random coefficients, this is just the coefficient on the random variable of interest. By not incorporating the scaling factor, the PEA diminishes the Probit scaling and enlarges the latent effect. This means that the PEA will be shifted up (since β_1 is positive) and flatter over the support of the x's. Second, the latent effect in the PEA does not include the impact of the part of the heterogeneous effect that is uncorrelated with the x's. This effect varies over the value of x_{1o} and can either enlarge or diminish the latent effect.

The main take away is that any patterns or significant biases caused by serial correlation in the latent error appear to be significantly muted when it comes to computing the APEs. A major contributing factor is the ability of the ME Probit estimator to somewhat preserve the consistency of the scaled coefficient estimates under specifications in which we would otherwise deem the procedure inconsistent. This calls for a theoretical investigation of the possible limitations of the consequences when miss-specifying the joint likelihood under serial correlation.

3.5.4 ASF

Figures F.10 - F.15 provide ASF estimates for DGP 1 and 2 over the relevant values of x_{1i} and fixing x_{2i} at its mean (1). There is very little difference between the two estimation procedures, over the different level of serial correlation or over the number of time observations. This again reiterates that because the ME Probit estimator is able to well estimate the scaled coefficient estimates, statistics such as the APE and ASF that only depend on the scaled coefficient estimates tend to be well estimated.

This simulation study has uncovered a surprising number of results which we will sum-

marize here. First, regardless of the estimation procedure, not specifying correlated random effects (i.e., allowing the random coefficients to be correlated with the covariates) significantly effects the results and interpretations of the results. Second, the Pooled Heteroskedastic Probit estimator has performed quite well in terms of the scaled parameter estimates, the APE estimates, and the ASF estimates. It produces estimates with fairly low bias and is much quicker, running in 15.3% of the time ME Probit takes to run. However, one of the main drawbacks to the pooled approach is that it is unable to identify the variance components and therefore some information is lost using this approach.

Alternatively, the ME Probit estimator is able to identify and estimate the variance components but relies on specifying the whole joint distribution which is generally assumed to be independent over time. We saw that there are biases in the de-scaled coefficient and variance component estimates on serial correlation. Therefore, even if the ME Probit can identify these parameters, interpretation should be taken lightly when one is concerned for the presence of serial correlation. But surprisingly these biases appear to counterbalance when calculating the scaled coefficient. This leads to good estimates of the APE and ASF under miss-specification of the joint likelihood. Finally, there appears to be efficiency gains using the JMLE approach regardless of whether or not the joint likelihood is misspecified. This is somewhat surprising since there are no theoretical results that would suggest efficiency under misspecification.

It should be reiterated that this apparent robustness results could be an artifact of the particular data generating processes considered. But we tried to provide a range of interesting DGP to investigate this results. Although beyond the scope of this chapter, there may be some theoretical result in terms of deriving the bias of the scaled coefficient for the JMLE under relatively simple dependency structures.

The next section examines whether the results found here are accordant with real data. Given these results, we would expect to see the Pooled Heteroskedastic Probit and ME probit approaches to provide similar results in their scaled coefficient estimates and APE estimates even when we find evidence of serial correlation.

3.6 Application

Our application utilizes data from Blattman, Jamison, and Sheridan (2017) (for the remainder of the chapter, referred to as BJS) where they study the effect of Cognitive Behavioral Therapy (CBT) on criminal and violent behavior of men in Liberia. After identifying and approaching potential high risk men, the research team obtained 999 men who agreed to enter the sample. Then treatment was assigned randomly within blocks as described in their accompanying appendix. The three possible treatments were: CBT, cash, and both CBT and cash. CBT works to make the patient aware of their automatic negative or self-destructive thoughts so they may be better able to actively change their behavior. Supplying cash should reduce criminal behavior for budget constrained individuals. BJS provides a more thorough discussion on what mechanisms these interventions may change behavioral and economic outcomes.

The data was collected as a series of 5 surveys. The initial survey provided baseline covariates on the men from the study and was taken prior to treatment. Table G.16 provide a summary of a section of these variables. The remaining four endline surveys were taken after 2 weeks, 5 weeks, 12 months and 13 months.[6] One of the major differences between our

[6]Because the surveys are taken unevenly over time, it would difficult to conclude that the dependency in the latent error follows a simple AR(1) process as we used in simulations. We believe that this would make any similarities found between the two procedures even more convincing that a robustness property may hold.

analysis and the initial work done in BJS, is they average the first two surveys and the last two surveys to construct short-run outcomes and long-run outcomes and calculate the effects separately while we treat it as a panel structure. By doing so they are able to investigate heterogeneity in treatment effect over time while we are more interested in heterogeneity that is correlated with the controls.

Although many different types of outcomes were recorded and analyzed by BJS, we will look at only some of the antisocial behavior outcomes in more detail. They define the antisocial behavior as, "disruptive or harmful acts toward others, such as crime or aggression." We will look at the binary outcomes of selling drugs, being arrest, and engaging in illicit activity.[7] Over all the observations, each outcome occurred on average around 10-13 percent of the population.

The last four variables (antisocial behavior index, perseverance index, reward responsiveness and impulsiveness index) are combinations of survey responses to capture the individuals inclination towards a particular characteristics. All are standardize to 0 mean and variance 1. These will be the dimensions in which we will investigate a heterogeneous treatment effect using a correlated random effects approach. BJS does investigate heterogeneous treatment in their appendix (Table E.7) but uses the endline survey responses to construct the outcome antisocial behavior index and only looks at heterogeneity correlated with the baseline antisocial behavior index and a baseline measure of self-control/patience.

To motivate a heterogeneous treatment in the nonlinear ME Probit and Pooled Heteroskedastic Probit models, Table G.17 provides the OLS estimates of a linear probability CRE model. In this setting, a simple linear analysis should provide fairly good estimates of the treatments effect because of the random assignment of the treatment.

[7] In the survey the respondents are asked if each of these outcomes occurred within the last two weeks.

Sells Drugs – All of the interventions decrease the probability of selling drugs in which the sum of the CBT and cash effects is comparable to the effect of both as a treatment. However the cash intervention is not very large and is also not statistically different from 0. The other two treatments are statistically significant. We also see strong evidence of heterogeneity in treatment over the antisocial behavior index and some evidence of heterogeneity over the perseverance index (not statistically significant). The direction of the heterogeneity suggests that those who initially demonstrate antisocial behavior (i.e., one standard deviation away from the average level of antisocial behavior), tend to have a stronger treatment effect (i.e., treatment effect of both CBT and cash changes from -0.0724 to -0.1432, almost doubling). This is consistent since one would expect CBT to have decreasing returns over the level of antisocial behavior (i.e., those who already display low levels of antisocial behavior do not gain much from CBT whereas those with high levels of antisocial behavior can gain much more). The treatments that include cash have more of an effect if the individual displayed poor perseverance (lower on the perseverance index). A possible explanation is that those with poor perseverance are more likely to have binding budget constraints compared to those with better perseverance.

Arrested – None of the interventions show a statistically significant effect on the probability of arrest. On top of that, the cash intervention led to a slightly positive-but not statistically significant- effect, opposite direction of what one would expect. Even so, there is statistically significant evidence of a heterogeneous effect for the treatment of both CBT and cash over the antisocial behavior index. It is important to note that being arrested is not just a measure of behavior but also a measure of the governments ability to enforce the law. The next outcome looks to isolate the effect on the behavior.

Illicit – All of the treatments are estimated to have the expected negative effect in which

the treatments including CBT have significant effects. Interestingly, the marginal effect of providing cash in addition to therapy is minimal since the treatment effects are estimated to be about the same. The treatment effects are heterogeneous in antisocial behavior for the interventions that include CBT and slightly heterogeneous in perseverance for only both CBT and cash. In BJS, the characteristic of perseverance is sometimes referred to as grit, and measured from the responses of seven questions on "the ability to press on in the face of difficulty" from the GRIT scale (Duckworth and Quinn (2009)). The positive direction of the heterogeneity in perseverance can be interpreted as: for an individual who has more perseverance than the average level by 1 standard deviation has a treatment effect from both cash and CBT of -0.0227 compared to -.0622 (almost a reduction of 2/3). This would suggest that perseverance may be a detriment in trying to change individuals' behaviors and actions.

It appears that overall, both CBT and cash produce stronger effects which may indicate that cash is necessary to loosen the budget constraint such that an individual may change their behavior influenced by the CBT. Moreover, most the heterogeneity seems to be captured by the antisocial behavior or the perseverance indexes. Finally, this panel structure suggests the possibility of serial correlation. Following the suggestion in Wooldridge (2010), we test for serial correlation in the linear model by regressing first differences of y_{it} on its lag and testing if the coefficient is equal to 0.5 (as implied by the case of no serial correlation). We are only able to reject the null hypothesis of no serial correlation for the outcome of physical fights (p-value = 0.7689).

Motivated that there is evidence of a treatment effect and possible heterogeneity in the treatment effect in a simpler linear model, Table G.18-G.20 provides the parameter estimates in the different Probit specifications. First note that the estimates are scaled parameter

estimates and therefore comparable between the different estimation methods. As per usual, the reported standard errors for the PMLE are robust to arbitrary serial correlation. But we also report the JMLE standard errors that are robust to arbitrary serial correlation. This is usually not done, since we assume serial independence for consistent estimation. However, given the results of the earlier simulation, we observed fairly accurate scaled coefficient and APE effects estimates from the JMLE even under serial correlation. Therefore we treat the estimator as if it were a quasi-MLE, knowing the likelihood is misspecified, and adjusting the inference accordingly.

For all three outcomes, the coefficient estimates on the treatments are quite different between the JMLE and PMLE procedures. For instance the cash treatment is estimated to have a negative coefficient for the outcomes of selling drugs and being arrested when estimated using the JMLE procedure but then estimated to have a positive coefficient in the PMLE approach. In the end, this may still have very little impact on the treatment effects estimates since they are also strongly determined by the heteroskedastic coefficients in the Pooled Heteroskedastic Probit model.

An explanation for the stark differences could be because the pooled estimator is quite inefficient with some standard error estimates approaching $4\times$ higher than the JMLE standard errors. Consequently, the JMLE coefficients are more frequently statistically significant from 0 whereas the PMLE coefficients are almost never statistically significant.

On the other hand, the efficiency of the JMLE also comes at a computational cost. The Pooled Heteroskedastic Probit estimator was always able to compute within a couple seconds while the ME Probit estimator was taking as long as 4 hours to compute. This makes bootstrapping for standard error, the common procedure when obtaining standard errors for ATE and ASF, impractical.

Given the evidence in the simulation studies of Section 5, one should not readily trust the variance component estimates in the ME Probit model. However, it is interesting to note that under the outcome of selling drugs and engaging illicit activity, the estimator would suggest that the cash or both the CBT and cash treatment do not have a random effect at all.

As for the CRE specifications, estimates of the coefficients for interaction terms appear more similar in direction, magnitude, and efficiency across the two estimators. For the outcome of selling drugs, the most important dimensions of heterogeneity appear to be antisocial behavior and perseverance, especially for the treatments that include give cash. As for being arrested, there is little heterogeneity in the treatment of CBT, but the treatment of cash is heterogeneous in perseverance and reward. Unlike the results of the linear specification, these results suggest that those with more perseverance are less likely to be arrested after given cash. The reward index compiles responses from eight survey questions to measure "whether [an individual is] motivated by immediate, typically emotional rewards." The results indicate the an individual more motivated by rewards is less likely to respond well to cash treatments in reducing the probability of being arrested. This may be because, without any changes in their behavior prior to the treatment, they were then rewarded with cash which provides positive reinforcement of their bad behavior.

Similar to what we observe in the linear specification, treatment is heterogeneous with respect to antisocial behavior and perseverance for the outcome of engaging in illicit activity. In particular, the treatment of CBT is fairly heterogeneous in antisocial behavior and the treatment of both CBT and cash is heterogeneous in perseverance in the same directions of the linear estimates.

The implications for the ATE can be seen in Table G.21. We find, using the OLS estimates

and ME Probit estimates allowing for correlation between the unobserved heterogeneity and the covariates tends to lower the ATE with very little cost to efficiency. It is more of a mixed bag when we look at the Pooled Heteroskedastic Probit estimator. Again, this may be due to the inefficiency of the estimator. Consequently, we tend to see stronger similarities between the ME Probit and OLS estimates. This reiterates the robustness of APEs using ME Probit seen in the simulation study.

In all outcomes, the strongest treatment among the three is both CBT and cash. For selling drugs, we find a statistically significant effect of both CBT and cash as well as therapy only. Interpreting the ME Probit estimates, both CBT and cash will reduce, on average, the probability of selling drugs in the future by 7.6 percentage points while the treatment of therapy only reduces the probability by 6.4 percentage point. The Pooled Heteroskedastic Probit Estimates differs slightly estimating a 4.9% decrease in probability for both CBT and cash. However, there appears to be a significant jump in the standard error so the difference between the two estimates are not likely to be statistically significant.

We find no statistically significant treatment effects for the outcome of arrested at any conventional levels of significance. For illicit activity we find fairly similar estimates between the ME Probit and Pooled Heteroskedastic Probit ATE.

An interesting result is that for some specifications and outcomes, we find the Pooled Heteroskedastic Probit estimator to be more efficient. This was not seen in our simulation results, were the ME Probit estimator appeared always more efficient (even under a misspecified log likelihood).

Since we saw ample statistical significance in the correlated random effects, Figures F.19 - F.21 show the surface of the treatment effects over relevant values of two characteristics. As discussed previously, the most influential characteristics are antisocial behavior and

perseverance, except in the case of being arrested in which reward has more impact than perseverance. When looking at the outcome for selling drugs, the first thing to note is that at combination of relevant characteristic values, there is a treatment that induces an effect in the desired direction. Moreover this figure tells us that those with low levels of perseverance require the treatment of both therapy and cash whereas those with higher perseverance are better served with just receiving therapy. Finally, both estimation procedures (JMLE and PMLE) produce similar figures with inconsequential differences in interpretation.

Moving to the treatment effects for being arrested, we find that there are areas in which no treatment is able to produce an effect in the desired direction. For those who are relatively better behaved initially and are very responsive to rewards, none of the treatments produce desirable effects. On the other end of the spectrum, a therapy and cash treatment produce a strong effect (i.e., those with antisocial behavior = 2 and reward = -2, the treatment of both therapy and cash reduces the probability of being arrested by approximately 25 to 30 percentage point). Although the broad conclusions are the same, there are small differences between the two estimators. Unlike the JMLE, the PMLE requires much lower values of reward and antisocial behavior to find cash to be the best treatment. Moreover the findings of the JMLE show a slightly larger area in which none of the treatments produce desirable effects compared to the PMLE.

The conclusions for the outcome of engaging in illicit activities are similar to those found in studying the outcome of selling drugs. Those with lower levels of perseverance require both therapy and cash while those with higher perseverance suffice with just therapy.

The conclusions of either only therapy or only cash as the optimal treatments may be unexpected as it suggests that is actually a marginal detriment in providing cash when also providing therapy (or vice versa). A possible explanation for this conclusion is the limitations

of the model specification. We have assumed that the heterogeneous treatment is linear in the individuals characteristics. Therefore the crossing from both therapy and cash to just therapy or just cash as the optimal treatment may be an unsubstantiated consequence of the marginally strong effect of therapy and cash over just therapy or over just cash on the other end of the characteristic spectrum.

A possible solution would be allow for a much more flexible specification of the correlated random effect. Instead of only specifying linear terms in random effect, we could also include higher order terms to capture any nonlinear relationship. However this will increase the dimensions of the parameter space fairly quickly, providing grounds for utilizing high dimensional approaches. For instance, including second order terms for the four characteristics (10 terms) for the intercept and each of the treatments will result in increasing the number of parameters by 40. One could extend the work of Wooldridge and Zhu (Forthcoming) who use a debiased estimator of a L1-penalized pooled probit with correlated random effects (only in the intercept). Unfortunately, to our knowledge, there are not any published commands in common statistical packages such as Stata or MATLAB that allow for either a penalized ME Probit (or any ME Generalized Linear Model) estimator or a penalized Heteroskedastic Probit pooled estimator.

3.7 Discussion

The results from the simulations and application leave some open ended questions that we wish to look at in more depth. First, we are concerned that the robustness of the JMLE when independence over time does not hold may be because we have introduce serial correlation in a fairly simplistic manner. We will examine DGP 1 under AR(2) process. This introduces

a much more complex model of serial correlation rather than merely a perception of more or less persistence.

Second, we are concerned that many researchers are attracted to the JMLE because it is able to identify the variance parameters. As we showed in simulation, the estimates are strongly biased under the presence of serial correlation. But in our simulations we have always assumed the presence of random effects. Alternative in this simulation, we consider what will happen when the coefficients are in fact non-random. We also find that the presence of serial correlation can mislead one to believe there are random effects when there are none. This further illustrates the caution that should be taken when interpreting the variance components from the ME Probit estimator.

Finally, there is a growing interest in utilizing a Logit model as an alternative to a Probit model. Therefore we repeat DGP1 but specify that the latent error is logistically distributed. The analogue of the ME Probit estimator is the ME Logit estimator in which we employ the command `melogit` in STATA. As in the Probit case, the random components are still assumed to be normally distributed and integrated out numerically. However this means there is no good analogue of the pooled approach with correct distribution assumptions since the logistic distribution does not mix well with the normal distribution. Consequently, this section will not focus on the comparison of the JMLE to the PMLE but rather whether the JMLE is itself consistent under serial correlation in terms of the parameter, variance components, and partial effect estimates.

3.7.1 AR(2)

Consider the following AR(2) process in the latent error

$$\varepsilon_{it} = 0.6\varepsilon_{it-1} - 0.3\varepsilon_{it-2} + e_{it} \tag{3.32}$$

where $e_{it} \sim N(0, 1 - 0.6^2 - 0.3^2)$. This means that each error is positively correlated with the first lagged error and then negatively correlated (conditional on the first lag) with the second lagged error. With a simple AR(1) process, serial correlation in the latent error appears similar to individual heterogeneity, and as we found in section 5, does not bias the scaled coefficient or APE estimates. But with a more complex AR(2) process, the correlation over time cannot be as easily mistaken as individual heterogeneity.

Table G.22 present the scaled coefficient estimates. Again, we find no strong bias in the ME Probit estimates even though the joint likelihood is misspecified. We do see an increase in bias as the number of time observation increases, which might suggest that the JMLE starts to waver in its capability of addressing a more complex correlation structure as more observations are present. But this holds true for the Pooled Heteroskedastic Probit estimator as well.

Turning to the APE estimates in Table G.23, both the ME Probit and the Pooled Heteroskedastic Probit estimates have low bias. We find that there are still fairly substantial efficiency gains by utilizing the JMLE even when the joint likelihood is misspecified.

So even when introducing a more complex structure to the serial correlation, we find that the bias in estimating the variance component σ_1^2 fully captures the consequences of the serial correlation in the latent error. Figures F.22 - F.24 show the empirical distribution of the ME Probit estimates for σ_1^2. As one would expect, there is an upward bias since

overall, the AR(2) process in equation (3.32) would induce positive correlation among the time observation.

Overall, this would help to further illustrate a possible robustness to serial correlation in the scaled parameter and ASF/APE estimates under JMLE. This should be theoretically investigated in further studies.

3.7.2 No Random Effects

Since the JMLE seems to be able to address serial correlation through the variance component σ_1^2. It would be interesting to observe what would occur when no random effects are actually present $\sigma_1^2 = \sigma_2^2 = 0$. Tables G.24 and G.25 present the computational results. Since $\sigma_1^2 = \sigma_2^2 = 0$ is at the boundary of the valid parameter space, we would expect the ME Probit estimator to struggle. Table G.24 presents the number of failed convergence of the estimator prior to obtaining 1,000 successes. When there is no serial correlation, there are upwards of 700 failures for the ME Probit estimator, but as serial correlation is introduced, the failures reduce dramatically. This is because the introduction of serial correlation allows for estimates of variance components away from the boundary. Table G.25 reports the estimation times. Now we see a much strong contrast between the Pooled Heteroskedastic Probit estimator and ME Probit estimator where the ME Probit can take up to 22 times longer.

Instead of looking at the estimates of α and β (which follow the trends of all the previous simulations) we will simply note that the APE, in Table G.26, are well estimated regardless of the estimation procedure used or whether a correlated random effects specified. Since there are no random coefficients, there cannot be correlation between the fixed parameters and the covariates. Consequently, specifying correlated random coefficients does not necessarily

hurt the estimators in terms of bias but it does result in a less efficient estimator as it calls for the inclusion of many irrelevant covariates.

We will focus on the estimates of variance components and whether or not standard LR tests are valid in detecting the presence of random coefficients under serial correlation. Tables G.27-G.28 present the average and standard deviations of the predicted variance components from ME Probit under specifications (1) and (2). When there is no serial correlation, both the estimates of σ_1^2 and σ_2^2 are quite close to zero, which is what we would hope for when there is no random coefficients actually present. As the level of serial correlation increases, the variance component σ_2^2 remains low while the estimates for σ_1^2 are increasingly biased up. This means that serial correlation can be miss-interpreted as individual heterogeneity. This leads us to caution any researcher that would like to make inference on the variance component estimates and use them interpretatively. One would hope that the LR test should be able to reject the model of random coefficients in favour of a more simple non-random coefficient Probit model. Table G.29 reports for the rejection rates at the 5% significance level. We find that under no serial correlation the test performs as expected but as the serial correlation increases the rejection rates also increase. In fact, with a correlation coefficient of 0.8, we reject 100% of the simulated samples.

So although the earlier simulation results are able to suggest that the ME Probit estimator is a favourable estimator given that it is more efficient than the pooled approach and there appears to not be much bias under the misspecification of the joint likelihood for scaled coefficients and APE estimates. But when there is in fact no random coefficient, we find the ME Probit estimator struggle computationally compared the Pooled Heteroskedastic Probit estimator. The ME Probit estimator takes much longer to compute and failing to converge at all in many instances. Moreover this simulation re-emphasizes the caution that should be

taken when interpreting the variance components.

3.7.3 Logit

This simulation utilizes a logistic distribution in the latent error compared to the normal distribution used in a Probit. Although it seems to be favoured particularly in applied work, the logistic distribution does not easily incorporate a mixed effects framework. The logistic distribution does not mix well with itself nor the normal distribution. This raises two issues:

1. Assuming that the random coefficients are normally distributed, as is usually done in the Mixed Effects literature, then there is no equivalent pooled approach. Specifically, the unobserved components:

$$u_{1i} + x_{1i}u_{2i} + \varepsilon_{it} \tag{3.33}$$

   are the sum of two normals and a logistic random variable whose distribution is generally unknown. Consequently we cannot evaluate the conditional distribution to obtain the contemporaneous conditional mean of y_{it} as we did in equation (3.16) when all the unobserved components are assumed to be normally distributed.

2. It is unclear how to implement AR(1) process to the logistic errors since the logistic distribution does not mix well with other logistically distributed random variables. We find that how the AR(1) process is implemented will vary greatly in how we can approach the estimation problem. We consider two approaches, with the following autoregressive process of order 1

$$\varepsilon_{it} = \rho\varepsilon_{it-1} + e_{\varepsilon it} \tag{3.34}$$

   we can either aim to insure $\varepsilon_{it}|\varepsilon_{it-1}$ is logistically distributed or that the marginal

distributions of all the time observation are identically logistically distributed.

These two challenges have to be considered when constructing our simulation study. With respect to the first point, in our simulation, we of course use the ME Logit estimator, but we also consider the Pooled Heteroskedastic Probit estimator where imposing normality in the latent error is used as an approximation to, what is usually an unknown latent distribution.

In addressing the second point, we run the simulations on two different implementations of the serial correlation. In the first case, which we will refer to a conditional logistic AR(1) we will generate the process from the following

$$\begin{aligned} \varepsilon_{i1} &\sim logistic(0, \sqrt{3}/\pi) \\ e_{\varepsilon it} &\sim logistic(0, \sqrt{3(1-\rho^2)}/\pi), \text{ for } t = 2, ..., T \end{aligned} \tag{3.35}$$

This means $\varepsilon_{it}|\varepsilon_{it-1}$ is logistically distributed but since the logistic distribution does not mix with itself, the marginal distributions are not identically distributed over t (although they will all have the same standardized first two moments). The second case, which we will refer to as a marginal logistic AR(1), will be generating from the following distributions

$$\begin{aligned} \varepsilon_{i1} &\sim logistic(0, \sqrt{3}/\pi) \\ e_{\varepsilon it} &\sim log\left(\frac{\sin(\rho U \pi)}{\sin(\rho(1-U)\pi)}\right) \text{ where } U \sim Uniform(0,1), \text{ for } t = 2, ..., T \end{aligned} \tag{3.36}$$

as proposed in Sim (1993). Now the marginal distributions will be identically logistically distributed. We feel that this would give more credence to a pooled, although misspecified in distribution, approach. Under no serial correlation ($\rho = 0$) both processes are identical and the ME Logit likelihood is correctly specified.

Tables G.30 and G.31 reports the de-scaled coefficient estimates. Similar to the Probit case, as the level of serial correlation increases, there is an increasingly positive bias for both AR(1) specifications. Tables G.32 and G.33 report the scaled coefficient estimates for

the conditional Logit AR(1) process and the marginal Logit AR(1) process. The ME Logit parameter estimates are scaled by $1/\sqrt{\pi^2/3 + \hat{\sigma}_1^2}$ which should match (at least in terms of scaling to) the Pooled Heteroskedastic Probit scaled coefficient estimates. As we saw in the Probit case, the bias of the ME Logit estimator is countered by bias in the variance component estimate, $\hat{\sigma}_1^2$, which results in unbiased scaled coefficient estimates. In fact, we see the ME Logit estimator is far superior to the Pooled Heteroskedastic Probit estimator in terms of bias and efficiency.

Finally, the APE estimates are presented in Tables G.34 and G.35. As we expected, since the marginal logistic AR(1) process produces identically distributed errors, the Pooled Heteroskedastic estimator performs better compared to the conditional logistic AR(1) data generating process. But in either case the ME Logit estimator has lower bias and is much more efficient than the pooled approach even though we know that the joint likelihood is misspecified.

The results from this simulation suggest that the robustness of the JMLE under serial correlation is not limited to the normal distribution.

3.8 Conclusion

This study has been an comprehensive investigation in the behaviors of PMLE and JMLE for panel random coefficient binary response models under serial correlation. After introducing the two estimators and the context in which they are usually implemented, we explored their potential in a diverse simulation study as well as sought to confirm our results with an application. Consistent with our initial intuition, there are several points that need to be considered when implementing these estimators.

First and foremost, specifying correlated random effects matters enormously whether you are consider a PMLE or JMLE approach. We saw this regardless of the data generating process (DGP1, DGP2, DGP3), the level of serial correlation ($\rho = 0, 0.4, 0.8$), the type of serial correlation (AR(1) vs AR(2)), or the distribution of the latent error (Probit vs Logit). Our biggest concern is that those who implement the Mixed Effects approaches may be swayed by the language in thinking that they are able to easily model the heterogeneity with such a flexible framework without considering potential correlation with the covariates.

Another expected results is that the pooled approach is much quicker to implement which may have more importance when considering much larger datasets with many more covariates. As we saw in our application the difference ranged between seconds for the PMLE approach and hours for the JMLE approach.

More intriguing, we find quite a number of surprising results that should change some of the perceptions of JMLE and PMLE.

JMLE estimates of the scaled coefficient, ASF, and APE appear to be robust to fairly simple specifications of serial correlation even though the presence of serial correlation implies that the joint likelihood is miss-specified. This is because the bias in the de-scale parameter estimates are countered by a bias in the variance component estimate. Consequently, interpretation of the de-scaled parameter estimates and variance components are ill-advised when one is concerned there may be serial correlation in the latent error.

In simulation, we repeatedly found the JMLE to be the efficient estimator even under miss-specification of the likelihood. There is no theoretical grounding as to why a Pooled estimator need not be more efficient than a miss-specified JMLE. In fact, we do see the PMLE become more efficient than the JMLE in the Application but we were unable to reproduce these results in simulations. Therefore the questions of efficiency and under what

settings the PMLE becomes more efficient than the JMLE remains open and requires further investigation.

In our discussion of the case of no random effects, it was surprising to see the large number of failures to converge for the ME Probit under no serial correlation and then the dramatic drop as the level of serial correlation increases. This adds to the computational advantage of the pooled approach. Although it does not exactly model the heterogeneity, it is much more likely to be able to converge when there is no random effects and at a much faster rate.

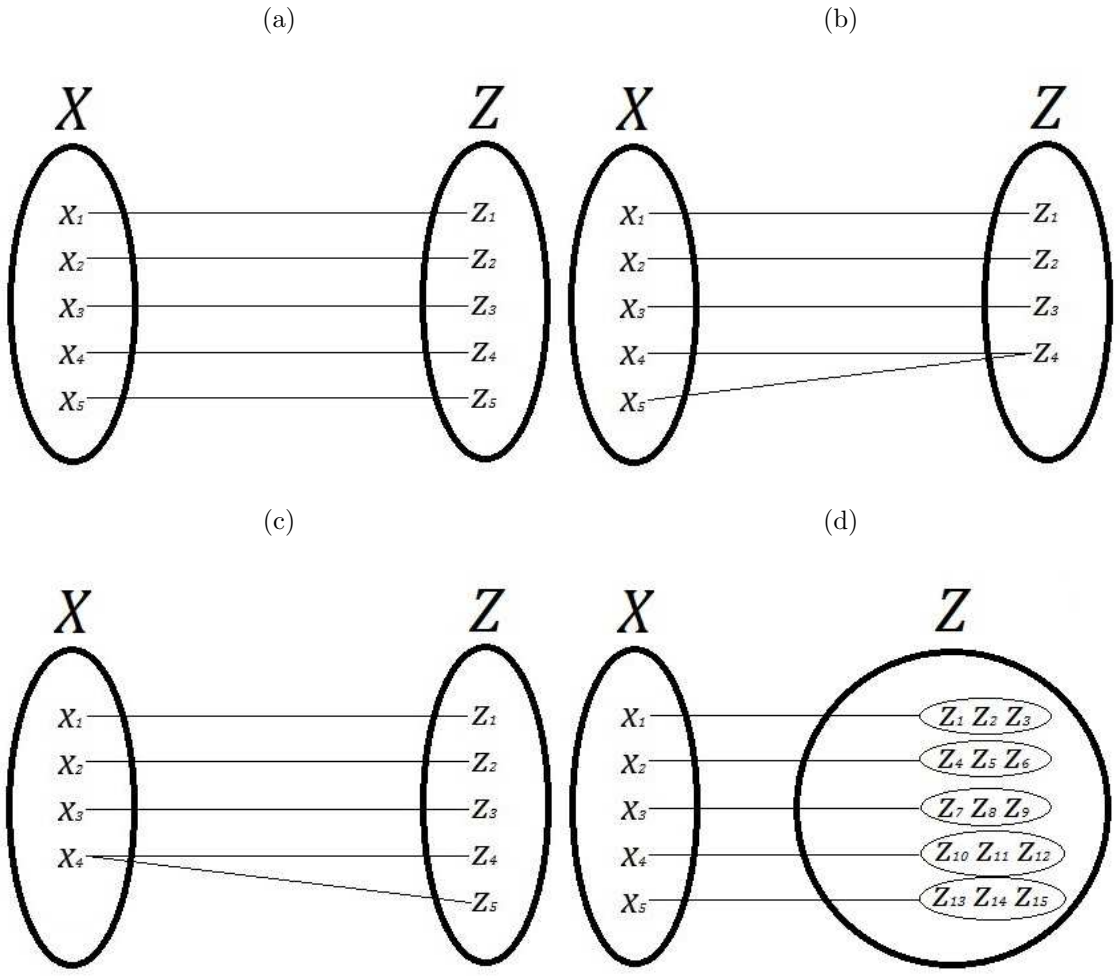
Should we really care about differentiating between serial correlation and random effects? One could argue that they are both ways for an econometrician to model persistence in the data and there is no particular reason to prefer one over the other. As we see in the simulation, this idea appears to be consistent with the robustness of the JMLE under serial correlation. But it would warn against making strict interpretation of the variance components. In the end they are capturing the variability of the persistence over time but this is not necessarily equivalent to the true variance of the random effect.

APPENDICES

APPENDIX A

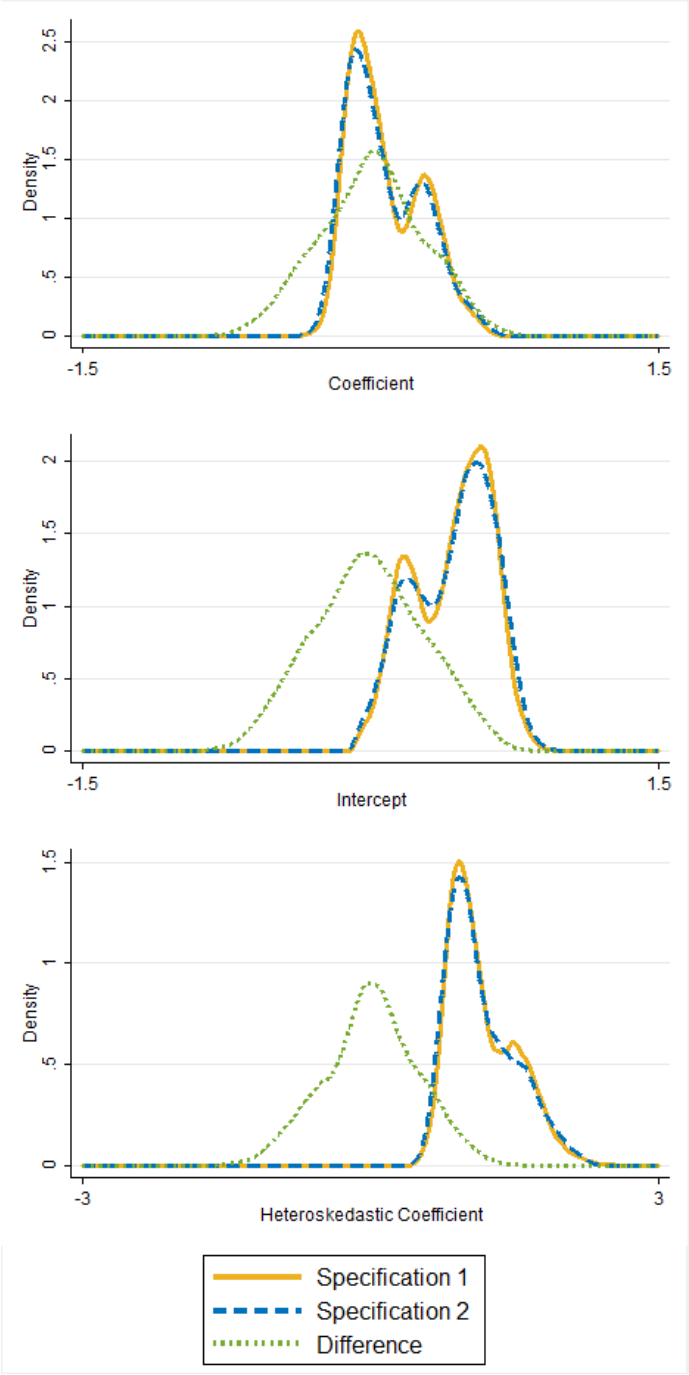
Figures for Chapter 1

Figure A.1: **Visual representation of bijective transformations**



Each oval represent the support of either X or Z, the objects inside represent possible realizations in the support, and the lines connecting the realizations represent pairs of realizations that occur in the joint support with positive probability. From left to right: (a) shows a bijective transformation from X to Z and equivalently a bijective transformation from Z to X. There is no variation in one of the random variables that cannot be perfectly described by the variation in the other. (b), (c), and (d) show examples where there is not a bijective transformation. In (b) there is extra variation in X that cannot be explained by Z and in (c) there is extra variation in Z that cannot be explained by X. The case of an exclusion restriction is presented in (d). Imagine there is an element in Z that is excluded in X that can take on 3 values. Then for every point in the support of X there are 3 possible realizations in Z that will occur with positive probability.

Figure A.2: **Parameter estimates from two observationally equivalent models**



From top to bottom: intercept estimates where the true values are 0 (Specification 1) or 0.5 (Specification 2), coefficient estimates where the true values are 0.5 (specification 1) or 0 (Specification 2), and the heteroskedastic coefficient estimates where the true values are 2 (Specification 1) or 1 (Specification 2). 2,000 simulations of sample size 1,000 using the Stata command `hetprobit`.

APPENDIX B

Proofs and Notation for Chapter 2

Identification

The following lemma is an extension of corollary 1.4.1 given in chapter 1 to allow for multivariate $\mathbf{X}$ and $\mathbf{Z}$.

Lemma B.1. *Let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be vectors of random variables with continuous support. Suppose the following conditions hold*

(i) *$\boldsymbol{Z}$ does not contain a constant.*

(ii) *$E(\boldsymbol{X}'\boldsymbol{X})$ is non-singular.*

(iii) *$E(\boldsymbol{Z}'\boldsymbol{Z})$ is non-singular.*

(iv) *β_o is non-zero*

(v) *Each element of $\boldsymbol{Z}$ is a polynomial function of an element in $\boldsymbol{X}$, such that*

$$Z_j = X_k^{p_j^k}$$

where K is the dimension of $\boldsymbol{X}$ and $p_j^k \in \{1, 2, 3, ...\}$ is the order of the polynomial on the k^{th} term in $\boldsymbol{X}$ that composes the j^{th} term in $\boldsymbol{Z}$.

Then for all parameters $(\beta, \delta) \in \Theta$ (the parameter space) if $\boldsymbol{X}(\beta_o - \exp(Z(\delta - \delta_o))\beta) = 0$ with probability 1, then $(\beta, \delta) = (\beta_o, \delta_o)$.

Proof. Suppose there is a $(\beta, \delta) \in \Theta$, such that

$$\mathbf{X}(\beta_o - \exp(\mathbf{Z}(\delta - \delta_o))\beta) = 0 \tag{B.1}$$

with probability 1. Then I can rearrange given condition (iv),

$$\mathbf{Z}(\delta - \delta_o) = \ln\left(\frac{\mathbf{X}\beta_o}{\mathbf{X}\beta}\right) \tag{B.2}$$

Let A denote the set of k such that the set $\{p_j^l : l = k\}$ is nonempty, then there exists a maximum polynomial order, $\tilde{p}_k = \max_j \{p_j^l : l = k\}$ for each $k \in A$. Then for each $k \in A$, take the partial derivative with respect to X_k, $\tilde{p}_k + 1$ times,

$$0 = (-1)^{\tilde{p}_k+1} \left[\left(\frac{\beta_{ko}}{\mathbf{X}\beta_o} \right)^{\tilde{p}_k+1} - \left(\frac{\beta_k}{\mathbf{X}\beta} \right)^{\tilde{p}_k+1} \right] \tag{B.3}$$

which implies $\frac{\beta_{ko}}{\mathbf{X}\beta_o} = \frac{\beta_k}{\mathbf{X}\beta}$. There are two cases: either for all k such that $\{p_j^l : l = k\}$ is not an empty set, $\beta_k = \beta_{ko} = 0$ or there exists at least one $\hat{k}$ such that $\beta_{\hat{k}} \neq 0$ and $\beta_{\hat{k}o} \neq 0$. In the first case, this reduces to the scenario that $\mathbf{X}$ and $\mathbf{Z}$ are not functionally related in which Theorem 1.2.1 of chapter 1 can be applied to obtain identification. In the second case, equation (B.3) implies $\frac{\beta_{ko}}{\beta_k} = \frac{\mathbf{X}\beta_o}{\mathbf{X}\beta}$ and plugging into equation (B.2),

$$\mathbf{Z}(\delta - \delta_o) = \ln \left(\frac{\beta_{ko}}{\beta_k} \right) \tag{B.4}$$

the right hand side is a constant. By conditions (i) and (iii) equation (B.4) can only hold if $\delta_o - \delta = 0$ and by condition (ii) this implies $\beta_o = \beta$.

Proof of Theorem 2.4.1:

Using Lemma B.1: Parts (i)-(iii) of Assumption 2.4.1 insure that

$$E((\mathbf{x}_i, h(v_{2i}, \mathbf{z}_i))'(\mathbf{x}_i, h(v_{2i}, \mathbf{z}_i))) \tag{B.5}$$

is non-singular (shown in the paper). Part (iv) restricts how the heteroskedastic function may be specified to avoid issues non-identification due to the non-linear setting and correspond to conditions (i), (iii), and (v) of Lemma B.1. Part (v) insures that there is identification of the heteroskedastic components and corresponds to condition (iv) of Lemma B.1. Applying Lemma B.1, identification follows.

Asymptotic for Parametric Estimator

Proof of Theorem 2.5.1:

Using Theorem 2.6 of Newey and McFadden (1994), since there is no weighting matrix, I merely need to show the following:

(i) $E(M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi, \theta)) = 0$ only if $\pi = \pi_o$ and $\theta = \theta_o$

(ii) $\pi_o \in \Pi$ and $\theta_o \in \Theta$, both of which are compact

(iii) $E(M(y_1, y_2, \mathbf{z}; \pi, \theta))$ is continuous at each $\pi \in \Pi$ and each $\theta \in \Theta$

(iv) $E(\sup_{(\pi,\theta)\in\Pi\times\Theta} ||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi, \theta)||) < \infty$

Identification, part (i), holds under Assumption 2.4.1. Part (ii) is assumed. Part (iii) is evident given the linear LS and Probit specifications and part (iv) is satisfied given the finite second moment conditions given in Assumption 2.4.1, see below for more details.

$$
\begin{aligned}
||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi, \theta)|| \leq & ||(y_{2i} - m(\mathbf{z}_i)\pi)m(\mathbf{z}_i)|| + ||S_i(\pi, \beta, \gamma, \delta)|| \\
\leq & ||(y_{2i} - m(\mathbf{z}_i)\pi)m(\mathbf{z}_i)|| + \left[\left| \frac{(y_{1i} - \Phi_i(\pi,\theta))\phi_i(\pi,\theta)}{\Phi_i(\pi,\theta)(1-\Phi_i(\pi,\theta))\exp(g_i\delta)} \right| \right. \\
& \left. \times (||\mathbf{x}_i|| + ||h_i(\pi)|| + ||(x_i\beta + h_i(\pi)\gamma)||||g_i||) \right] \\
\leq & ||(y_{2i} - m(\mathbf{z}_i)\pi)m(\mathbf{z}_i)|| + \left[\frac{\max(\lambda_i(\pi,\beta,\gamma,\delta), \lambda_i(\pi,-\beta,-\gamma,\delta))}{\exp(g_i\delta)} \right. \\
& \left. \times (||\mathbf{x}_i|| + ||h_i(\pi)|| + ||(\mathbf{x}_i\beta + h_i(\pi)\gamma)||||g_i||) \right] \\
\leq & ||(y_{2i} - m(\mathbf{z}_i)\pi)m(\mathbf{z}_i)|| + \left[\frac{1}{\exp(g_i\delta)} C \left(1 + \left| \frac{\mathbf{x}_i\beta + h_i(\pi)\gamma}{\exp(g_i\delta)} \right| \right) \right. \\
& \left. \times (||\mathbf{x}_i|| + ||h_i(\pi)|| + ||(\mathbf{x}_i\beta + h_i(\pi)\gamma)||||g_i||) \right]
\end{aligned}
$$

where $\lambda_i(\pi, \theta)$ is the inverse mills ratio and notationally, $h_i(\pi) = h(y_{2i} - m(\mathbf{z}_i)\pi, \mathbf{z}_i)$ and

$g_i = g(y_{2i}, \mathbf{z}_i)$. Therefore, $E(\sup_{(\pi,\theta)\in\Pi\times\Theta} ||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi, \theta)||)$ is finite as long as the second moments of $m(z_i), \mathbf{x}_i, h_i(\pi),$ and g_i are bounded (which is presumed under Assumption 2.4.1).

Proof of Theorem 2.5.2:

Using Theorem 6.1 of Newey and McFadden (1994), I merely need to show the following:

(i) $\pi_o \in \text{int}(\Pi)$ and $\theta_o \in \text{int}(\Theta)$, both of which are compact

(ii) $M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi, \theta)$ is continuously differentiable in a neighborhood of (π_o, θ_o) with probability approaching one.

(iii) $E(M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \theta_o)) = 0$

(iv) $E(||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \theta_o)||^2)$ is finite;

(v) $E(\sup_{(\pi,\theta)\in\Pi\times\Theta} || \bigtriangledown_{(\pi,\theta)} M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi, \theta)||) < \infty$

(vi) $G'G$ is non-singular

where $G = E(\bigtriangledown_{(\pi,\theta)} M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \theta_o))$. Part (i) is assumed and part (ii) is evident given the linear LS and Probit specifications. Part (iii) holds by Assumption 2.3.1 (correct conditional mean specification in the first stage and Fischer consistency in the second stage).

Part (iv) can be verified

$$
\begin{aligned}
||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \theta_o)||^2 &= ||(y_{2i} - m(\mathbf{z}_i)\pi_o)m(\mathbf{z}_i)||^2 + ||S_i(\pi_o, \beta_o, \gamma_o, \delta_o)||^2 \\
&= ||(y_{2i} - m(\mathbf{z}_i)\pi_o)m(\mathbf{z}_i)||^2 + \left(\frac{(y_{1i} - \Phi_i(\pi_o, \theta_o))^2 \phi_i(\pi_o, \theta_o)^2}{\Phi_i(\pi_o, \theta_o)^2 (1 - \Phi_i(\pi_o, \theta_o))^2 \exp(2 g_i \delta_o)} \right) \\
&\quad \times (||\mathbf{x}_i||^2 + ||h_i(\pi_o)||^2 + ||(\mathbf{x}_i\beta_o + h_i(\pi_o)\gamma_o)||^2 ||g_i||^2)
\end{aligned}
$$

applying law of iterated expectations,

$$
E\Big(||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \theta_o)||^2 \Big| \mathbf{z}_i, y_{2i} \Big)
$$

$$
\begin{aligned}
=&||(y_{2i} - m(\mathbf{z}_i)\pi_o)m(\mathbf{z}_i)||^2 + \left(\frac{E\big((y_{1i} - \Phi_i(\pi_o, \theta_o))^2 | \mathbf{z}_i, y_{2i}\big)\phi_i(\pi_o, \theta_o)^2}{\Phi_i(\pi_o, \theta_o)^2(1 - \Phi_i(\pi_o, \theta_o))^2 \exp(2g_i\delta_o)}\right) \\
&\times (||\mathbf{x}_i||^2 + ||h_i(\pi_o)||^2 + ||(\mathbf{x}_i\beta_o + h_i(\pi_o)\gamma_o)||^2||g_i||^2) \\
=&||(y_{2i} - m(\mathbf{z}_i)\pi_o)m(\mathbf{z}_i)||^2 + \left(\frac{\Phi_i(\pi_o, \theta_o)(1 - \Phi_i(\pi_o, \theta_o))\phi_i(\pi_o, \theta_o)^2}{\Phi_i(\pi_o, \theta_o)^2(1 - \Phi_i(\pi_o, \theta_o))^2 \exp(2g_i\delta_o)}\right) \\
&\times (||\mathbf{x}_i||^2 + ||h_i(\pi_o)||^2 + ||(\mathbf{x}_i\beta_o + h_i(\pi_o)\gamma_o)||^2||g_i||^2) \\
=&||(y_{2i} - m(\mathbf{z}_i)\pi_o)m(\mathbf{z}_i)||^2 + \left(\frac{\lambda_i(\pi_o, \beta_o, \gamma_o, \delta_o)\lambda_i(\pi_o, -\beta_o, -\gamma_o, \delta_o)}{\exp(2g_i\delta_o)}\right) \\
&\times (||\mathbf{x}_i||^2 + ||h_i(\pi_o)||^2 + ||(\mathbf{x}_i\beta_o + h_i(\pi_o)\gamma_o)||^2||g_i||^2
\end{aligned}
$$

Since $\lambda_i(\pi_o, \beta_o, \gamma_o, \delta_o)\lambda_i(\pi_o, -\beta_o, -\gamma_o, \delta_o)$ is bounded (and bounded away from 0), taking an expectation of the above equation, $E(||M(y_{1i}, y_{2i}, \mathbf{z}_i; \pi_o, \theta_o)||^2)$ is finite as long as the second moments of $m(\mathbf{z}_i), \mathbf{x}_i, h_i(\pi),$ and g_i are bounded (which is presumed under Assumption 2.4.1). Part (v) follows from boundedness of the first derivative of the inverse mills ratio and finite second moments of $m(\mathbf{z}_i), \mathbf{x}_i, h_i(\pi),$ and g_i. In showing (vi), let $G = (G_\pi, G_\theta)$ where

$$
G_\pi = \begin{pmatrix} G_{1\pi} \\ G_{2\pi} \end{pmatrix} = \begin{pmatrix} E(m(\mathbf{z}_i)'m(\mathbf{z}_i)) \\ E(\Gamma_i(\pi_o, \theta_o)\omega_i(\pi_o, \theta_o)'m(\mathbf{z}_i)) \end{pmatrix}
$$

$$
G_\theta = \begin{pmatrix} G_{1\theta} \\ G_{2\theta} \end{pmatrix} = \begin{pmatrix} 0 \\ -E\left(\Delta_i(\pi_o, \theta_o)\omega_i(\pi_o, \theta_o)'\omega_i(\pi_o, \theta_o)\right) \end{pmatrix}
$$

and

$$
\begin{aligned}
\Gamma_i(\pi_o, \theta_o) &= (\partial h_i(\pi_o)/\partial v_2)\gamma_o\Delta_i(\pi_o, \theta_o) \\
\Delta_i(\pi_o, \theta_o) &= \frac{\phi_i(\pi_o, \theta_o)^2}{\Phi_i(\pi_o, \theta_o)(1 - \Phi_i(\pi_o, \theta_o))\exp(2g_i\delta_o)} \\
\omega_i(\pi_o, \theta_o) &= \begin{pmatrix} \mathbf{x}_i, & h_i(\pi_o, \theta_o), & -(\mathbf{x}_i\beta + h_i(\pi_o, \theta_o)\gamma_o)g_i(\pi_o, \theta_o) \end{pmatrix}
\end{aligned}
$$

Then

$$G'G = \begin{pmatrix} G'_\pi G_\pi & G'_\pi G_\theta \\ G'_\theta G_\pi & G'_\theta G_\theta \end{pmatrix}$$

$$\begin{aligned} G'_\pi G_\pi =& E(m(\mathbf{z}_i)'m(\mathbf{z}_i))E(m(\mathbf{z}_i)'m(\mathbf{z}_i)) \\ &+ E(\Gamma_i(\pi_o,\theta_o)m(\mathbf{z}_i)'\omega_i(\pi_o,\theta_o))E(\Gamma_i(\pi_o,\theta_o)\omega_i(\pi_o,\theta_o)'m(\mathbf{z}_i)) \\ G'_\pi G_\theta =& - E(\Gamma_i(\pi_o,\theta_o)m(\mathbf{z}_i)'\omega_i(\pi_o,\theta_o))E\left(\Delta_i(\pi_o,\theta_o)\omega_i(\pi_o,\theta_o)'\omega_i(\pi_o,\theta_o)\right) \\ G_\theta G_\theta =& E\left(\Delta_i(\pi_o,\theta_o)\omega_i(\pi_o,\theta_o)'\omega_i(\pi_o,\theta_o)\right)E\left(\Delta_i(\pi_o,\theta_o)\omega_i(\pi_o,\theta_o)'\omega_i(\pi_o,\theta_o)\right) \end{aligned}$$

which is easily non-singular by Assumption 2.4.1.

Identification and Asymptotic for Semi-Parametric Estimator

Proof of Theorem 2.7.1:

Write $u_{1i} = h_o(\mathbf{z}_i, v_{2i}) + \epsilon_i$ where $\text{Med}(\epsilon_i|\mathbf{z}_i, v_{2i}) = 0$. Plugging into equation (2.1) and redefining: $\tilde{x}_i = (\mathbf{x}_i, h_o(\mathbf{z}_i, v_{2i}))$ and $\tilde{\beta}'_o = (\beta'_o, 1)$, one can apply Theorem 2.1 of Khan (2013) to

$$y_i = 1\{\tilde{\mathbf{x}}_i\tilde{\beta}_o + \epsilon \geq 0\} \tag{B.6}$$

and obtain the observational equivalence result.

Proof of Theorem 2.7.2:

Identification of $m_o(\cdot)$ in the first stage is immediate from part (i) of Assumption 2.7.6. For identification of the second stage parameters and functions, suppose there are $\beta_{-k} \in \mathcal{B}$,

$h(\cdot) \in \mathcal{H}$, and $g(\cdot) \in \mathcal{G}$ such that $(\beta_{-k}, h(\cdot), g(\cdot)) \neq (\beta_{-ko}, h_o(\cdot), g_o(\cdot))$ and

$$\frac{\mathbf{x}_{-ki}\beta_{-ko} + x_{ki} + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))} = \frac{\mathbf{x}_{-ki}\beta_{-k} + x_{ki} + h(v_{2i}, \mathbf{z}_i)}{\exp(g(y_{2i}, \mathbf{z}_i))} \tag{B.7}$$

with probability 1. By Assumption 2.7.6 (iii), x_{ki} conditional on $\mathbf{x}_{-ki}$ and $h(v_{2i}, \mathbf{z}_i)$ had density with respect to the Lebesgue measure that is positive on $\Re$ for any $h(\cdot) \in \mathcal{H}$. So for any realization $\mathbf{x}_{-ki}, v_{2i}, z_i$, there exists a x_{ki} such that,

$$\mathbf{x}_{-ki}\beta_{-ko} + x_{ki} + h_o(v_{2i}, z_i) > 0 \text{ and } \mathbf{x}_{-ki}\beta_{-k} + x_{ki} + h(v_{2i}, z_i) < 0 \tag{B.8}$$

and since the scaling by $\exp(g(y_{2i}, z_i))$ is always positive for any $g(\cdot) \in \mathcal{G}$, this is a contradiction. I maintain separate identification of β_o and $h_o(v_{2i}, z_i)$ by parts (ii) and the CMR (part (iv)) of Assumption 2.7.6.

Before providing the remaining proofs from Section 2.7.2, I will briefly outline some notation. Let $\mathbf{a} = (a_1, ..., a_k)$ be a $1 \times k$ vector of non-negative integers, then the $|\mathbf{a}|$-th derivative with respect to a function $f : \Re^k \to \Re$ is defined as,

$$\bigtriangledown^{\mathbf{a}} f(\mathbf{x}) = \frac{\partial^{|\mathbf{a}|}}{\partial x_1^{a1} \cdots \partial x_k^{ak}} \tag{B.9}$$

where $|\mathbf{a}| = \sum_{i=1}^{k} a_i$. For any $s > 0$, let $[s]$ denote the largest integer smaller than s. Define the s-th Holder norm, $||\cdot||_{\Lambda} s$, as

$$||f||_{\Lambda} s = \sum_{|a| \leq [s]} \sup_{\mathbf{x} \in \mathcal{X}} |\bigtriangledown^{\mathbf{a}} f(\mathbf{x})| + \sum_{|a| = [s]} \sup_{\mathbf{x} \neq \bar{\mathbf{x}}} \frac{|\bigtriangledown^{\mathbf{a}} f(\mathbf{x})| - |\bigtriangledown^{\mathbf{a}} f(\bar{\mathbf{x}})|}{||\mathbf{x} - \bar{\mathbf{x}}||^{s-[s]}} \tag{B.10}$$

where $||\cdot||$ denotes the euclidean norm. Define a Holder space with smoothness s as

$$\Lambda^s(\mathcal{X}) = \{f \in C^{s-[s]}(\mathcal{X}) : ||f||_{\Lambda} s < \infty\} \tag{B.11}$$

where $C^r(\mathcal{X})$ is the set of continuous function on $\mathbf{X}$ that have continuous first r-th derivatives. Define a weighted Holder ball with radius c, smoothness s, and weight function $(1+||\cdot||^2)^{-w/2}$

with $w > 0$,

$$\Lambda_c^s(\mathcal{X}, w) = \{f \in \Lambda^s(\mathcal{X}) : ||f(\cdot)(1 + ||\cdot||^2)^{-w/2}||_{\Lambda^s} \leq c < \infty\} \tag{B.12}$$

Finally, define the following two norms

$$||f(\mathbf{x})||_2 = \left(\int_{\mathcal{X}} f(\mathbf{x})^2 \mathrm{d}F_{\mathbf{x}}\right) \tag{B.13}$$

$$||f(\mathbf{x})||_{\infty,w} = \sup_{\mathbf{x}\in\mathcal{X}} \left|f(\mathbf{x})(1 + ||\mathbf{x}||^2)^{-w/2}\right| \tag{B.14}$$

Proof of Corollary 2.7.1:

By the triangle inequality and definition of $||\cdot||_{\infty,w_1}$,

$$\begin{aligned}
&\left|\left|n^{-1}\sum_{i=1}^n \phi\left(\frac{\mathbf{x}_i\hat{\beta} + \hat{h}(\hat{v}_{2i}, \mathbf{z}_i)}{\exp(\hat{g}(y_{2i}, \mathbf{z}_i))}\right)\frac{\hat{\beta}_j}{\exp(\hat{g}(y_{2i}, \mathbf{z}_i))}\right.\right. \\
&\quad \left.\left. -E\left(\phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\right|\right| \\
&\qquad \leq \left|\left|n^{-1}\sum_{i=1}^n \phi\left(\frac{\mathbf{x}_i\hat{\beta} + \hat{h}(\hat{v}_{2i}, \mathbf{z}_i)}{\exp(\hat{g}(y_{2i}, \mathbf{z}_i))}\right)\frac{\hat{\beta}_j}{\exp(\hat{g}(y_{2i}, \mathbf{z}_i))}\right.\right. \\
&\qquad\quad \left.\left. - n^{-1}\sum_{i=1}^n \phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right|\right| \\
&\qquad\quad + \left|\left|n^{-1}\sum_{i=1}^n \phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right.\right. \\
&\qquad\quad \left.\left. - E\left(\phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\right|\right| \\
&\qquad \leq \left|\left|\phi\left(\frac{\mathbf{x}_i\hat{\beta} + \hat{h}(\hat{v}_{2i}, \mathbf{z}_i)}{\exp(\hat{g}(y_{2i}, \mathbf{z}_i))}\right)\frac{\hat{\beta}_j}{\exp(\hat{g}(y_{2i}, \mathbf{z}_i))}\right.\right. \\
&\qquad\quad \left.\left. - \phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right|\right|_{\infty,w_1} \\
&\qquad\quad + \left|\left|n^{-1}\sum_{i=1}^n \phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right|\right|
\end{aligned}$$

$$- E\left(\phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\Bigg\|$$

and the first term is $o_p(1)$ by the results of Theorem 2.7.3 and the second term is $o_p(1)$ using the Weak Law of Large Numbers, noting

$$Var\left(\phi\left(\frac{\mathbf{x}_i\beta_o + h_o(v_{2i}, \mathbf{z}_i)}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right)\frac{\beta_{jo}}{\exp(g_o(y_{2i}, \mathbf{z}_i))}\right) \tag{B.15}$$

is bounded.

APPENDIX C

Simulation Details for Chapter 2

General Control Function in the Demand for Premium Cable

First I constructed the distribution of markets and operators in which there is one operator in each market (but the operators serve multiple markets). The market ID was assigned using a truncated (from 0 to 172) exponential distribution with mean 40 such that the higher the market id (rounded up to the integer), the smaller the market size.[1] I only allow for two operators, to mimic the competition between Time Warner and AT&T (assigned with equal probability to each market). The product characteristics: number of premium channels offered (z_{11m}) and cost shifter (z_{2m}), are the same within a market. The number of premium channels was drawn from a truncated (0 to 10) $Normal(4.5, 6.25)$. The cost shifter was a function of the quality (number of channels) and the operator (efficient/inefficient operator):

$$cost_m = 10 + 5o_m + 2numch_m + \epsilon_{1m}$$

where $o_m \in \{0, 1\}$ is the operator in the market and $\epsilon_{1m} \sim Normal(0, 1)$ is a market level cost shock. The endogenous variable price is constructed as a function of the number of channels, cost, and unobserved quality (v_{2m}).

$$p_m = 7.5 + 2numch_m + cost_m + v_{2m} \tag{C.1}$$

where v_{2m} is drawn from a $Uniform(-8, 8)$. Then to construct the consumer characteristics (z_{12i}), I draw e_1, e_2 from independent standard normal distributions where age and income

[1] The market identifiers could be constructed any way, this was just so there was a good range in market size.

are constructed as,

$$
\begin{aligned}
age_i &= \lceil 20 + 40\Phi(0.4e_1 - 0.1(e_1^2 - 1) + \sqrt{0.85}e_2) \rceil \\
d_{1i} &= 1\{\Phi(e_1) < .196\} \\
d_{2i} &= 1\{\Phi(e_1) \geq 0.196 \text{ and } \Phi(e_1) < 0.44\} \\
d_{3i} &= 1\{\Phi(e_1) \geq 0.44 \text{ and } \Phi(e_1) < 0.685\} \\
d_{4i} &= 1\{\Phi(e_1) \geq 0.685 \text{ and } \Phi(e_1) < 0.86\} \\
d_{5i} &= 1\{\Phi(e_1) \geq 0.86\} \\
income_i &= \sum_{g=1}^{5} d_{gi} inc_g
\end{aligned}
$$

where e_{2i} is drawn from a

$$
\begin{aligned}
inc_1 &\sim Uniform(10, 25) \\
inc_2 &\sim Uniform(25, 50) \\
inc_3 &\sim Uniform(50, 75) \\
inc_4 &\sim Uniform(75, 100) \\
inc_5 &\sim 20Exp(1) + 100
\end{aligned}
$$

Consequently age and income are positively correlated. The last consumer characteristic, household size, is constructed as the following function of age and income,

$$
hhs_i = \lceil \exp(-0.75 + 0.0015 income_i + 0.03 age_i + \epsilon_{2i} \rceil
$$

where ϵ_{2i} is drawn from a truncated (-1 to 1) $Normal(0, 0.45)$. In violation of CF-CI, the conditional distribution of u_{1i} is,

$$
u_{1i} | z_{11m}, z_{12i}, z_{2m}, v_{2m} \sim N\left(0.32 v_{2m} + 0.15 v_{2m} \times numch_m - 0.02 v_{2m} \times age_i, 1\right) \quad \text{(C.2)}
$$

So there is no heteroskedasticity in the latent error but the unobserved product attributes (advertisement) has an interactive effect with number of channels (the addition of more channels matters more if this was advertised) and age (younger consumers may be more susceptible to advertisement). Finally, the binary dependent variable, y_{1i} is calculated from,

$$\begin{aligned} y_{1i} = 1\{ & -9.8 - 0.14p_m + 0.017p_m d_{2i} + 0.03p_m d_{3i} + .035p_m d_{4i} \\ & + 0.045p_m d_{5i} + 0.01numch_m + 0.005income_i + 0.03hhs_i \\ & + 0.005age_i + 0.006age_1^2 + u_{1i} > 0\} \end{aligned} \tag{C.3}$$

The summary statistics are presented in Table E.1. Similar to the real data in PT, conditional on choosing cable, about 1/3 of the sample selects premium cable.

ASF Estimates for the Effect of Income on Home-ownership

In the construction of the exogenous variables, first z_{11i} (age), z_{21i} and z_{22i}(education of wife) are determined (z_{11i} independent of z_{21i} and z_{22i}, z_{21i} and z_{22i} are mutually exclusive) and then z_{12i} (children in household) and z_{23i} (wife working) are functions of the other exogenous variables to induce correlation. Since the sample consisted on 981 married men aged 30 to 50, z_{11i} is drawn from a truncated $Normal(41.8, 60)$ with a lower bound of 30 and an upper bound of 50. Let e be drawn from a $Uniform(0, 1)$, then the education of the wife was determined by,

$$z_{21i} = \begin{cases} 1 & \text{if } 0.482 < e_1 \leq .897 \\ 0 & \text{if } else \end{cases} \qquad z_{22i} = \begin{cases} 1 & \text{if } e_2 > 0.897 \\ 0 & \text{if } else \end{cases}$$

Since it would seem reasonable for having young children in household be negatively correlated with age, and higher education,z_{12i} is calculated as the following,

$$z_{12i} = 1\{261.2 - 5z_{11i} - 20z_{21i} - 50z_{22i} + \varepsilon_{12i} > 0\} \tag{C.4}$$

where ε_{12i} is drawn from a $Normal(0, 30)$. Since it would seem reasonable that the probability of a wife working is negatively correlated with having young children in the household and age but positively correlated with higher education, z_{23i} is calculated as the following,

$$z_{23i} = 1\{17.6 - .3z_{11i} - 5z_{12i} + 3z_{21i} + 10z_{22i} + \varepsilon_{23i} > 0\} \tag{C.5}$$

where ε_{23i} is drawn from a $N(0, 5)$. The conditional mean of y_{2i} is,

$$y_{2i} = 7.2 + 0.0117z_{11i} + 0.0911z_{12i} + 0.0642z_{21i} + 0.1291z_{22i} + 0.0911z_{23i} + v_{2i};$$

where v_{2i} is drawn from a $N(0, 0.088)$. The linear index is,

$$x_i\beta_o = 3.8y_{2i} + 0.09z_{11i} + z_{12i} \tag{C.6}$$

so the conditional distribution of u_{1i} is only a function of the linear index and v_{2i}

$$u_{1i}|v_{2i}, z_{11i}, z_{12i}, z_{21i}, z_{22i}z_{23i} \sim N\Bigg(- 2v_{2i} - 2v_{2i}x_i\beta_o, \exp\Big(2(0.01(x_i\beta_o)\Big)\Bigg)$$

and the binary dependent variable is calculated from,

$$y_{1i} = 1\{-34 + x_i\beta_o + u_{1i} > 0\} \tag{C.7}$$

Table E.4 present the summary statistics of the simulated data as well as the summary statistics from Rothe (2009) as a comparison.

The SML estimator proposed in Rothe (2009) maximizes the following log likelihood,

$$\hat{\beta}_{SML} = \arg\max_{\beta} \sum_{i=1}^{n} y_{1i} \log\Big(\hat{G}(x_i\beta, \hat{v}_{2i})\Big) + (1 - y_{1i}) \log\Big(1 - \hat{G}(x_i\beta, \hat{v}_{2i})\Big) \tag{C.8}$$

where $\hat{G}(x_i\beta, \hat{v}_{2i}) = \Phi(\sum_{j=1}^{n} K_h([x_j\beta, \hat{v}_{2j}] - [x_i\beta, \hat{v}_{2i}])y_{1j} / \sum_{j=1}^{n} K_h([x_j\beta, \hat{v}_{2j}] - [x_i\beta, \hat{v}_{2i}]))$ and $K_h(\cdot)$ is a bivariate kernel based on bandwidth h and scaled by that bandwidth. In order to eliminate the asymptotic bias, the SML estimator requires the use of higher order kernels. However, Rothe finds, utilizing lower order kernels tend to perform better with finite samples. Therefore I use a first order Gaussian kernel. The normal CDF transformation insures the estimates fall between 0 and 1 and imposes the correct distribution as a transformation. I find that this helps the parameter estimates. In determining optimal bandwidths, Rothe suggests maximizing the above likelihood with respect to both the parameters β and the bandwidths h. I find that this can result in a number of extreme outliers that corrupt the analysis. Therefore I equate the bandwidths to the optimal value (given the distribution is truly normal) as a function of the parameters $h = 1.06\sqrt{Var([x_i\beta, \hat{v}_{2i}])}n^{-1/5}$. This is then plugged into the likelihood and maximize with respect to the parameters β.

The proposed Het Probit (GCF) maximizes the following likelihood

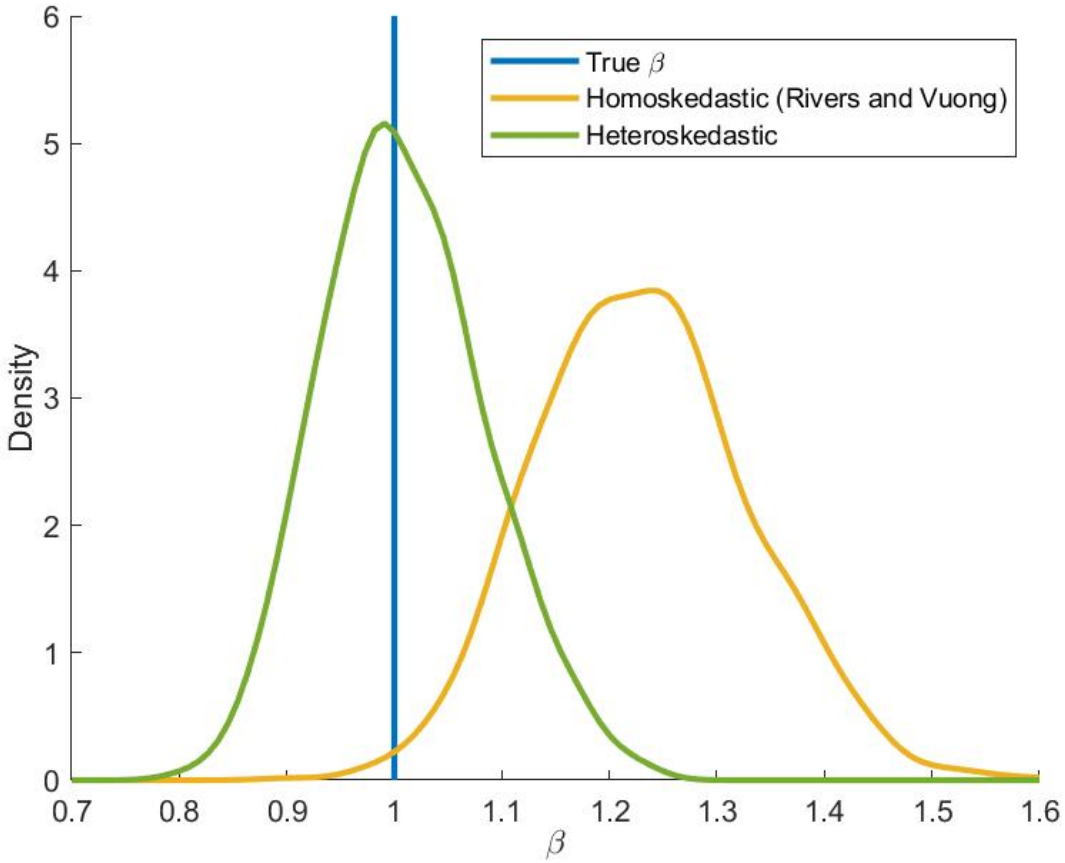
$$\begin{aligned}(\hat{\beta}, \hat{\gamma}, \hat{\delta})_{HetProbit(GCF)} = \arg\max_{\beta,\gamma,\delta} \sum_{i=1}^{n} & y_{1i} \log\left(G(x_i, v_{2i}; \beta, \gamma, \delta)\right) \\ & + (1 - y_{1i}) \log\left(1 - G(x_i, v_{2i}; \beta, \gamma, \delta)\right)\end{aligned} \tag{C.9}$$

where $G(x_i, v_{2i}; \beta, \gamma, \delta) = \Phi\left(\frac{x_i\beta + \hat{v}_{2i}\gamma_1 + \hat{v}_{2i}x_i\gamma_2}{\exp(x_i\delta)}\right)$. I found that the estimates are sensitive to the starting values therefore I used $[1, 0.5, 0.75, 1.5, 2] \times (\beta_o, \gamma_o, \delta_o)$ as the starting values and choose the estimates with the largest log likelihood.

APPENDIX D

Figures for Chapter 2

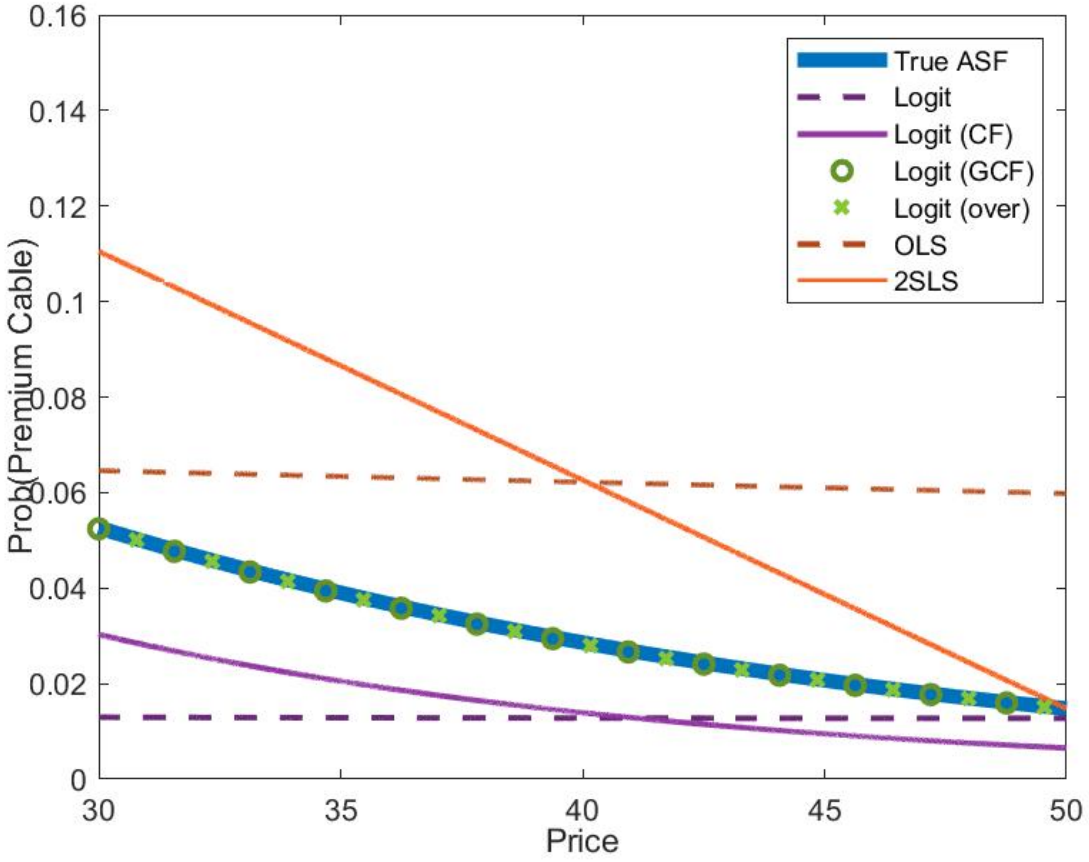
Figure D.1: **Effect of Heteroskedasticity on Parameter Estimate**



Shows empirical distribution for estimates of β in the model $y_{1i} = 1\{y_{2i}\beta + u_{1i} > 0\}$ and $y_{2i} = z_i + v_{2i}$. The unobserved heterogeneity are generated from a heteroskedastic bivariate normal as in equation (2.6) where $\rho(z_i) = 0.6, \sigma_1(z_i) = \sigma_2(z_i) = \exp(0.25z_i)$. The Homoskedastic estimator assumes the data generating process in equation (2.2) while the heteroskedastic estimator correctly scales by the true conditional variance. Calculated from 1,000 simulations of sample size 1,000.

General Control Function in the Demand for Premium Cable

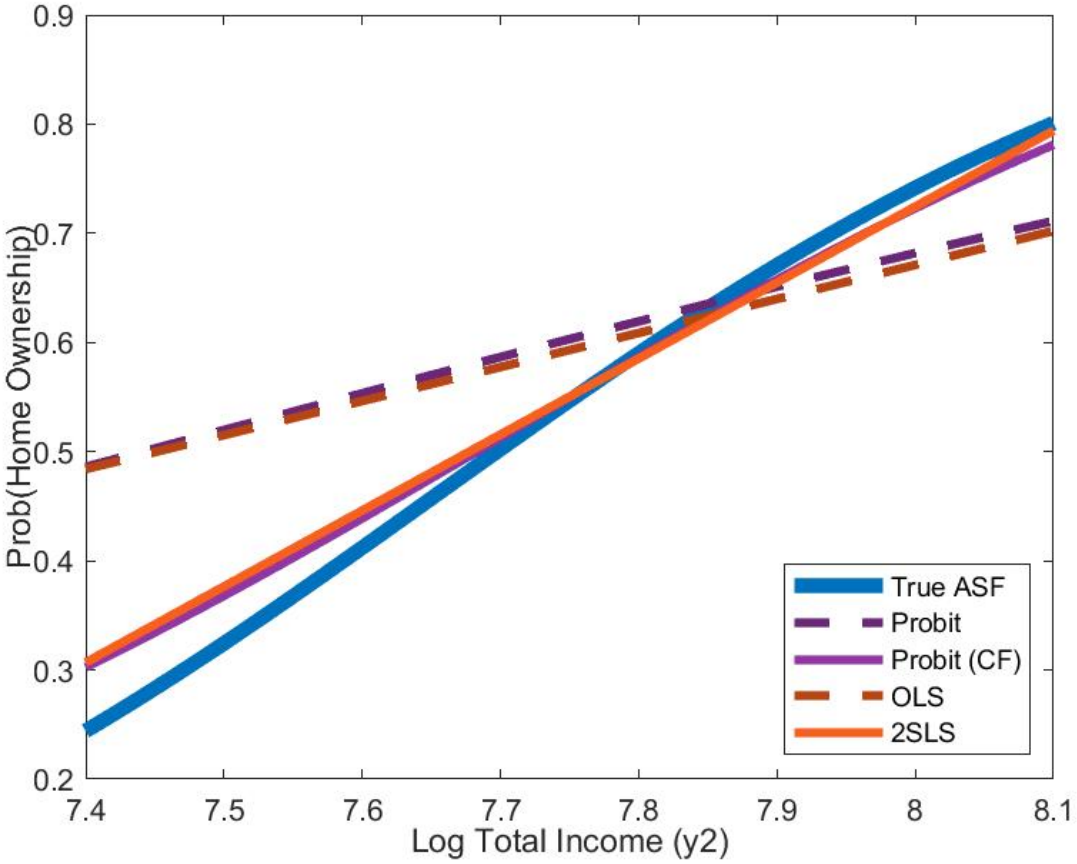
Figure D.2: **ASF for Income equal to $85,000**



ASF is evaluated over different prices for an additional 5 channels of premium cable offered to a consumer who is 35 years old in a family of 3 with income equal to $85,000.

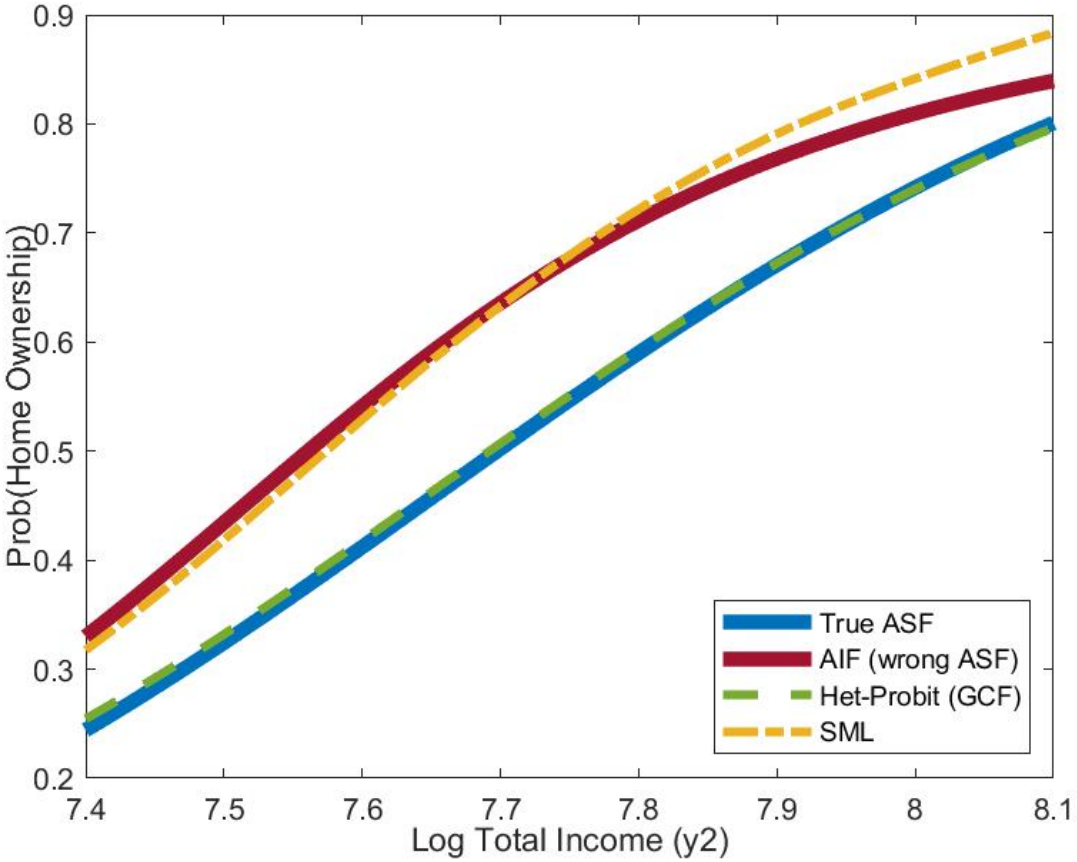
ASF Estimates for the Effect of Income on Home-ownership

Figure D.3: **ASF Estimates for Misspecified Models**



ASF evaluated over different levels of log(total income) for a 40 year old with children under the age of 16 in the household.

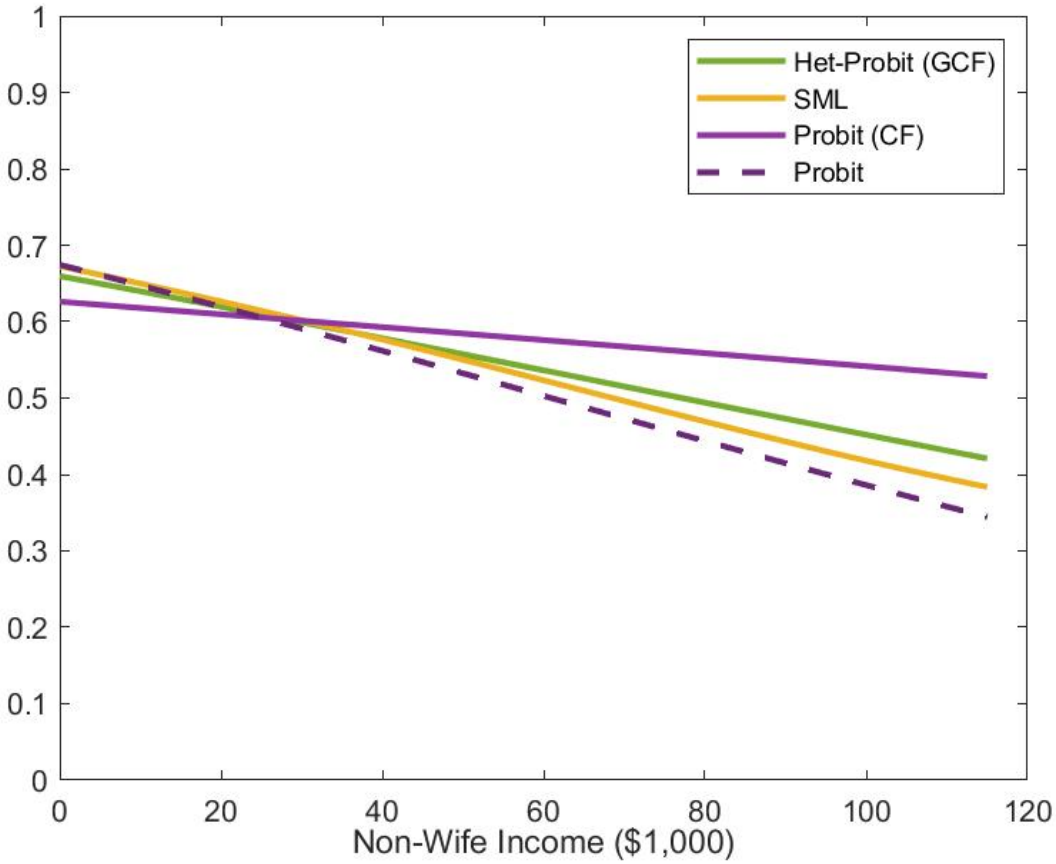
Figure D.4: **Consequence of CF-LI Assumption on ASF Estimates**



ASF evaluated over different levels of log(total income) for a 40 year old with children under the age of 16 in the household.

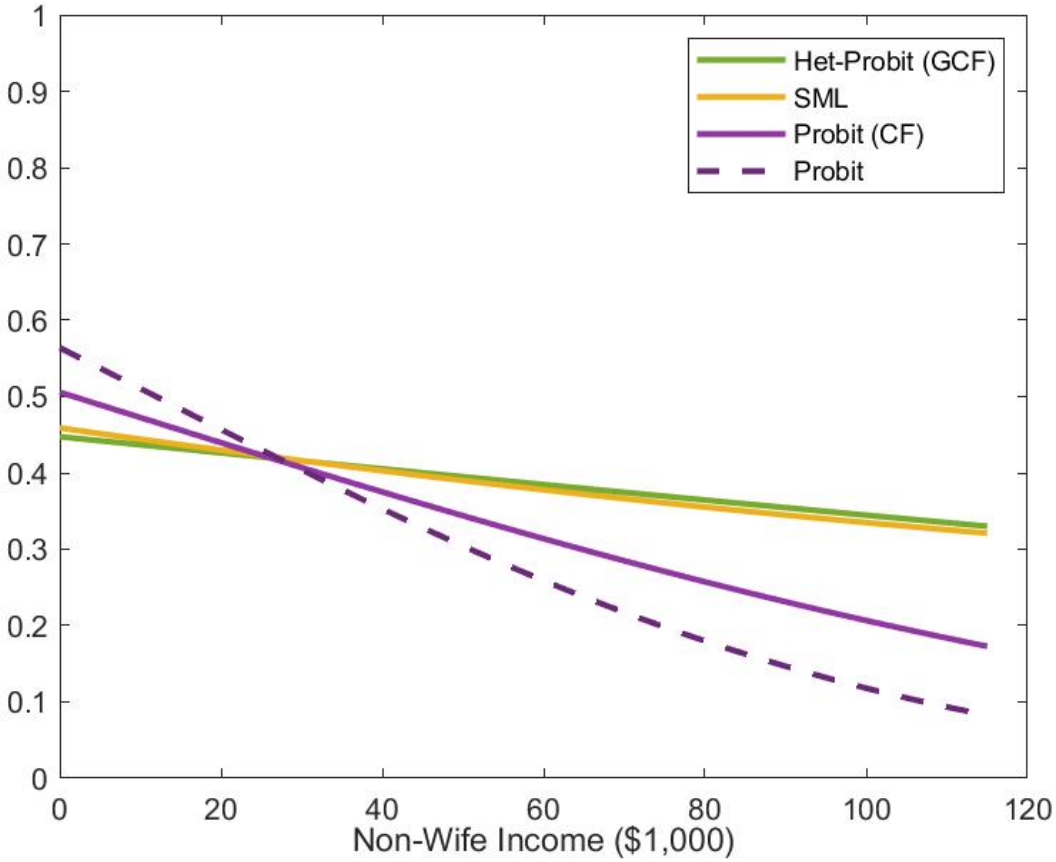
Empirical Example

Figure D.5: **Comparison of ASF for Families with No Children**



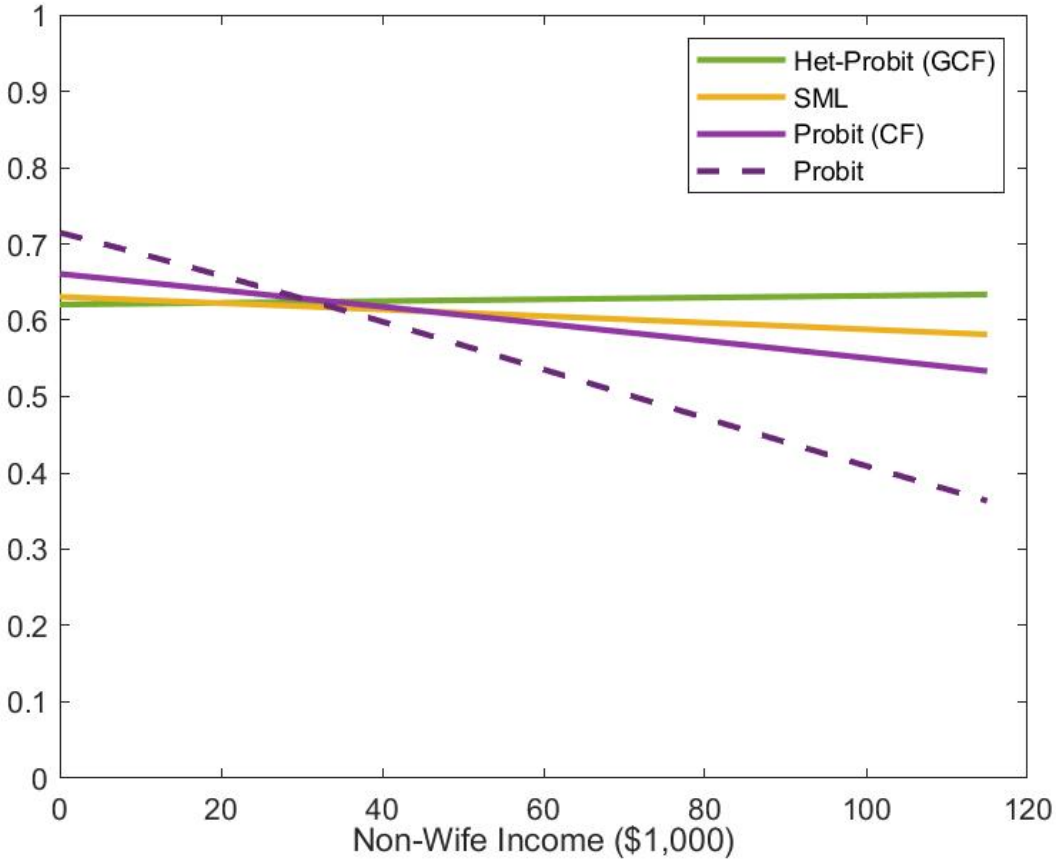
1991 CPS data on Married Women Labor force participation. ASF evaluated over difference levels of Non-Wife Income for family with no children in the household.

Figure D.6: **Comparison of ASF for Families with Young Children Only**



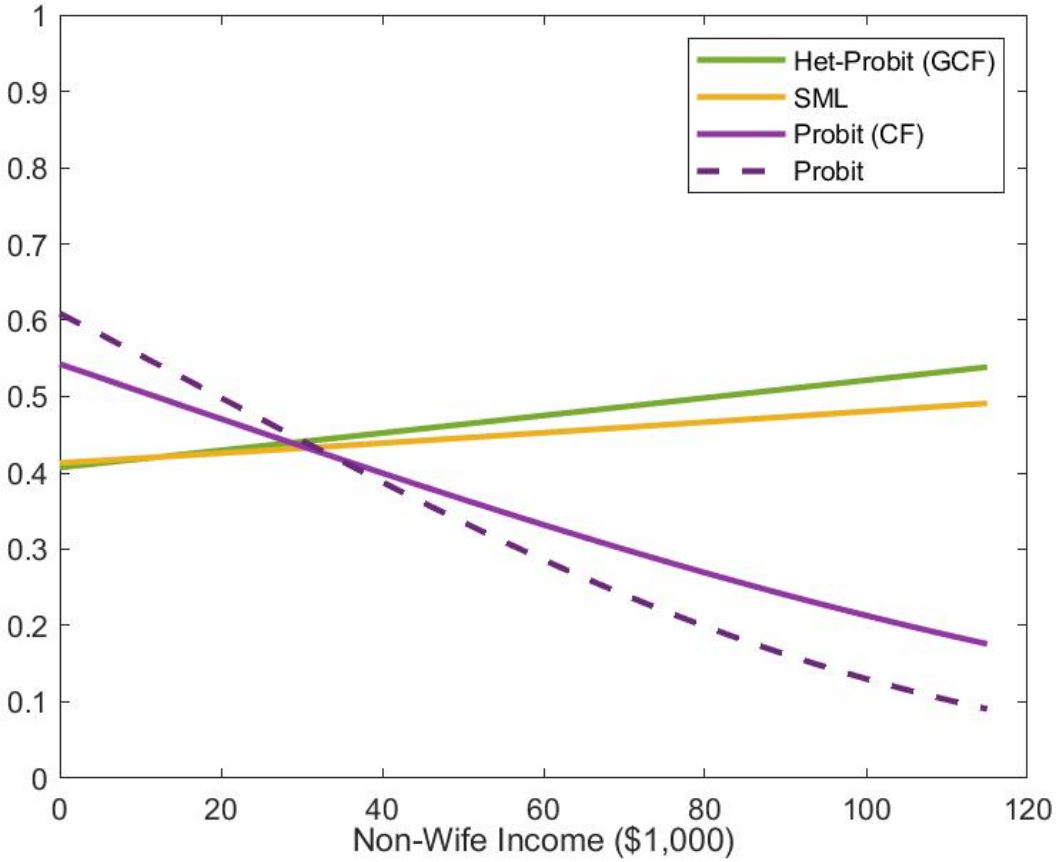
1991 CPS data on Married Women Labor force participation. ASF evaluated over difference levels of Non-Wife Income for family with only young (under 6) children in the household.

Figure D.7: **Comparison of ASF for Families with Old Children Only**



1991 CPS data on Married Women Labor force participation. ASF evaluated over difference levels of Non-Wife Income for family with only old (over 6) children in the household.

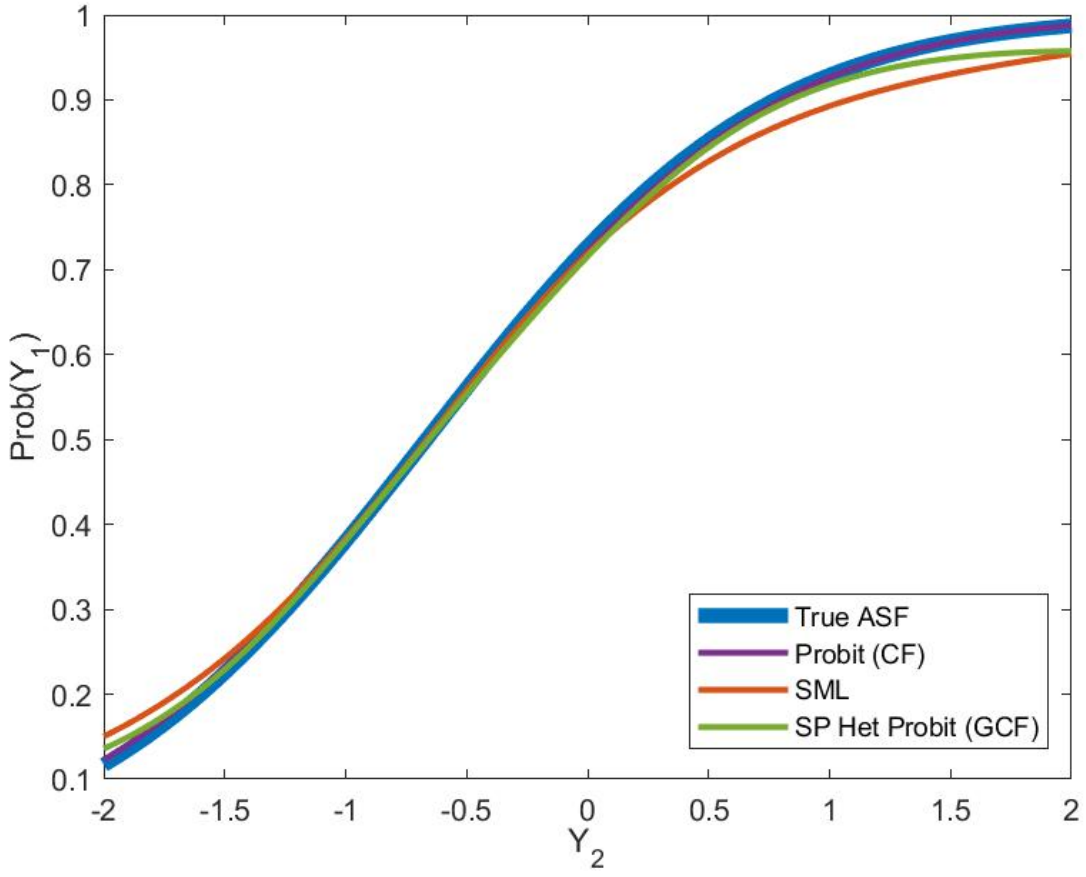
Figure D.8: **Comparison of ASF for Families with Both Young and Old Children**



1991 CPS data on Married Women Labor force participation. ASF evaluated over difference levels of Non-Wife Income for family with both old (over 6) and young (under 6) children in the household.

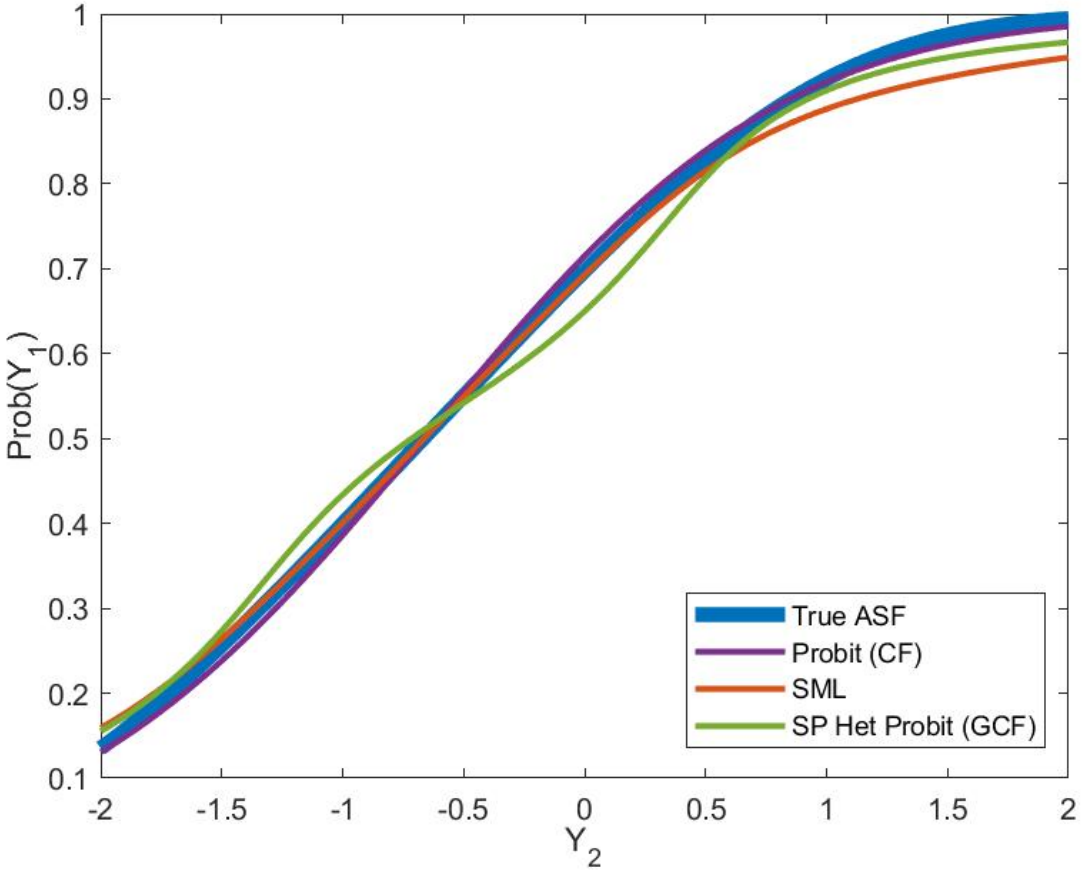
Extension: Semi-Parametric Distribution Free Estimator

Figure D.9: **Logistic Distribution** ($h_o^1 = v_{2i}$)



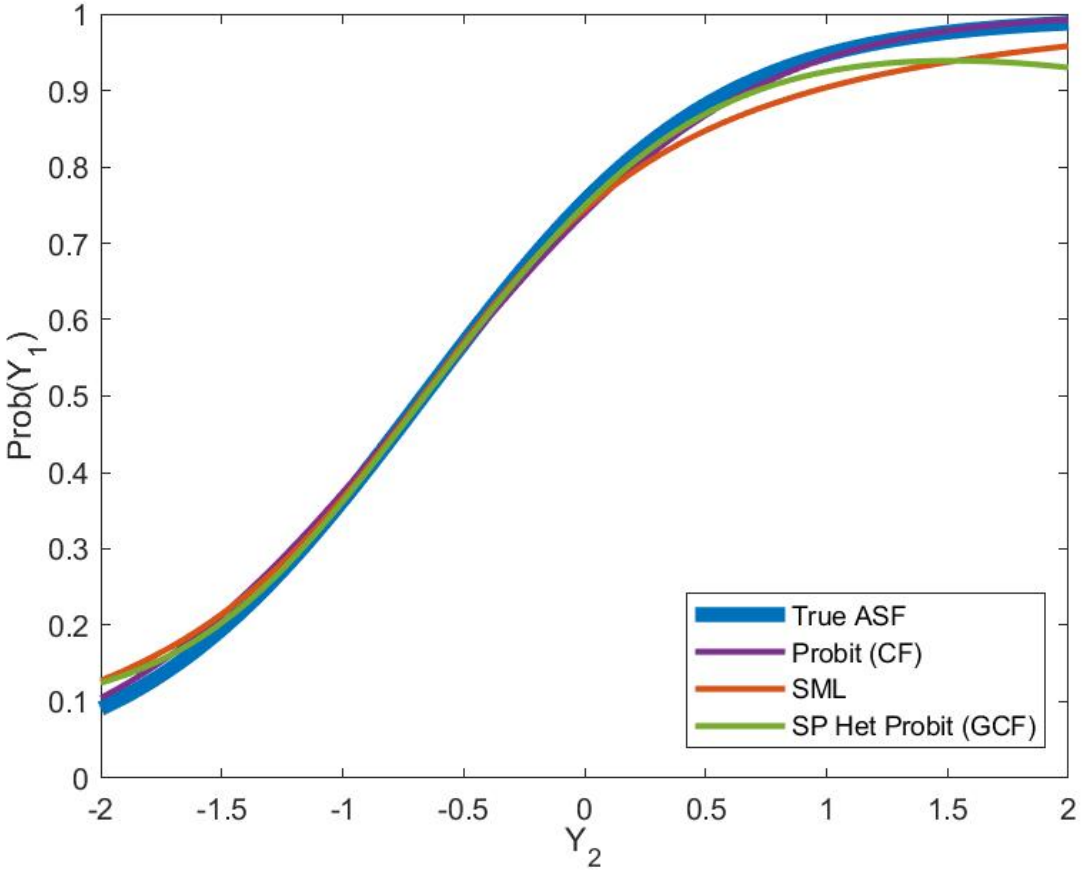
1,000 simulations of Sample size 1,000.

Figure D.10: **Uniform Distribution** ($h_o^1 = v_{2i}$)



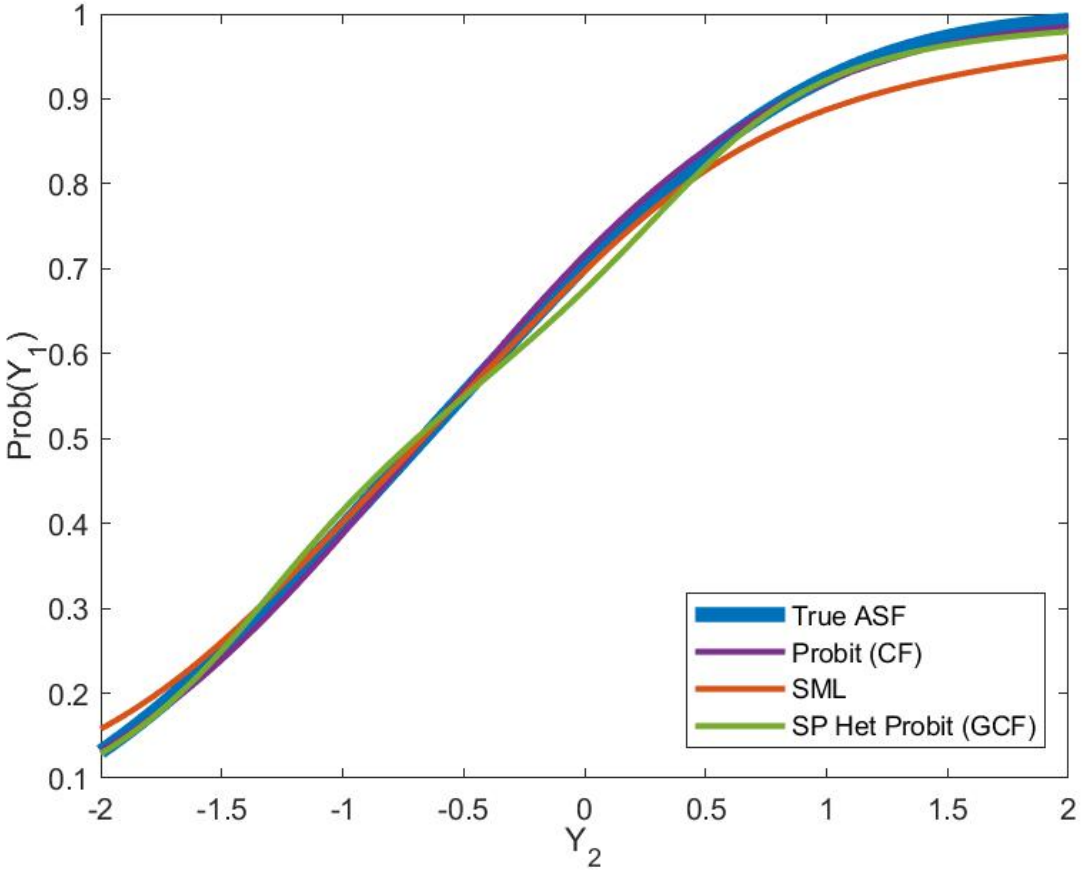
1,000 simulations of Sample size 1,000.

Figure D.11: **Student T Distribution** ($h_o^1 = v_{2i}$)



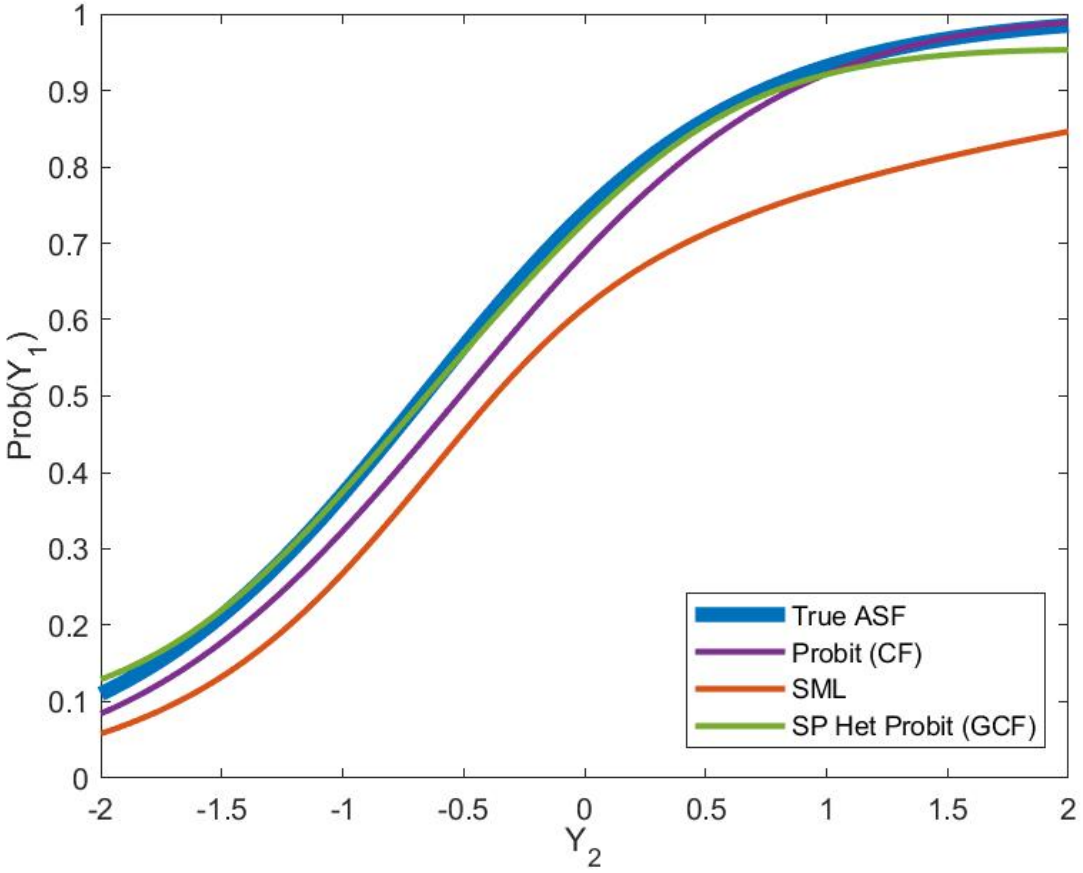
1,000 simulations of Sample size 1,000.

Figure D.12: **Gaussian Mixture Distribution** ($h_o^1 = v_{2i}$)



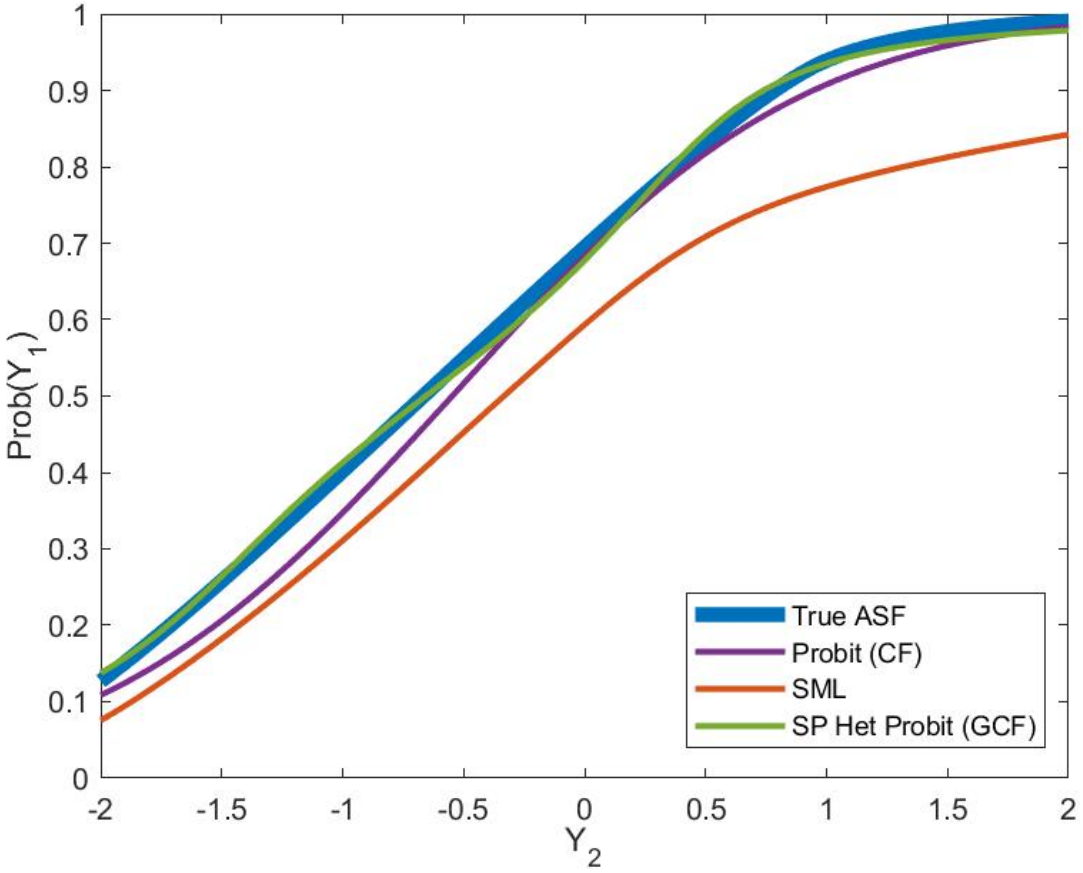
1,000 simulations of Sample size 1,000.

Figure D.13: **Logistic Distribution with Linear GCF (h_o^2)**



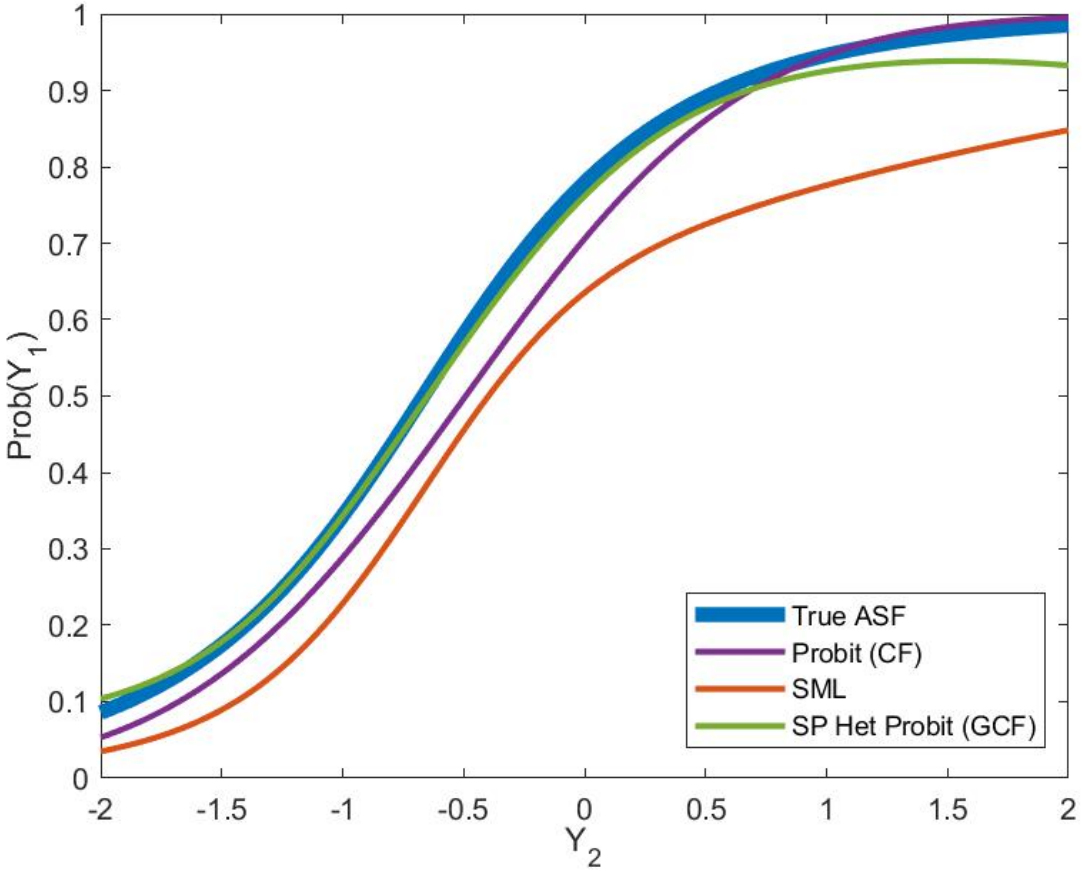
1,000 simulations of Sample size 1,000.

Figure D.14: **Uniform Distribution with Linear GCF (h_o^2)**



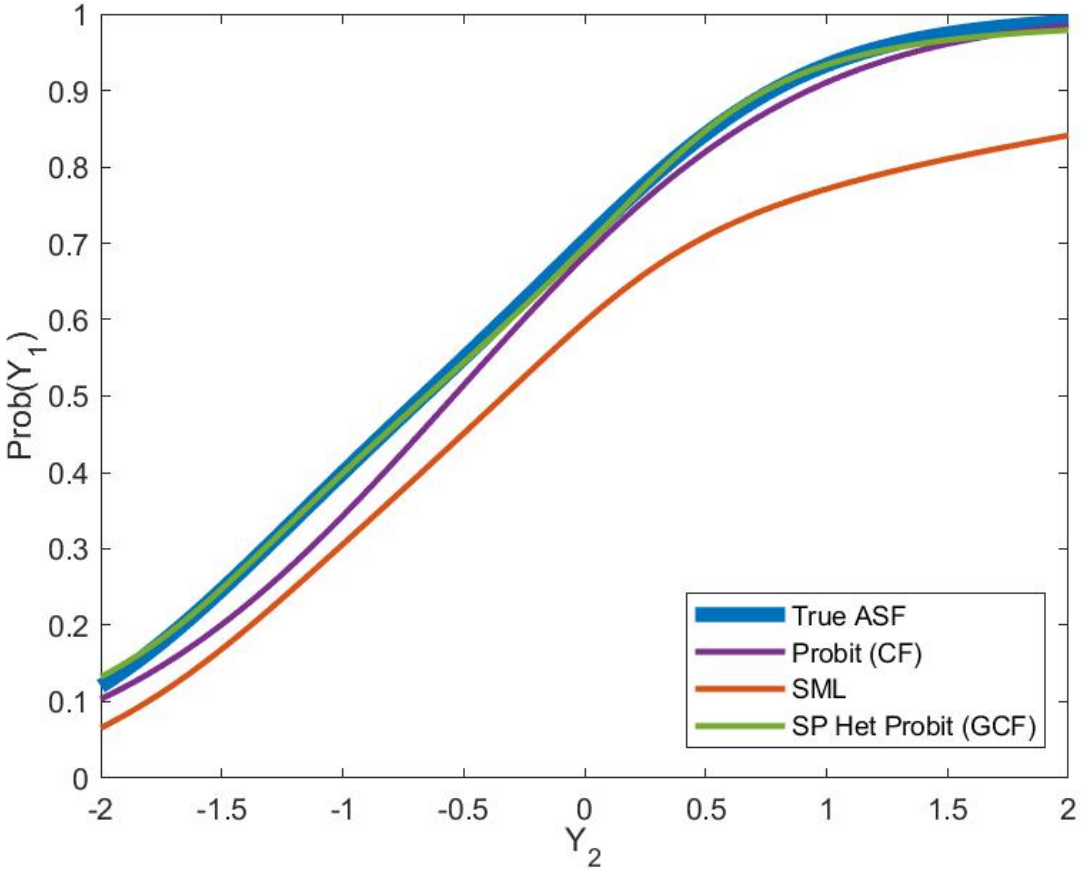
1,000 simulations of Sample size 1,000.

Figure D.15: **Student T Distribution with Linear GCF (h_o^2)**



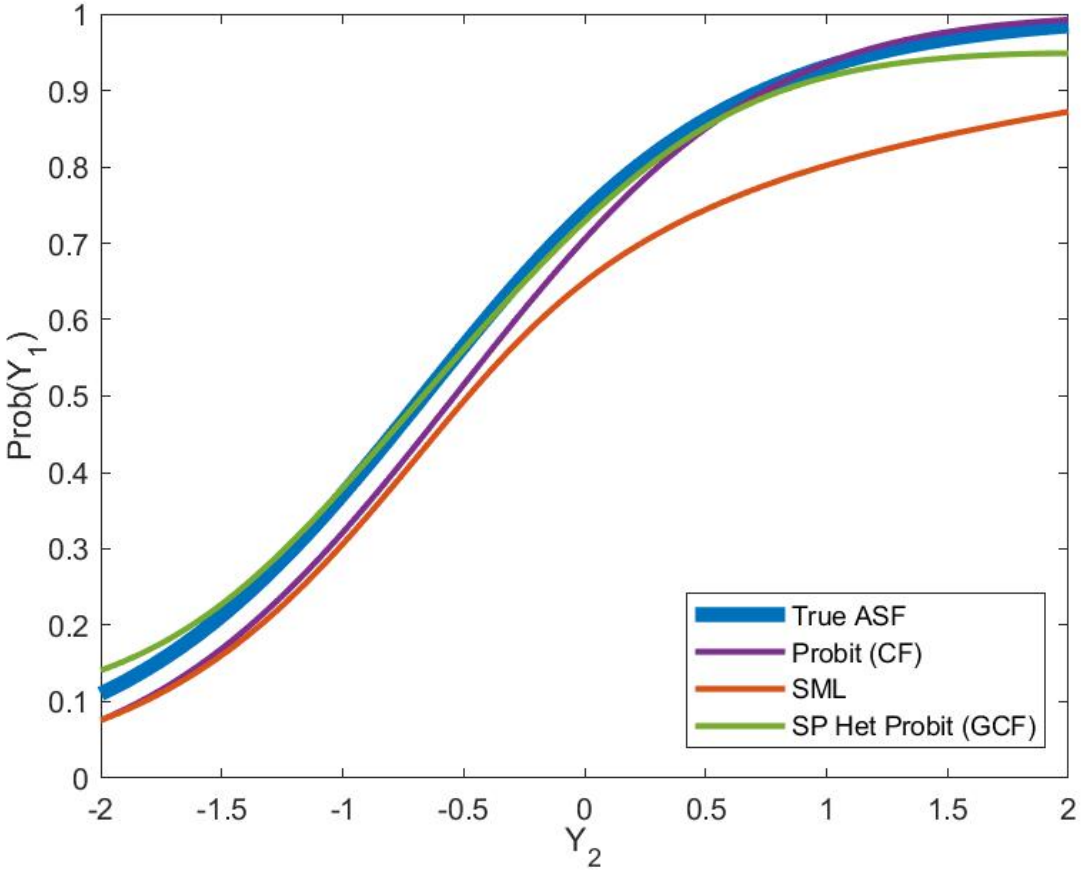
1,000 simulations of Sample size 1,000.

Figure D.16: **Gaussian Mixture Distribution with Linear GCF** (h_o^2)



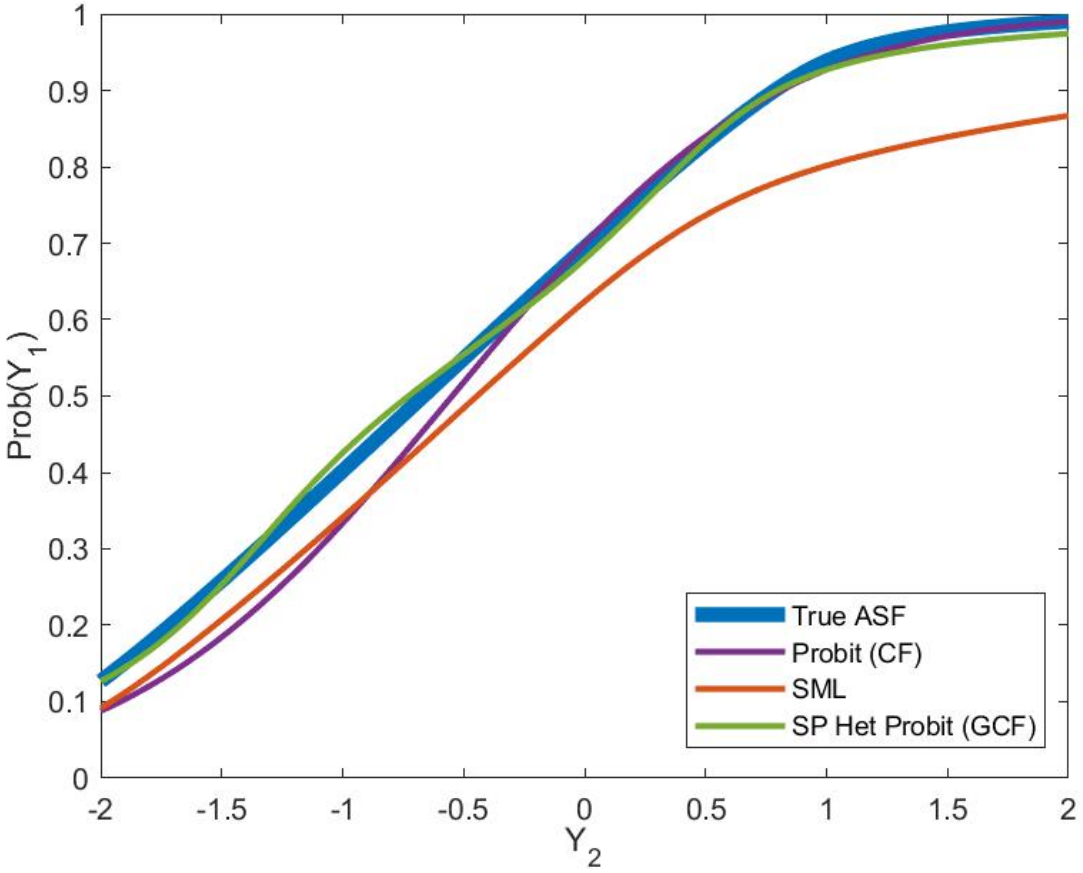
1,000 simulations of Sample size 1,000.

Figure D.17: **Logistic with Non-Parametric GCF** (h_o^3)



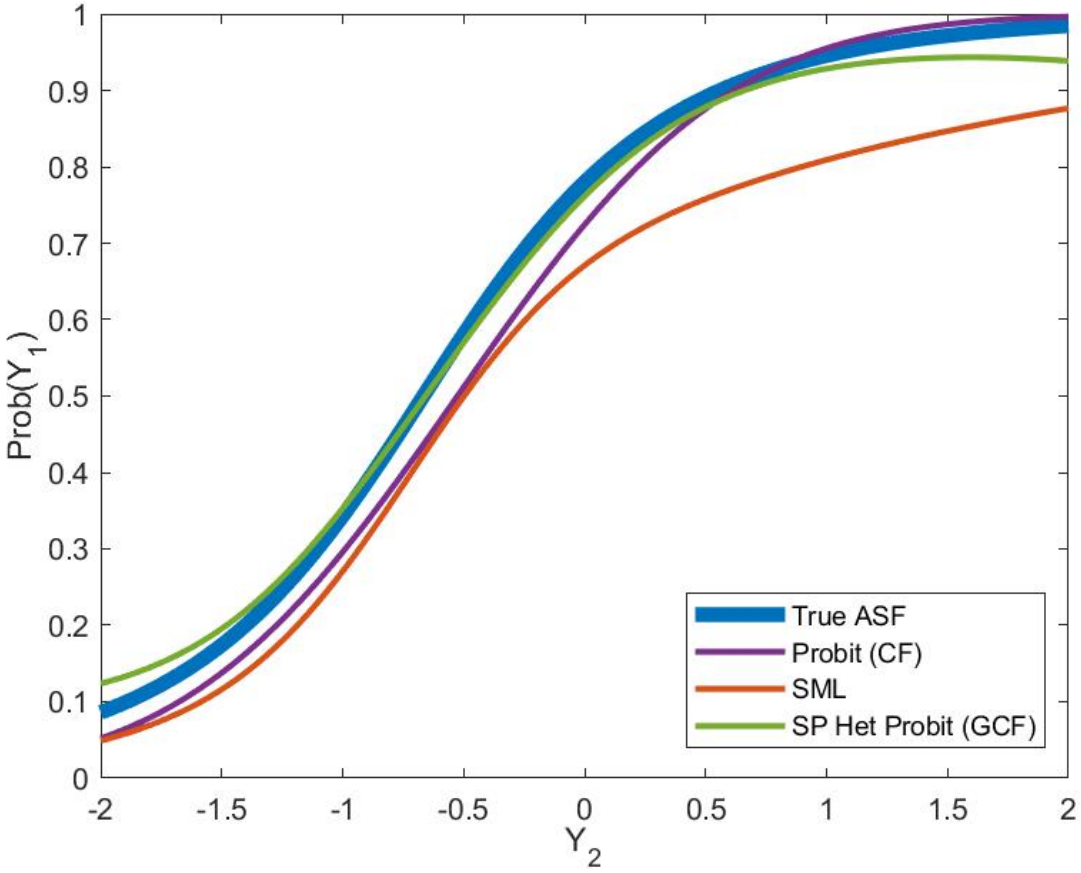
1,000 simulations of Sample size 1,000.

Figure D.18: **Uniform with Non-Parametric GCF** (h_o^3)



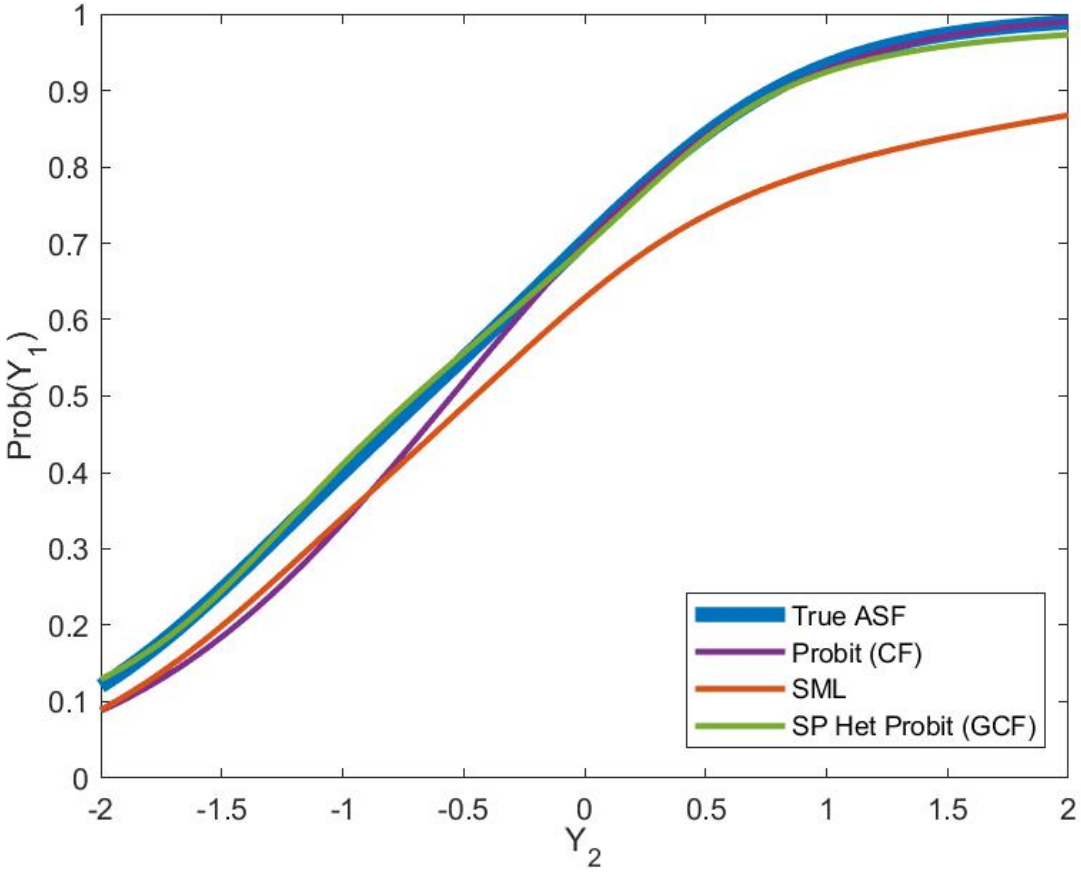
1,000 simulations of Sample size 1,000.

Figure D.19: **Student T with Non-Parametric GCF (h_o^3)**



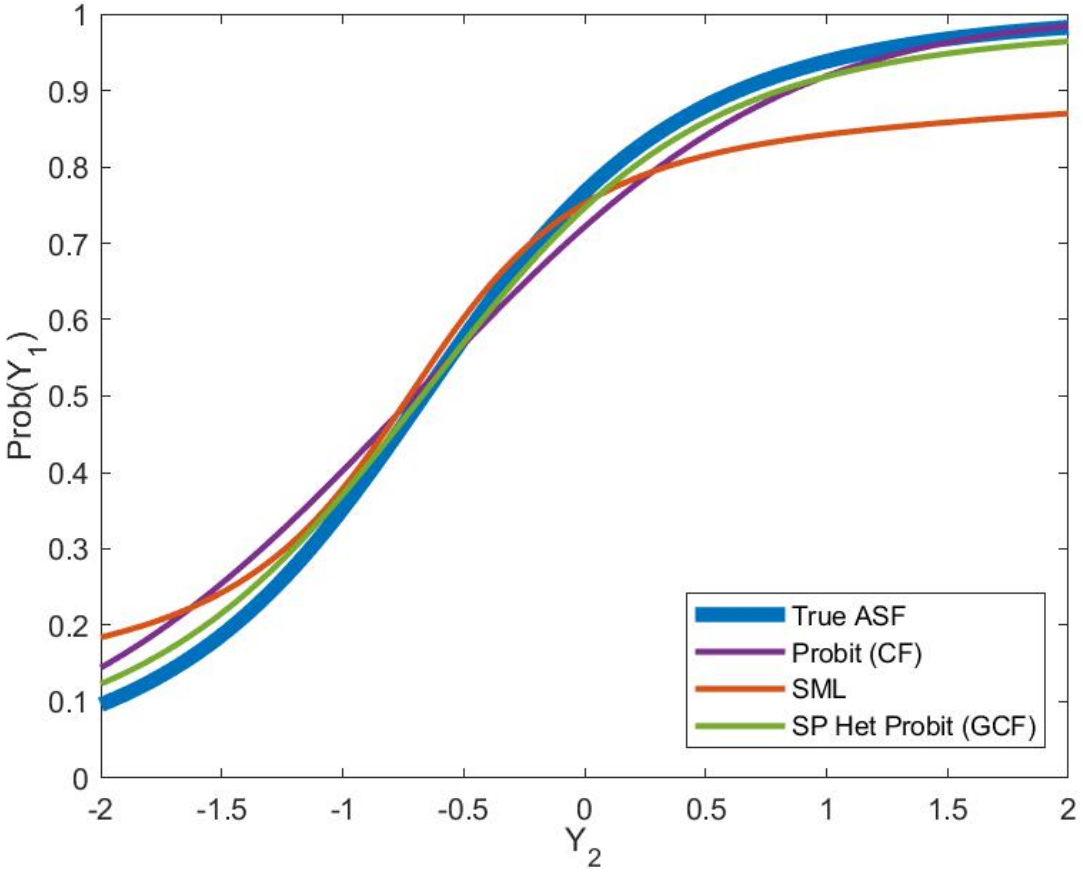
1,000 simulations of Sample size 1,000.

Figure D.20: **Gaussian Mixture with Non-Parametric GCF** (h_o^3)



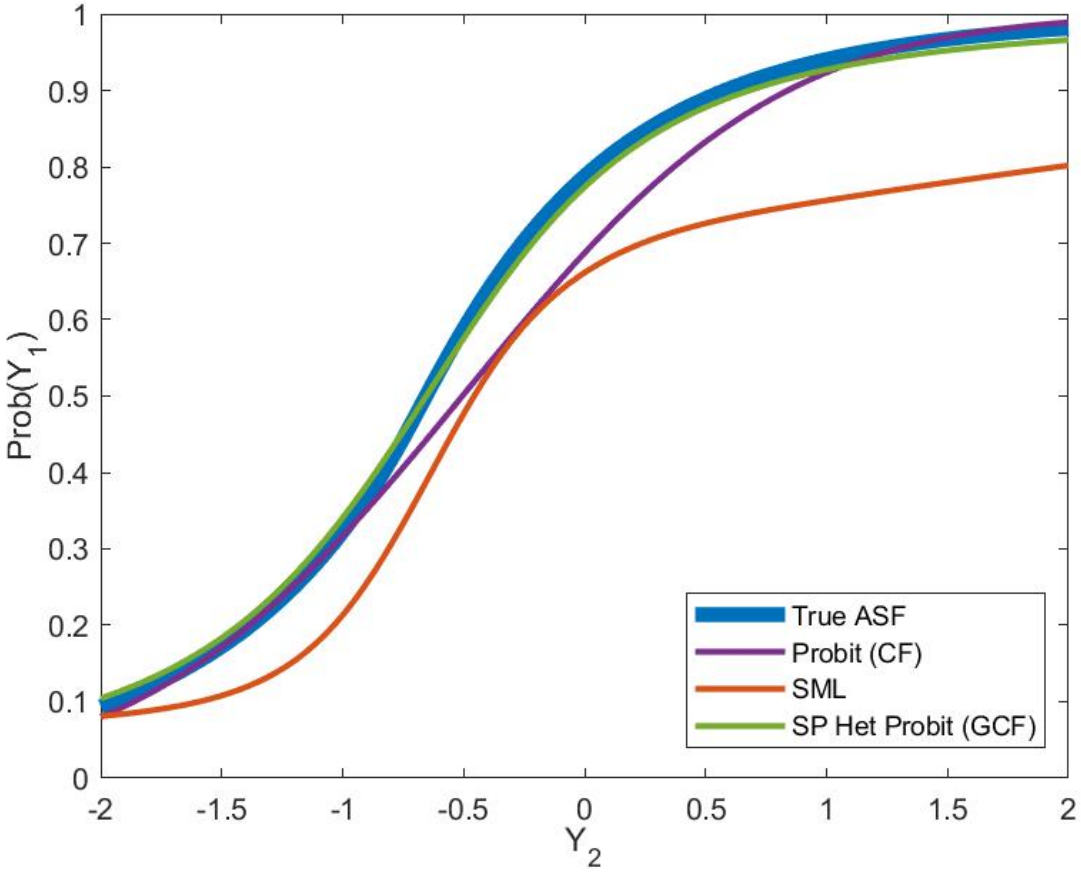
1,000 simulations of Sample size 1,000.

Figure D.21: **Heteroskedastic Logistic** ($h_o^1 = v_{2i}$)



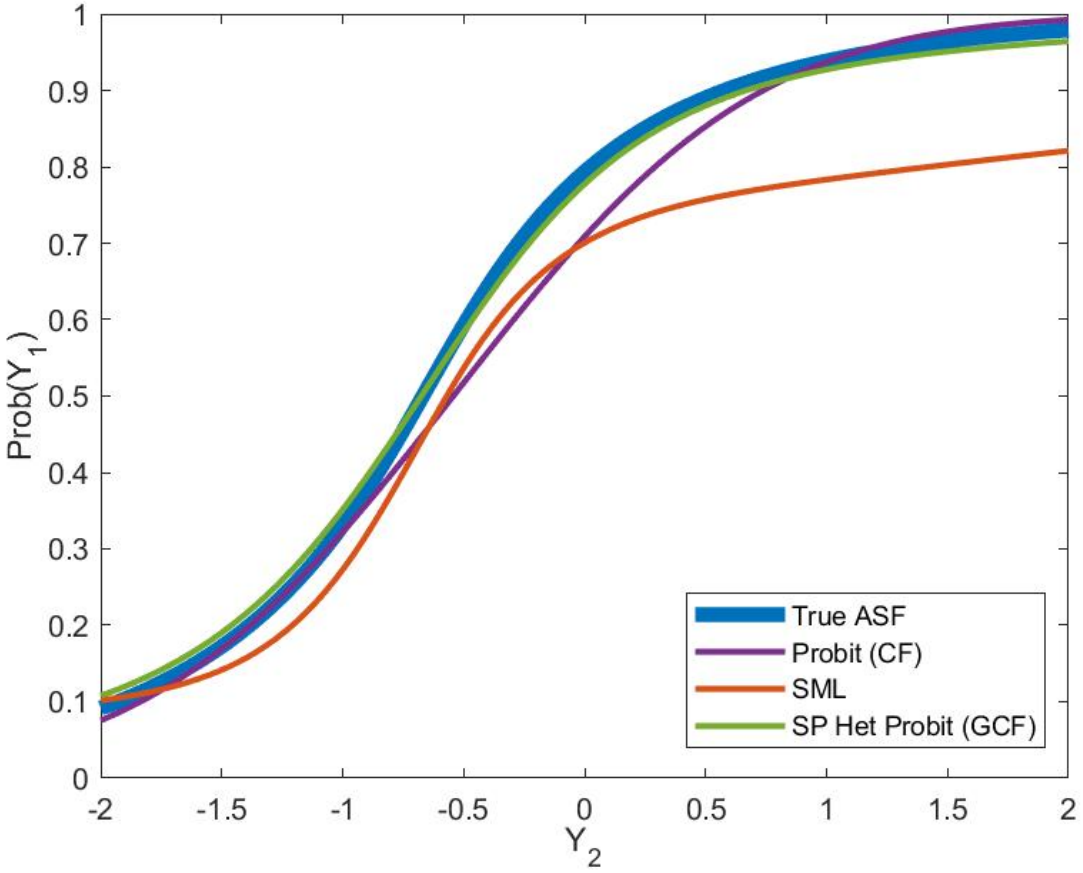
1,000 simulations of Sample size 1,000.

Figure D.22: **Heteroskedastic Logistic with Linear GCF** (h_o^2)



1,000 simulations of Sample size 1,000.

Figure D.23: **Heteroskedastic Logistic with Non-Parametric GCF (h_o^3)**



1,000 simulations of Sample size 1,000.

APPENDIX E

Tables for Chapter 2

General Control Function in the Demand for Premium Cable

Table E.1: **Summary Statistics**

Variables		Mean	Std. dev.
Premium Cable	(y_{1im})	0.329	0.470
Age	(z_{12i})	40.598	11.513
Income (in thousands)	(z_{12i})	60.019	34.782
Family Size	(z_{12i})	2.596	1.387
Price	(y_{2m})	40.691	11.594
Number of Channels	(z_{11m})	5.139	2.420
Cost	(z_{2m})	22.912	6.086

1,000 simulations of sample size 7,677.

Table E.2: **Comparison of Logit Parameter Estimates**

Variables	Logit (1)	Logit (CV) (2)	Logit (GCF) (3)	Logit (Over) (4)	TRUE
Price	-0.032	-0.109	-0.141	-0.141	-0.14
	(0.011)	(0.018)	(0.021)	(0.021)	
Price × Income Group 2	0.015	0.015	0.017	0.017	0.017
	(0.005)	(0.005)	(0.006)	(0.006)	
Price × Income Group 3	0.026	0.026	0.030	0.030	0.03
	(0.006)	(0.006)	(0.007)	(0.007)	
Price × Income Group 4	0.031	0.031	0.035	0.035	0.035
	(0.008)	(0.008)	(0.008)	(0.009)	
Price × Income Group 5	0.040	0.040	0.045	0.045	0.045
	(0.010)	(0.010)	(0.011)	(0.011)	
Number of Channels	-0.231	0.094	0.011	0.011	0.01
	(0.046)	(0.076)	(0.091)	(0.091)	
Income	0.002	0.002	0.005	0.005	0.005
	(0.004)	(0.004)	(0.004)	(0.004)	
Household Size	0.024	0.024	0.031	0.031	0.03
	(0.038)	(0.038)	(0.044)	(0.044)	
Age	0.077	0.073	0.025	0.024	0.005
	(0.114)	(0.115)	(0.147)	(0.147)	
Age^2	0.004	0.004	0.006	0.006	0.006
	(0.001)	(0.001)	(0.002)	(0.002)	

1,000 simulations of sample size 7,677 and standard deviations are given in parenthesis. Logit (CV) only includes the control variable v_{2i} to address the issue of endogeneity, Logit (GCF) uses the correct specification that allows for a general control function, and Logit (Over) over specifies the control function by including terms that are not in the true specification.

Table E.3: **Comparison of Price Elasticity Estimates**

Estimator	Mean	Std. dev.
OLS	0.485	12.009
CF	0.082	55.989
Logit	-0.386	0.267
Logit (CV)	-2.536	0.489
Logit (GCF)	-3.348	0.571
Logit (Over)	-3.350	0.571
TRUE	-3.320	

1,000 simulations of sample size 7,677.

ASF Estimates for the Effect of Income on Homeownership

Table E.4: **Comparison of Summary Statistics**

		Rothe		Simulated Data	
Variable		Mean	Std. dev.	Mean	Std. dev.
Homeowner	(y_1)	0.599	0.490	0.608	0.488
log(total income)	(y_2)	7.853	0.324	7.857	0.316
Age	(z_{11})	40.613	5.374	40.633	5.330
Children in HH	(z_{12})	0.848	0.359	0.851	0.356
Education of Wife					
Intermediate Degree	(z_{21})	0.415	0.493	0.422	0.494
High Degree	(z_{22})	0.103	0.304	0.111	0.314
Wife Working	(z_{23})	0.699	0.459	0.689	0.463

1,000 simulations of sample size 981.

Table E.5: **Comparison of Parameter Estimates**

	Rothe				Simulated Data				
	RF	Probit	Probit (CV)	SML	RF	Probit	Probit (CV)	SML	Het-Probit (GCF)
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
log(Income) (y_2)		2.1343	4.7923	3.8533		1.3789	4.1969	3.9605	3.9202
		(0.5571)	(1.5135)	(1.3338)		(0.0122)	(0.0725)	(0.0237)	(0.0511)
Age (z_{11})	0.0117	0.2076	0.0863	0.0982	0.0117	0.1084	0.0971	0.0925	0.0907
	(0.0117)	(0.0257)	(0.0209)	(0.0889)	(0.0001)	(0.0008)	(0.0010)	(0.0007)	(0.0008)
Child (z_{12})	0.0911	1	1	1	0.0911	1	1	1	1
	(0.0194)				(0.0009)				
CV ($\hat{v}_2$)			-3.0348				-2.6510		
			(1.3048)				(0.0600)		
Ed. of Wife									
Intm. (z_{21})	0.0642				0.0646				
	(0.0185)				(0.0008)				
High (z_{22})	0.1291				0.1286				
	(0.0298)				(0.0012)				
Wife Emp (z_{23})	0.0911				0.0914				
	(0.0194)				(0.0008)				
R^2	0.1072				0.1252				

1,000 simulations of sample size 981. Standard errors (for Rothe) and standard deviations (for Simulated Data) are given in parenthesis. RF is the reduced form first stage estimates, Probit does not address endogeneity at all, Probit (CV) is the Rivers and Vuong (1988) estimator that is a Probit model that only includes the control variable $\hat{v}_{2i}$ as an additional covariate to address endogeneity SML is the estimator proposed in Rothe (2009). Het-Probit (GCF) is the proposed heteroskedastic Probit with a flexible control function. Since coefficients are only identified to scale, I normalize the coefficients in columns (2)-(4) and (6)-(9) so the coefficient on Children in HH is 1. This allows for comparisons across the different specifications. True values of coefficients on log(Income) and Age are 3.80 and 0.09 respectively.

Table E.6: **APE Results and Simulated Distribution (True APE = 0.6448)**

Specification	Mean	SD	10%	25%	50%	75%	90%
Het-Probit (GCF)	0.6260	0.0034	0.4839	0.5603	0.6350	0.7017	0.7528
SML (Sieve)	0.6996	0.0025	0.5976	0.6475	0.6972	0.7512	0.8035
Probit	0.2851	0.0015	0.2247	0.2546	0.2858	0.3170	0.3434
Probit (CV)	0.5802	0.0037	0.4274	0.5098	0.5922	0.6643	0.7184
Lin. Prob. (OLS)	0.3117	0.0016	0.2462	0.2795	0.3122	0.3451	0.3739
Lin. Prob. (2SLS)	0.6960	0.0062	0.4553	0.5620	0.6886	0.8213	0.9475

1,000 simulations of sample size 981.

Empirical Example

Table E.7: **Summary Statistics**

Variables		Mean	Std Dev.	Mean (If Employed=0)	Mean (If Employed=1)
Employed	(y_1)	0.583	0.493		
Non-Wife Inc ($1000)	(y_2)	30.269	27.212	34.771	27.053
Education	(z_{11})	12.984	2.615	12.395	13.405
Experience	(z_{12})	20.444	10.445	22.080	19.274
Has kids (age<6)	(z_{13})	0.279	0.449	0.324	0.247
Has kids (age≥6)	(z_{14})	0.308	0.462	0.259	0.342
Husband's Education	(z_2)	13.148	2.977	12.811	13.388
Observations		5,634		2,348	3,286

1991 CPS data on Married Women Labor force participation.

Table E.8: **Coefficient Estimates for Married Women's LFP**

Variables	RF (1)	Probit (2)	Het Probit (3)	Probit (CV) (4)	Het Probit (CV) (5)	Probit (GCF) (6)	Het Probit (GCF) (7)	SML (Sieve) (8)
Non-wife Income		-0.071	-0.071	-0.024	-0.011	-0.073	-0.061	-0.072
		(0.011)	(0.010)	(0.029)	(0.026)	(1.873)	(0.025)	(0.004)
Non-wife Income		-0.058	-0.034	-0.069	-0.045	0.010	0.030	0.031
× Has Kids (Age<6)		(0.016)	(0.015)	(0.021)	(0.019)	(3.128)	(0.037)	(0.003)
Non-wife Income		-0.005	-0.010	-0.008	-0.017	0.068	0.065	0.059
× Has Kids (Age≥6)		(0.014)	(0.011)	(0.017)	(0.015)	(3.825)	(0.035)	(0.007)
Education	0.002	1	1	1	1	1	1	1
	(2.42×10^{-4})							
Experience	0.003	-0.168	-0.134	-0.224	-0.208	-0.221	-0.223	-0.223
	$(2.68\text{E}\times10^{-4})$	(0.028)	(0.036)	(0.052)	(0.047)	(2.946)	(0.062)	(0.011)
Has Kids (Age<6)	1.09×10^{-4}	-2.782	-2.870	-3.432	-3.886	-5.693	-6.263	-6.031
	(2.40×10^{-5})	(0.795)	(0.813)	(1.015)	(0.996)	(363.525)	(1.816)	(0.259)
Has Kids (Age≥6)	1.10×10^{-4}	1.102	1.337	1.036	1.322	-1.274	-1.196	-1.306
	(2.3×10^{-5})	(0.665)	(0.614)	(0.770)	(0.774)	(484.991)	(1.142)	(0.117)
Husband's Education	0.002							
	(4.54×10^{-4})							
Husband's Education								
× Has Kids (Age<6)	0.014							
Husband's Education	(0.004)							
× Has Kids (Age<6)	0.012							
	(0.002)							

1991 CPS data on Married Women Labor force participation. Standard errors given in parenthesis and calculated using 100 bootstraps. F-test in Reduced Form is a joint test of significant for the terms that include the instrument husband's education.

Table E.8 (cont'd)

Variables	RF (1)	Probit (2)	Het Probit (3)	Probit (CV) (4)	Het Probit (CV) (5)	Probit (GCF) (6)	Het Probit (GCF) (7)	SML (Sieve) (8)
Control Function								
$\hat{v}_{2i}$				-0.062	-0.080	-0.006	-0.025	
				(0.035)	(0.031)	(4.006)	(0.032)	
$\hat{v}_{2i}\times$ Has Kids (Age<6)						-0.093	-0.088	
						(5.647)	(0.056)	
$\hat{v}_{2i}\times$ Has Kids (Age≥6)						-0.088	-0.091	
						(7.727)	(0.045)	
Heteroskedasticity								
Non-wife Income			-0.004		-0.004		-0.004	
			(0.001)		(0.001)		(0.001)	
Education			1.25×10^{-4}		0.003		0.000	
			(1.92×10^{-4})		(0.005)		(0.001)	
Experience			0.008		0.009		0.005	
			(0.006)		(0.006)		(0.006)	
F-Stat	45.304							

1991 CPS data on Married Women Labor force participation. Standard errors given in parenthesis and calculated using 100 bootstraps. F-test in Reduced Form is a joint test of significant for the terms that include the instrument husband's education.

Table E.9: **Wald Test Results**

Null Hypothesis	Non-Wife Income is Exogenous				Control Variable		Homoskedasticity		
Alternative Hypothesis	Non-Wife Income is Endogenous				General Control Function		Heteroskedasticity		
Wald Statistic	4.582	12.219	4.749	12.987	4.949	7.109	15.213	15.851	10.658
p-value	0.032	0.007	0.029	0.005	0.084	0.029	0.002	0.001	0.014
Additional Assumptions:									
Homoskedasticity	x		x		x				
Endogeneity (CV)	x	x						x	
Endogeneity (GCF)			x	x					x

1991 CPS data on Married Women Labor force participation. Wald Statistics calculated using bootstrapped standard errors.

Table E.10: **APE Estimates for Non-Wife Income effect on Wife's LFP**

Estimators	APE	SE
OLS	-0.00333	0.00023
2SLS	-0.00253	0.00095
Probit	-0.00266	0.00032
Het Probit	-0.00331	0.00020
Probit (CV)	-0.00155	0.00076
Het Probit (CV)	-0.00116	0.00088
Probit (GCF)	-0.00180	0.00093
Het Probit (GCF)	0.00027	0.00103

1991 CPS data on Married Women Labor force participation. Standard errors given in parenthesis and calculated using 100 bootstraps.

Extension: Semi-Parametric Distribution Free Estimator

Table E.11: **Logistic Distribution** ($h_o^1 = v_{2i}$)

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0131	0.1717	0.1721	0.8875	0.9988	1.0947
	SML	-0.0352	0.1778	0.1811	0.8595	0.9702	1.0829
	SP Het Probit (GCF)	-0.0183	0.2210	0.2216	0.8603	0.9873	1.1191
500	Probit (CF)	-0.0073	0.1192	0.1194	0.9172	0.9992	1.0758
	SML	-0.0219	0.1236	0.1255	0.9024	0.9817	1.0661
	SP Het Probit (GCF)	-0.0124	0.1563	0.1567	0.8975	0.9995	1.0864
1,000	Probit (CF)	0.0007	0.0821	0.0820	0.9499	1.0022	1.0593
	SML	-0.0075	0.0840	0.0843	0.9404	0.9962	1.0521
	SP Het Probit (GCF)	0.0004	0.1246	0.1245	0.9285	1.0052	1.0831

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.12: **Uniform Distribution** ($h_o^1 = v_{2i}$)

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0210	0.1808	0.1819	0.8676	0.9862	1.1053
	SML	-0.0402	0.1898	0.1939	0.8400	0.9663	1.0985
	SP Het Probit (GCF)	-0.0354	0.2598	0.2621	0.8240	0.9773	1.1336
500	Probit (CF)	-0.0056	0.1223	0.1224	0.9180	0.9991	1.0734
	SML	-0.0124	0.1225	0.1230	0.9089	0.9907	1.0730
	SP Het Probit (GCF)	-0.0939	0.2908	0.3055	0.7547	0.9294	1.0814
1,000	Probit (CF)	0.0008	0.0856	0.0855	0.9484	1.0047	1.0613
	SML	-0.0021	0.0848	0.0848	0.9444	1.0007	1.0559
	SP Het Probit (GCF)	-0.0996	0.2097	0.2320	0.7794	0.9149	1.0371

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.13: **Student T Distribution** ($h_o^1 = v_{2i}$)

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0123	0.1611	0.1614	0.8928	1.0019	1.0989
	SML	-0.0190	0.1571	0.1582	0.8853	0.9850	1.0902
	SP Het Probit (GCF)	-0.0136	0.1872	0.1876	0.8841	0.9917	1.1018
500	Probit (CF)	-0.0083	0.1087	0.1090	0.9202	0.9888	1.0709
	SML	-0.0122	0.1077	0.1083	0.9157	0.9875	1.0648
	SP Het Probit (GCF)	0.0042	0.1254	0.1254	0.9161	1.0096	1.0915
$1,000$	Probit (CF)	-0.0031	0.0761	0.0761	0.9466	0.9960	1.0500
	SML	-0.0045	0.0735	0.0736	0.9475	0.9975	1.0454
	SP Het Probit (GCF)	0.0100	0.1023	0.1027	0.9473	1.0168	1.0764

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.14: **Gaussian Mixture Distribution** ($h_o^1 = v_{2i}$)

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0135	0.1840	0.1845	0.8783	1.0023	1.1095
	SML	-0.0373	0.1972	0.2006	0.8345	0.9765	1.1025
	SP Het Probit (GCF)	-0.0203	0.2463	0.2470	0.8425	0.9995	1.1400
500	Probit (CF)	-0.0130	0.1201	0.1207	0.9066	0.9910	1.0759
	SML	-0.0248	0.1230	0.1254	0.8923	0.9794	1.0653
	SP Het Probit (GCF)	-0.0608	0.2383	0.2458	0.8090	0.9548	1.0908
$1,000$	Probit (CF)	-0.0021	0.0877	0.0877	0.9380	1.0011	1.0558
	SML	-0.0085	0.0887	0.0891	0.9321	0.9948	1.0488
	SP Het Probit (GCF)	-0.0440	0.1658	0.1715	0.8562	0.9715	1.0676

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.15: **Logistic Distribution with Linear GCF (h_o^2)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0119	0.1533	0.1537	0.8913	0.9948	1.0912
	SML	-0.0710	0.1539	0.1694	0.8296	0.9213	1.0314
	SP Het Probit (GCF)	-0.0127	0.1985	0.1988	0.8643	0.9936	1.1169
500	Probit (CF)	-0.0064	0.1135	0.1136	0.9271	1.0038	1.0647
	SML	-0.0602	0.1044	0.1205	0.8737	0.9397	1.0111
	SP Het Probit (GCF)	-0.0059	0.1502	0.1503	0.9154	0.9995	1.0917
1,000	Probit (CF)	-0.0013	0.0765	0.0765	0.9509	0.9996	1.0500
	SML	-0.0517	0.0696	0.0867	0.9014	0.9509	0.9972
	SP Het Probit (GCF)	0.0033	0.1120	0.1120	0.9293	1.0067	1.0820

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.16: **Uniform Distribution with Linear GCF (h_o^2)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0310	0.1794	0.1820	0.8609	0.9757	1.0863
	SML	-0.0868	0.1722	0.1927	0.8028	0.9128	1.0223
	SP Het Probit (GCF)	-0.0145	0.2111	0.2115	0.8513	0.9923	1.1239
500	Probit (CF)	-0.0166	0.1194	0.1205	0.9068	0.9889	1.0649
	SML	-0.0644	0.1114	0.1286	0.8672	0.9370	1.0094
	SP Het Probit (GCF)	-0.0168	0.2088	0.2094	0.8467	0.9911	1.1106
1,000	Probit (CF)	-0.0141	0.0836	0.0848	0.9311	0.9869	1.0437
	SML	-0.0604	0.0762	0.0972	0.8886	0.9413	0.9911
	SP Het Probit (GCF)	-0.0203	0.1715	0.1726	0.8724	0.9954	1.0915

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.17: **Student T Distribution with Linear GCF (h_o^2)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0033	0.1421	0.1421	0.9128	1.0040	1.0904
	SML	-0.0563	0.1309	0.1425	0.8626	0.9486	1.0311
	SP Het Probit (GCF)	-0.0060	0.1674	0.1674	0.8890	1.0017	1.1013
500	Probit (CF)	-0.0034	0.0992	0.0992	0.9317	0.9959	1.0652
	SML	-0.0529	0.0889	0.1034	0.8892	0.9460	1.0049
	SP Het Probit (GCF)	0.0002	0.1208	0.1208	0.9257	1.0039	1.0789
1,000	Probit (CF)	-0.0002	0.0695	0.0695	0.9519	0.9985	1.0479
	SML	-0.0479	0.0636	0.0796	0.9083	0.9505	0.9957
	SP Het Probit (GCF)	0.0086	0.0967	0.0970	0.9473	1.0118	1.0684

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.18: **Gaussian Mixture Distribution with Linear GCF (h_o^2)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0195	0.1773	0.1783	0.8781	0.9915	1.1023
	SML	-0.0778	0.1768	0.1931	0.8175	0.9289	1.0282
	SP Het Probit (GCF)	-0.0377	1.0300	1.0302	0.8759	1.0019	1.1342
500	Probit (CF)	-0.0189	0.1163	0.1178	0.9042	0.9863	1.0605
	SML	-0.0721	0.1076	0.1294	0.8592	0.9273	0.9984
	SP Het Probit (GCF)	-0.0177	0.1930	0.1937	0.8674	0.9967	1.1047
1,000	Probit (CF)	-0.0094	0.0822	0.0827	0.9351	0.9904	1.0469
	SML	-0.0578	0.0760	0.0954	0.8902	0.9393	0.9943
	SP Het Probit (GCF)	-0.0014	0.1396	0.1395	0.9048	1.0022	1.0942

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.19: **Logistic Distribution with Non-Parametric GCF (h_o^3)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	0.0349	0.1565	0.1603	0.9421	1.0387	1.1367
	SML	-0.0525	0.1739	0.1816	0.8363	0.9467	1.0589
	SP Het Probit (GCF)	0.0117	0.2313	0.2315	0.9123	1.0244	1.1405
500	Probit (CF)	0.0354	0.1106	0.1161	0.9657	1.0395	1.1084
	SML	-0.0451	0.1245	0.1324	0.8776	0.9594	1.0376
	SP Het Probit (GCF)	0.0453	0.1537	0.1601	0.9581	1.0551	1.1472
1,000	Probit (CF)	0.0426	0.0774	0.0883	0.9929	1.0465	1.0950
	SML	-0.0344	0.0843	0.0910	0.9101	0.9649	1.0245
	SP Het Probit (GCF)	0.0010	0.1231	0.1230	0.9219	1.0074	1.0872

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.20: **Uniform Distribution with Non-Parametric GCF (h_o^3)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	0.0270	0.1695	0.1716	0.9199	1.0313	1.1433
	SML	-0.0710	0.1966	0.2089	0.8151	0.9353	1.0599
	SP Het Probit (GCF)	-0.0159	0.2418	0.2422	0.8469	1.0006	1.1395
500	Probit (CF)	0.0381	0.1110	0.1173	0.9612	1.0390	1.1127
	SML	-0.0488	0.1249	0.1341	0.8735	0.9541	1.0344
	SP Het Probit (GCF)	-0.0063	0.2364	0.2363	0.8518	1.0080	1.1458
1,000	Probit (CF)	0.0424	0.0789	0.0895	0.9878	1.0442	1.0971
	SML	-0.0403	0.0870	0.0958	0.9030	0.9643	1.0211
	SP Het Probit (GCF)	-0.0606	0.1781	0.1881	0.8281	0.9525	1.0616

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.21: **Student T Distribution with Non-Parametric GCF (h_o^3)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	0.0363	0.1431	0.1475	0.9511	1.0409	1.1333
	SML	-0.0338	0.1559	0.1595	0.8733	0.9677	1.0657
	SP Het Probit (GCF)	0.0241	0.1870	0.1885	0.9318	1.0314	1.1473
500	Probit (CF)	0.0373	0.0990	0.1058	0.9747	1.0376	1.1035
	SML	-0.0291	0.1045	0.1085	0.9026	0.9692	1.0401
	SP Het Probit (GCF)	0.0577	0.1205	0.1336	0.9807	1.0638	1.1440
1,000	Probit (CF)	0.0387	0.0700	0.0800	0.9883	1.0408	1.0856
	SML	-0.0264	0.0740	0.0786	0.9244	0.9737	1.0243
	SP Het Probit (GCF)	0.0078	0.1036	0.1038	0.9395	1.0132	1.0760

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.22: **Gaussian Mixture Distribution with Non-Parametric GCF (h_o^3)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	0.0303	0.1708	0.1733	0.9220	1.0385	1.1441
	SML	-0.0668	0.1945	0.2056	0.8071	0.9462	1.0685
	SP Het Probit (GCF)	0.0056	0.2327	0.2327	0.8857	1.0238	1.1484
500	Probit (CF)	0.0317	0.1130	0.1173	0.9565	1.0307	1.1115
	SML	-0.0581	0.1247	0.1375	0.8525	0.9403	1.0303
	SP Het Probit (GCF)	0.0106	0.2048	0.2050	0.9050	1.0290	1.1443
1,000	Probit (CF)	0.0389	0.0828	0.0915	0.9825	1.0408	1.0979
	SML	-0.0443	0.0922	0.1023	0.8920	0.9565	1.0225
	SP Het Probit (GCF)	-0.0296	0.1561	0.1588	0.8595	0.9771	1.0797

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.23: **Heteroskedastic Logistic ($h_o^1 = v_{2i}$)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0687	0.1798	0.1925	0.8222	0.9295	1.0513
	SML	-0.1041	0.1935	0.2197	0.7730	0.8895	1.0225
	SP Het Probit (GCF)	0.1213	0.2858	0.3104	0.9453	1.1030	1.2574
500	Probit (CF)	-0.0593	0.1319	0.1446	0.8557	0.9456	1.0256
	SML	-0.0926	0.1460	0.1728	0.8136	0.9040	1.0116
	SP Het Probit (GCF)	-0.0263	0.1532	0.1554	0.8836	0.9712	1.0713
1,000	Probit (CF)	-0.0516	0.0854	0.0998	0.8951	0.9497	1.0034
	SML	-0.0816	0.0953	0.1254	0.8529	0.9231	0.9824
	SP Het Probit (GCF)	-0.0043	0.1160	0.1160	0.9203	0.9940	1.0705

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.24: **Heteroskedastic Logistic with Linear GCF (h_o^2)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	-0.0171	0.1649	0.1657	0.8795	0.9869	1.0861
	SML	-0.0843	0.1601	0.1809	0.8114	0.9116	1.0197
	SP Het Probit (GCF)	0.1245	0.2358	0.2665	0.9792	1.1112	1.2494
500	Probit (CF)	-0.0148	0.1204	0.1213	0.9132	0.9870	1.0650
	SML	-0.0609	0.1163	0.1313	0.8621	0.9365	1.0193
	SP Het Probit (GCF)	-0.0019	0.1350	0.1349	0.9110	0.9996	1.0843
1,000	Probit (CF)	-0.0105	0.0816	0.0822	0.9380	0.9910	1.0399
	SML	-0.0557	0.0780	0.0958	0.8905	0.9449	0.9964
	SP Het Probit (GCF)	0.0026	0.1025	0.1025	0.9343	1.0019	1.0628

1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

Table E.25: **Heteroskedastic Logistic with Non-Parametric GCF (h_o^3)**

N	Estimators	Bias	Std. Dev.	RMSE	25%	50%	75%
250	Probit (CF)	0.0044	0.1618	0.1618	0.8972	1.0054	1.1102
	SML	-0.0846	0.1823	0.2009	0.7910	0.8995	1.0314
	SP Het Probit (GCF)	0.1307	0.2867	0.3149	0.9974	1.1174	1.2534
500	Probit (CF)	0.0051	0.1182	0.1182	0.9308	1.0111	1.0815
	SML	-0.0721	0.1360	0.1539	0.8405	0.9251	1.0204
	SP Het Probit (GCF)	0.0208	0.1264	0.1281	0.9343	1.0210	1.1102
$1,000$	Probit (CF)	0.0110	0.0795	0.0803	0.9568	1.0131	1.0662
	SML	-0.0646	0.0906	0.1112	0.8753	0.9398	0.9999
	SP Het Probit (GCF)	0.0042	0.1106	0.1106	0.9242	1.0001	1.0810

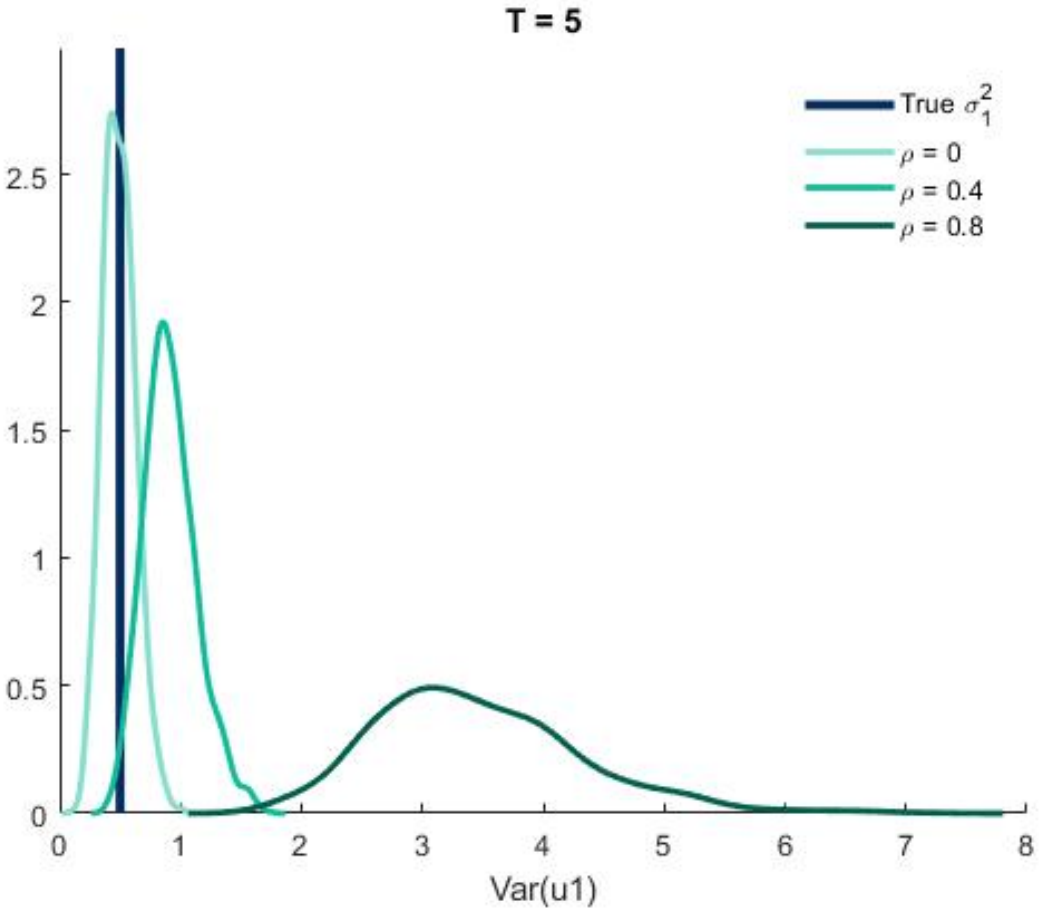
1,000 simulations. Root mean square error is reported in the third column and the 25th, 50th and 75th percentiles of the empirical distribution are reported in the last three columns.

APPENDIX F

Figures for Chapter 3

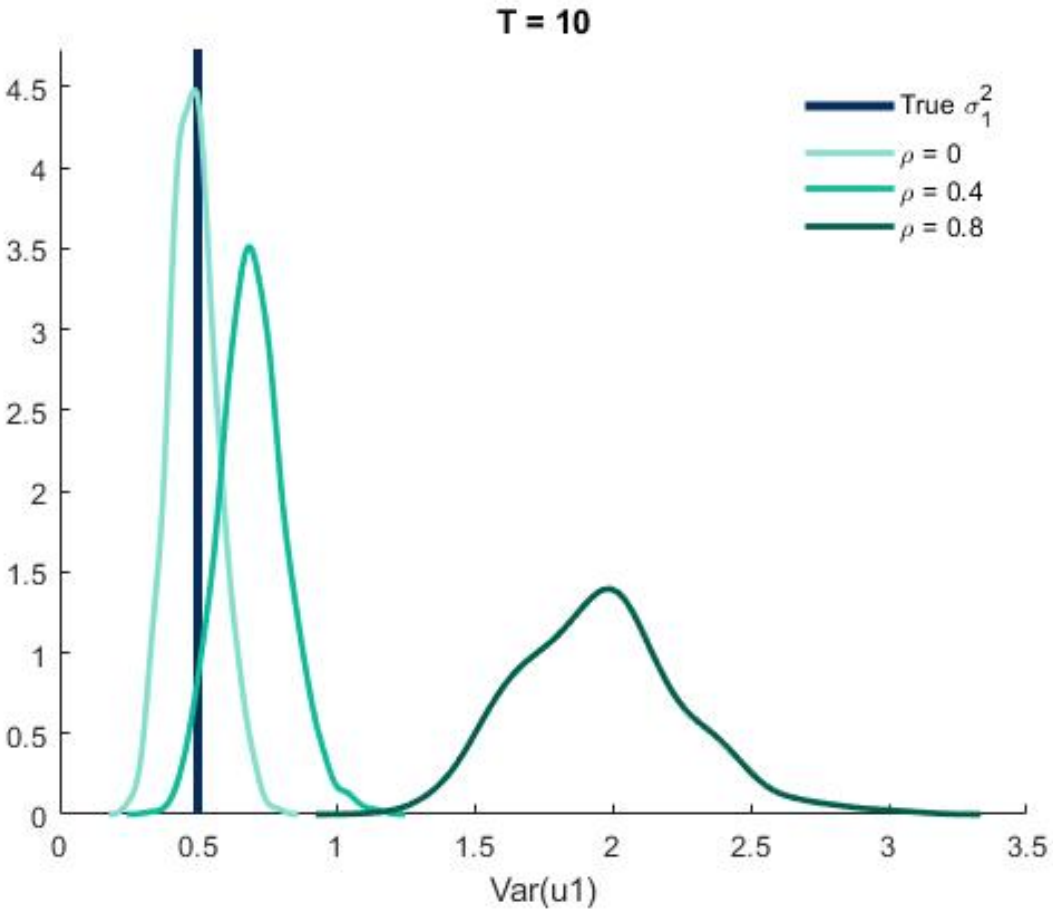
Simulation

Figure F.1: **Distribution of $\hat{\sigma}_1^2$ for T=5 under DGP1**



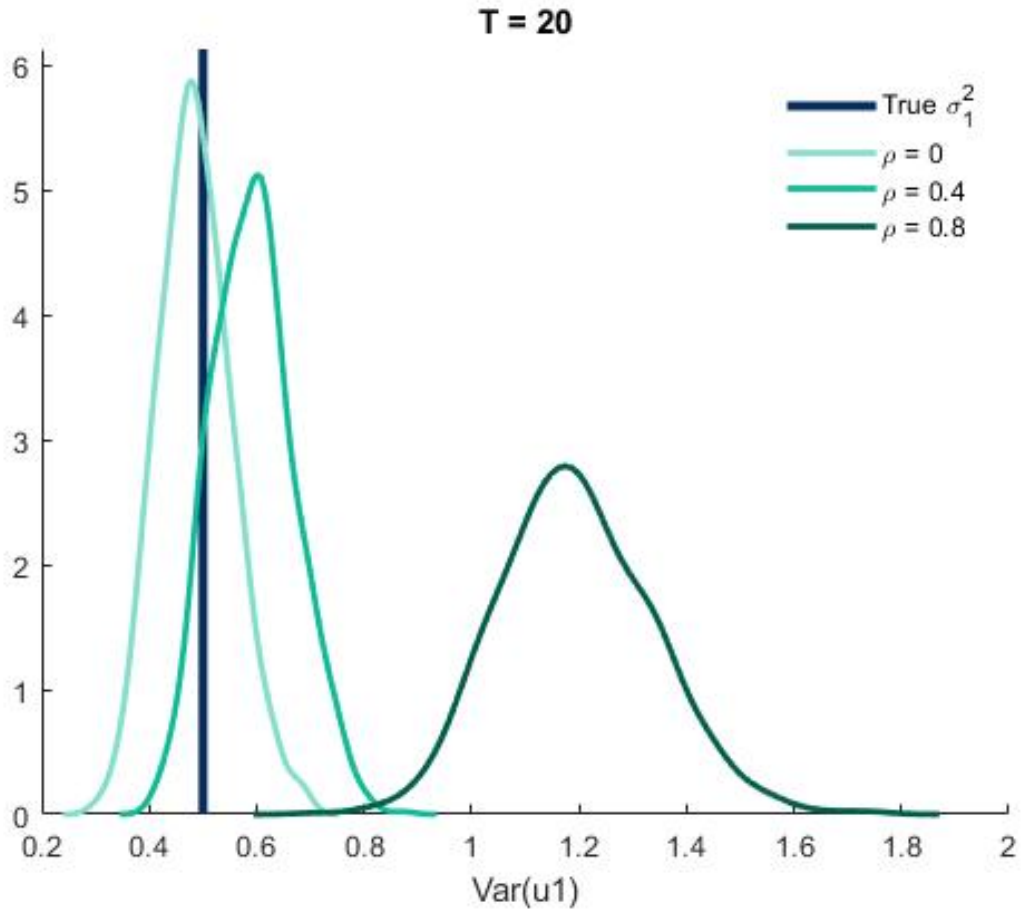
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.2: **Distribution of $\hat{\sigma}_1^2$ for T=10 under DGP1**



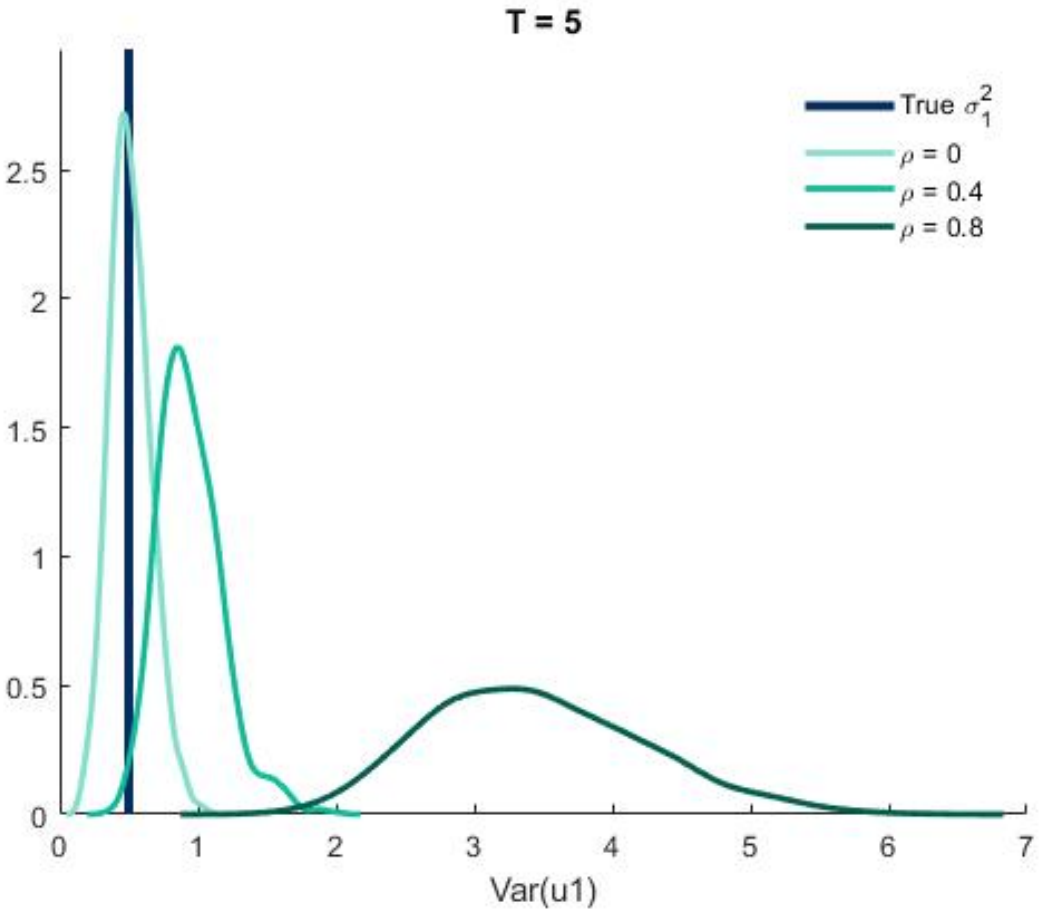
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.3: **Distribution of $\hat{\sigma}_1^2$ for T=20 under DGP1**



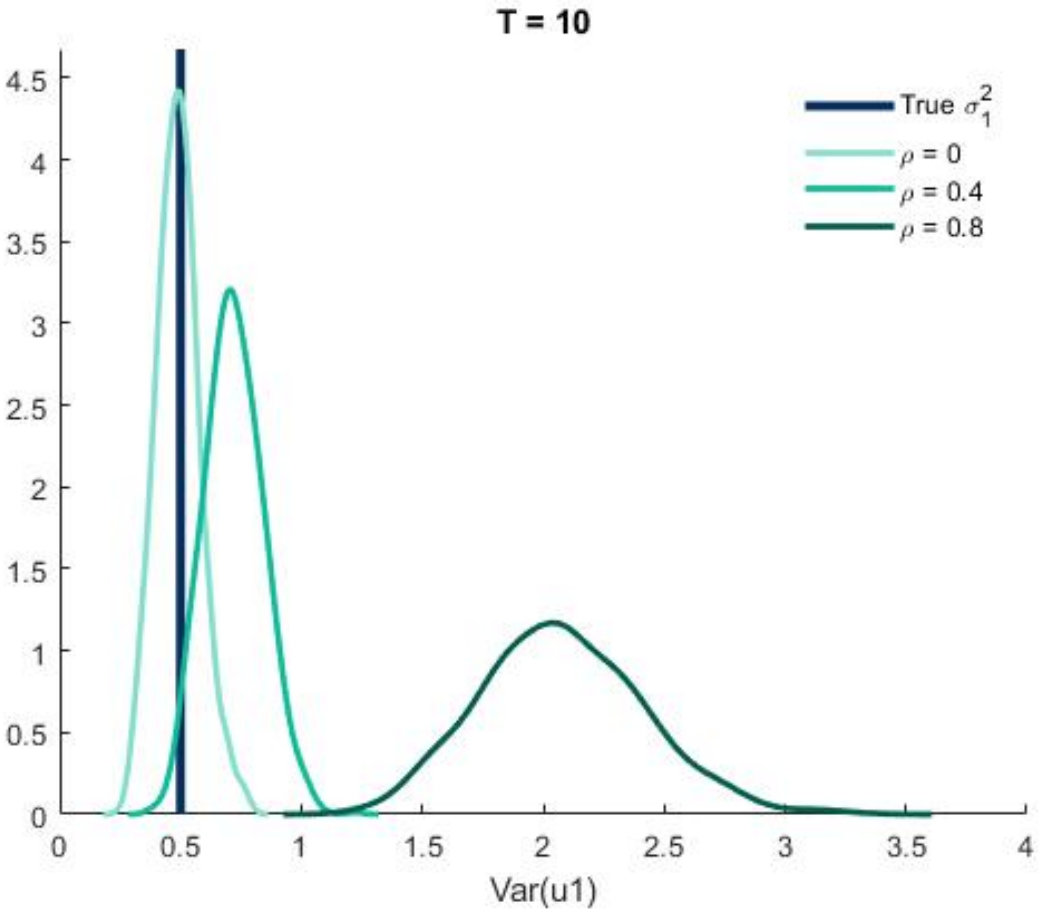
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.4: **Distribution of $\hat{\sigma}_1^2$ for T=5 under DGP2**



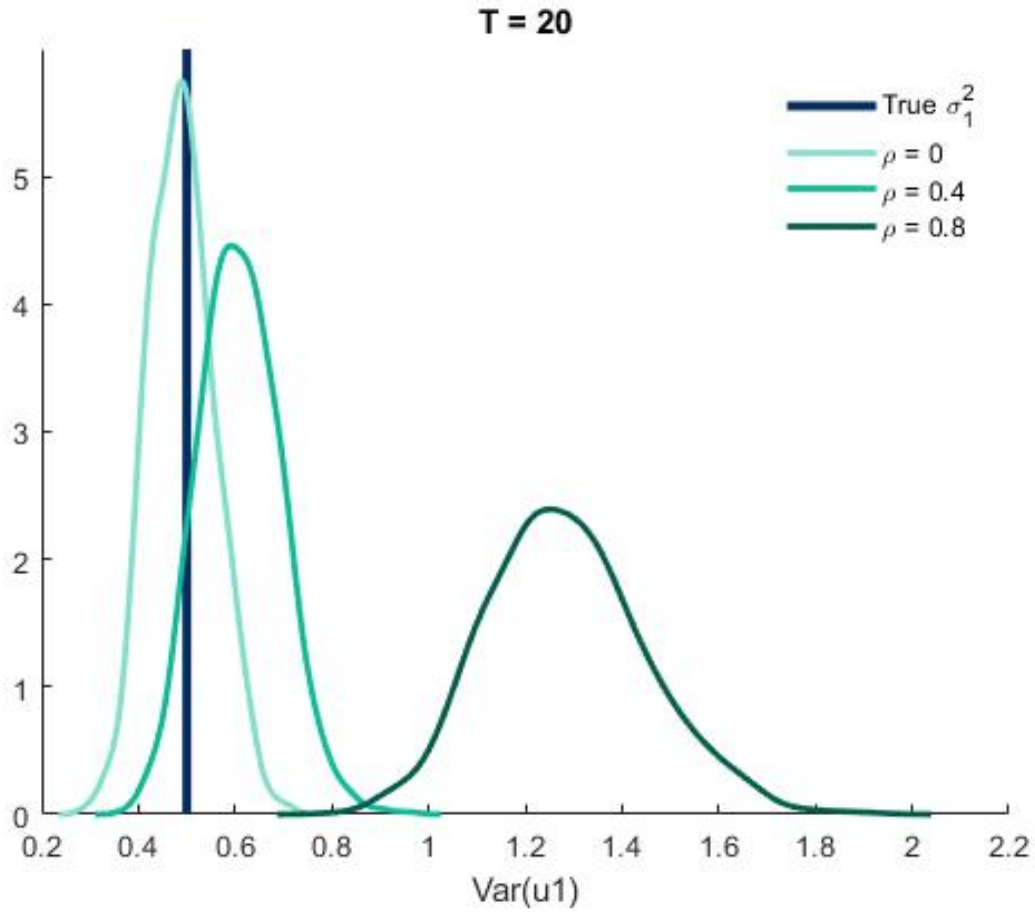
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.5: **Distribution of $\hat{\sigma}_1^2$ for T=10 under DGP2**



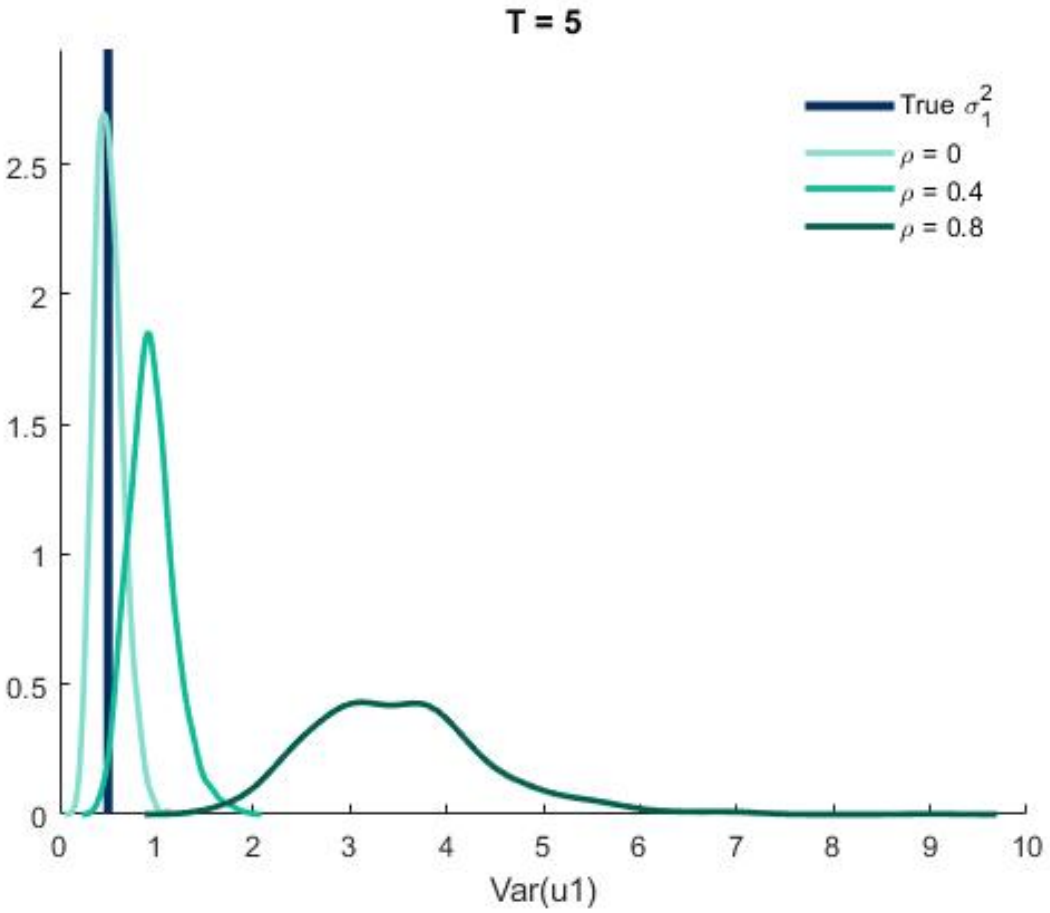
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.6: **Distribution of $\hat{\sigma}_1^2$ for T=20 under DGP2**



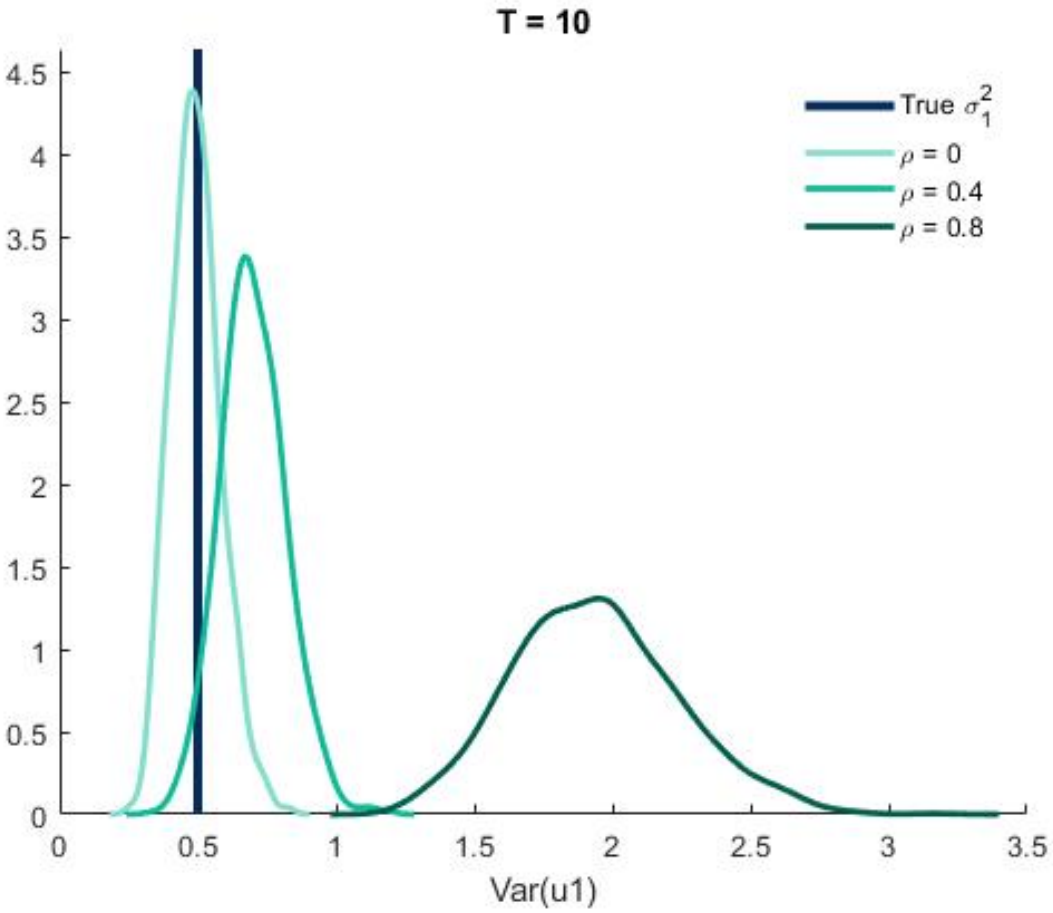
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.7: **Distribution of $\hat{\sigma}_1^2$ for T=5 under DGP3**



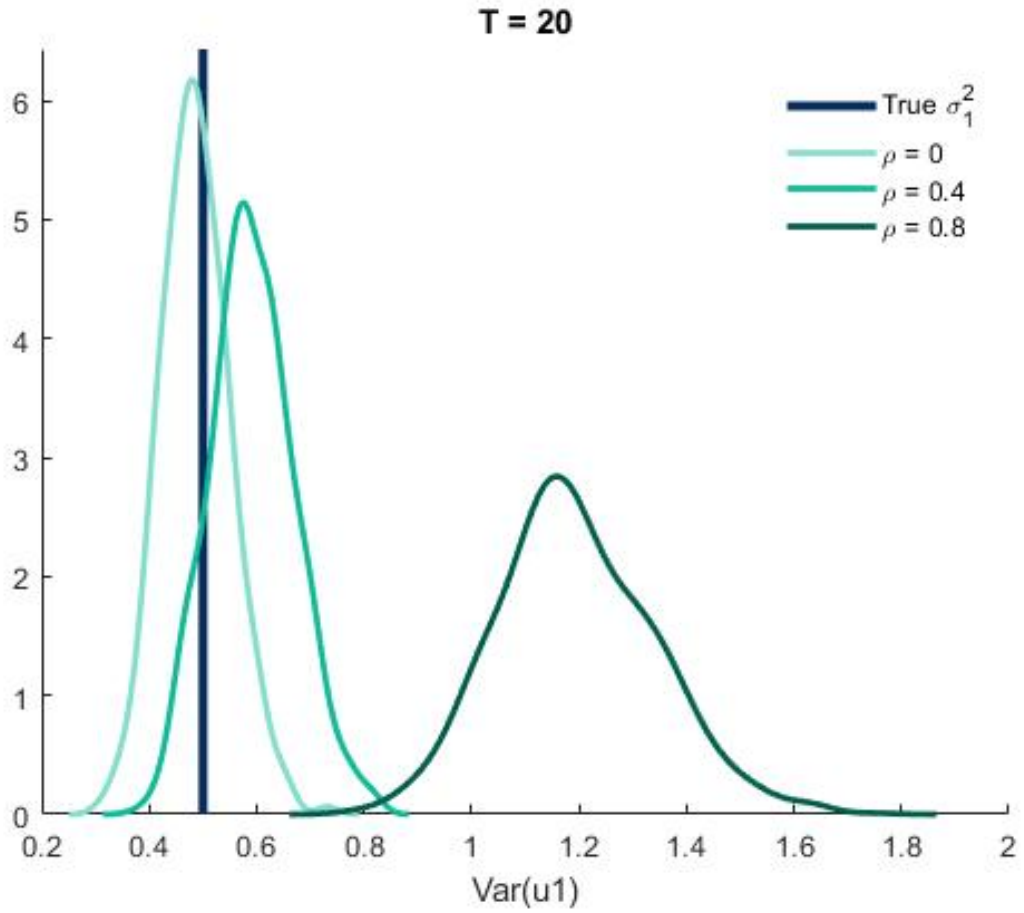
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.8: **Distribution of $\hat{\sigma}_1^2$ for T=10 under DGP3**



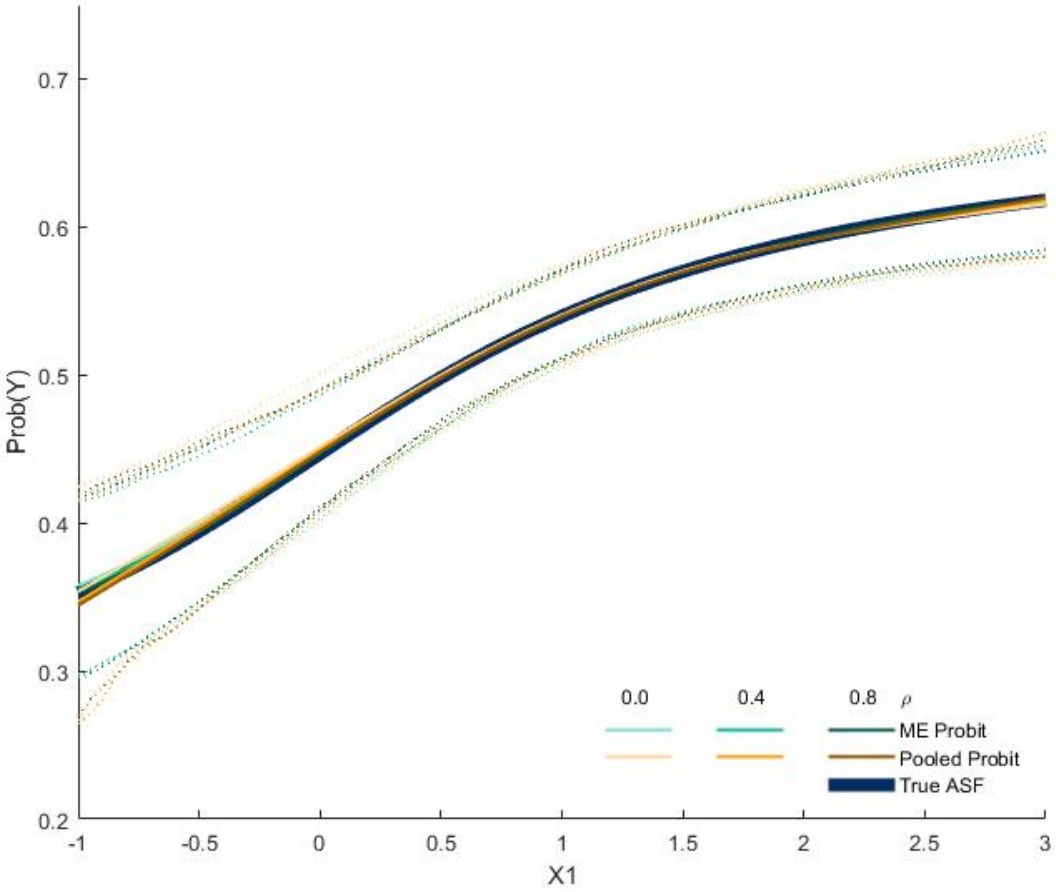
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.9: **Distribution of $\hat{\sigma}_1^2$ for T=20 under DGP3**



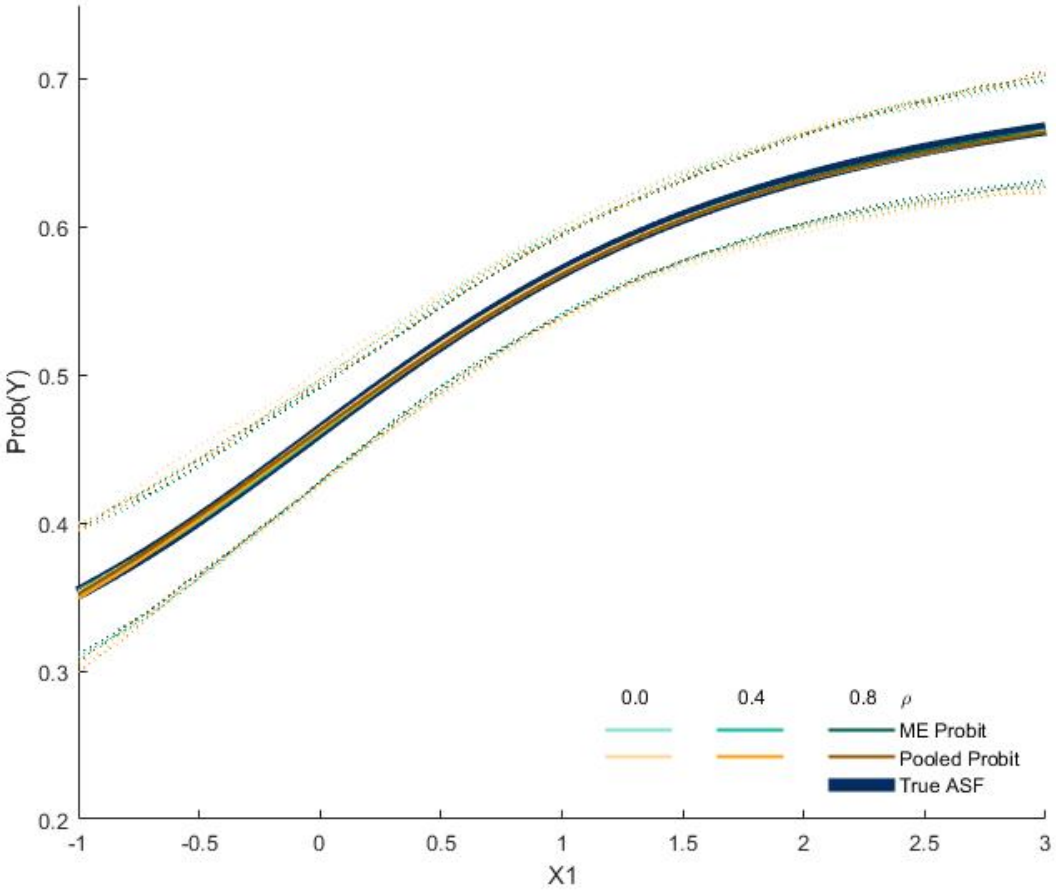
1,000 simulations of N=300. Increasing serial correlation as color lightens.

Figure F.10: **ASF Estimates for T=5 under DGP1**



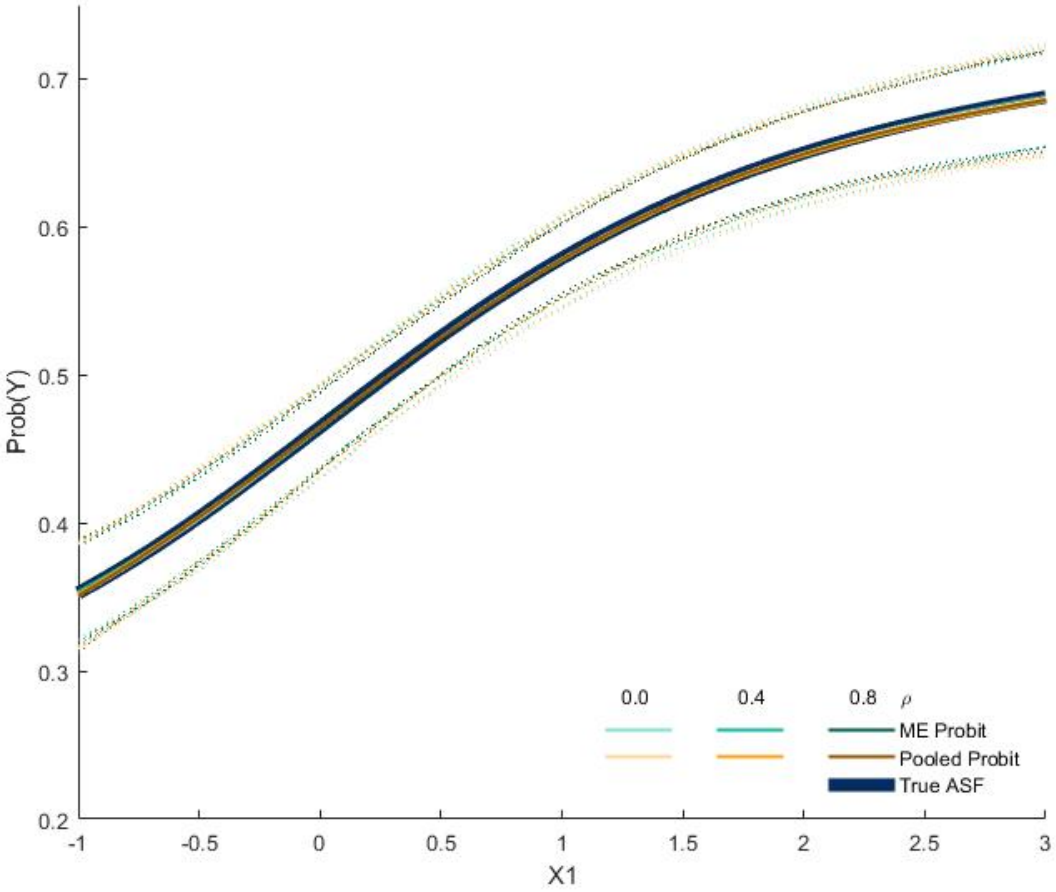
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.11: **ASF Estimates for T=10 under DGP1**



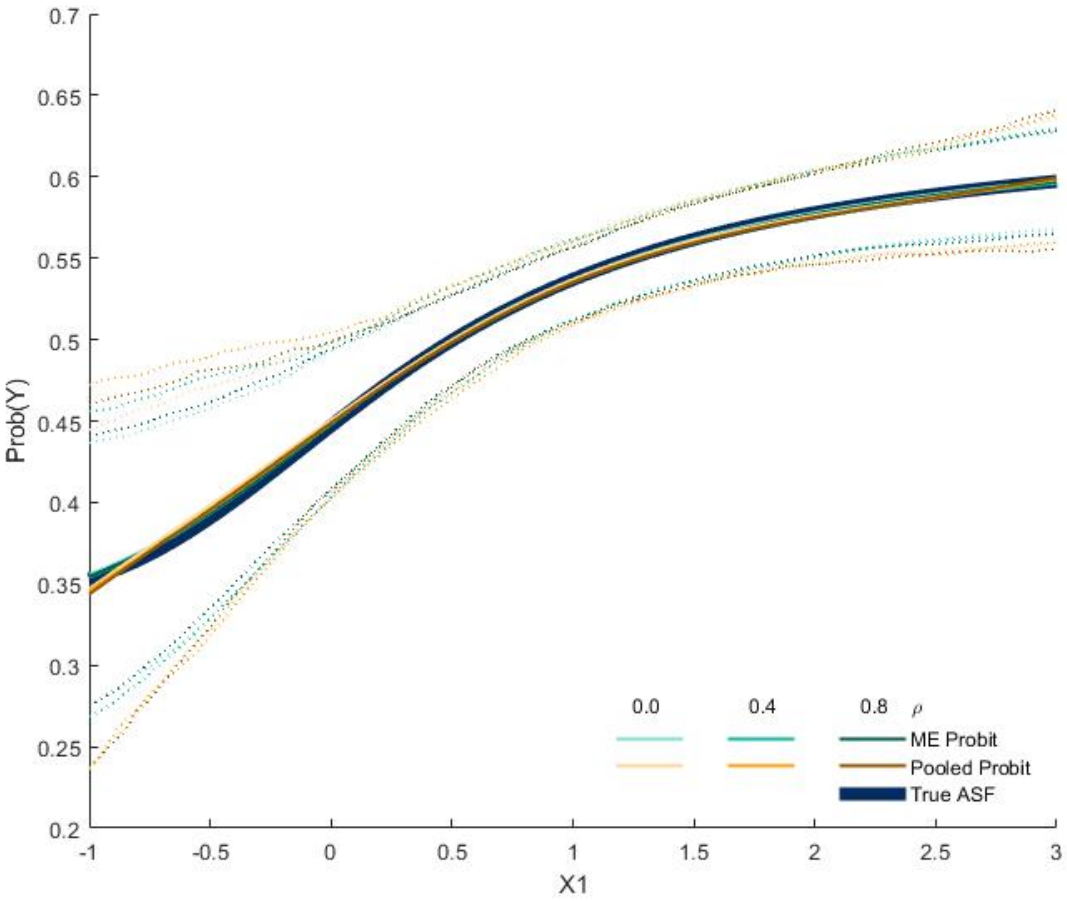
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.12: **ASF Estimates for T=20 under DGP1**



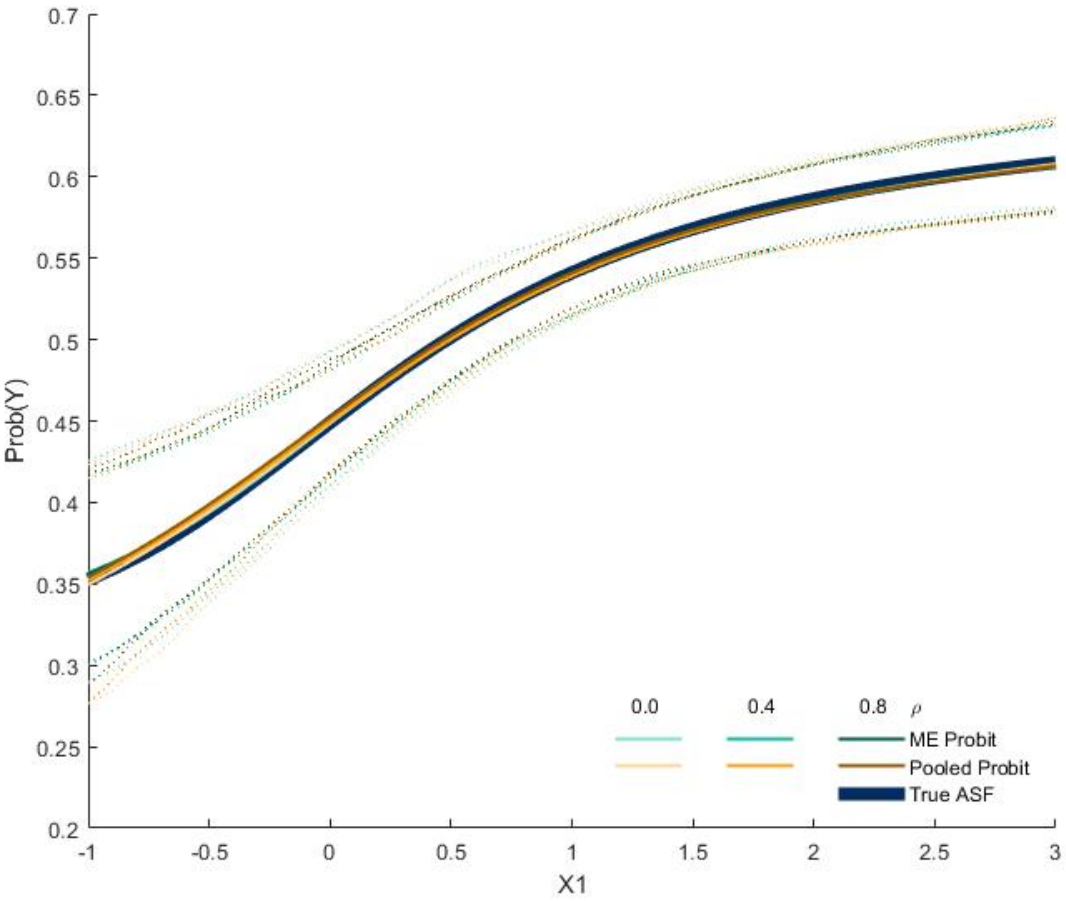
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.13: **ASF Estimates for T=5 under DGP2**



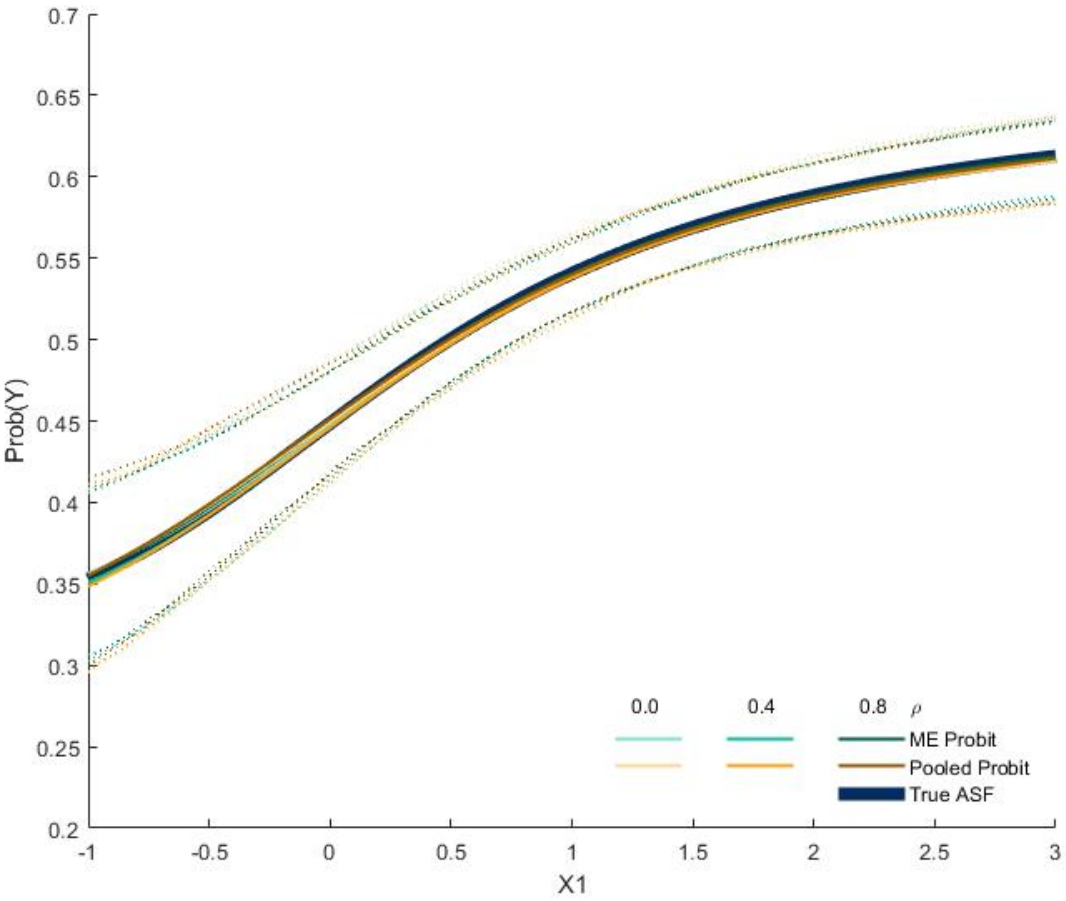
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.14: **ASF Estimates for T=10 under DGP2**



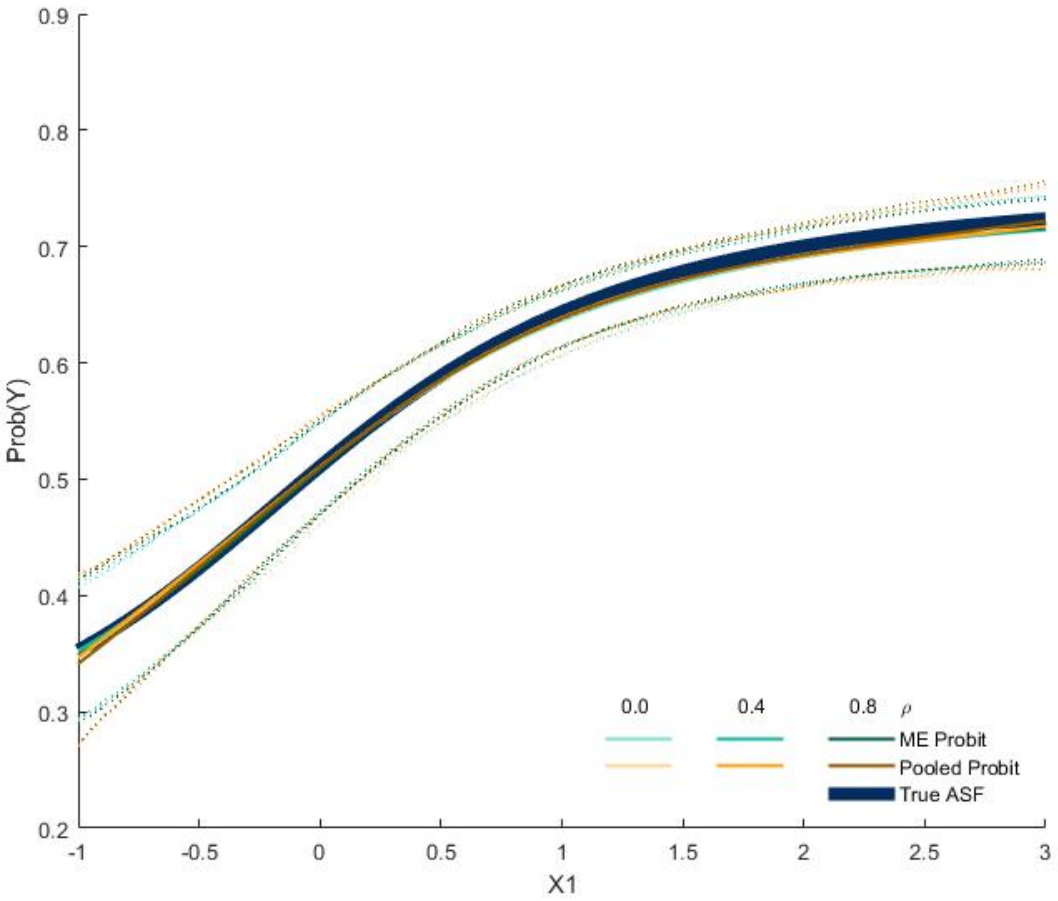
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.15: **ASF Estimates for T=20 under DGP2**



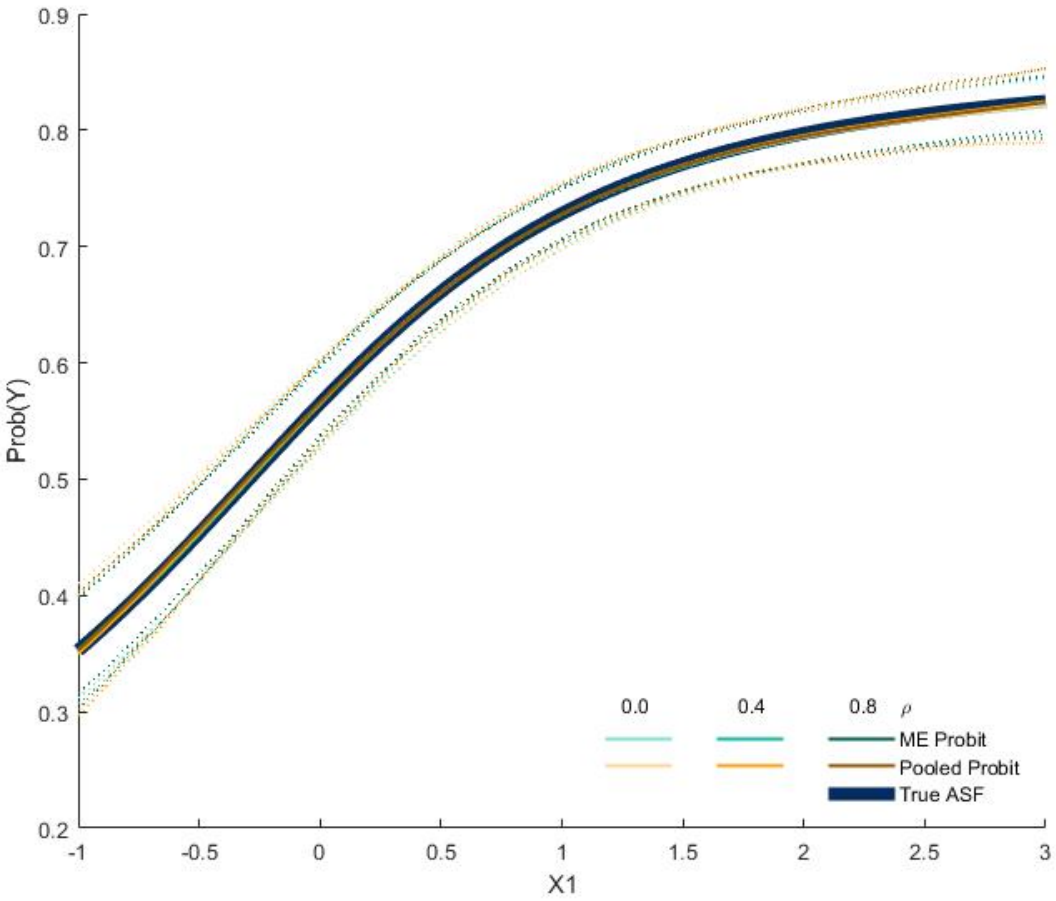
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.16: **ASF Estimates for T=5 under DGP3**



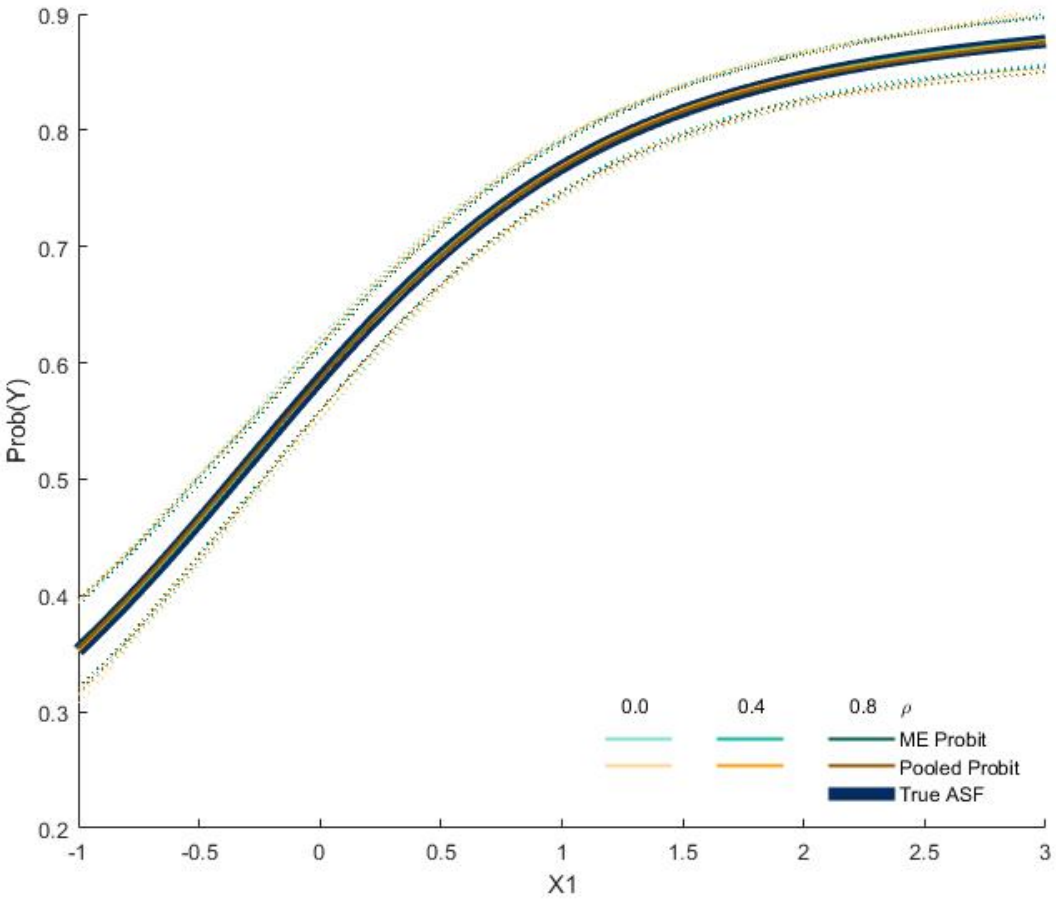
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.17: **ASF Estimates for T=10 under DGP3**



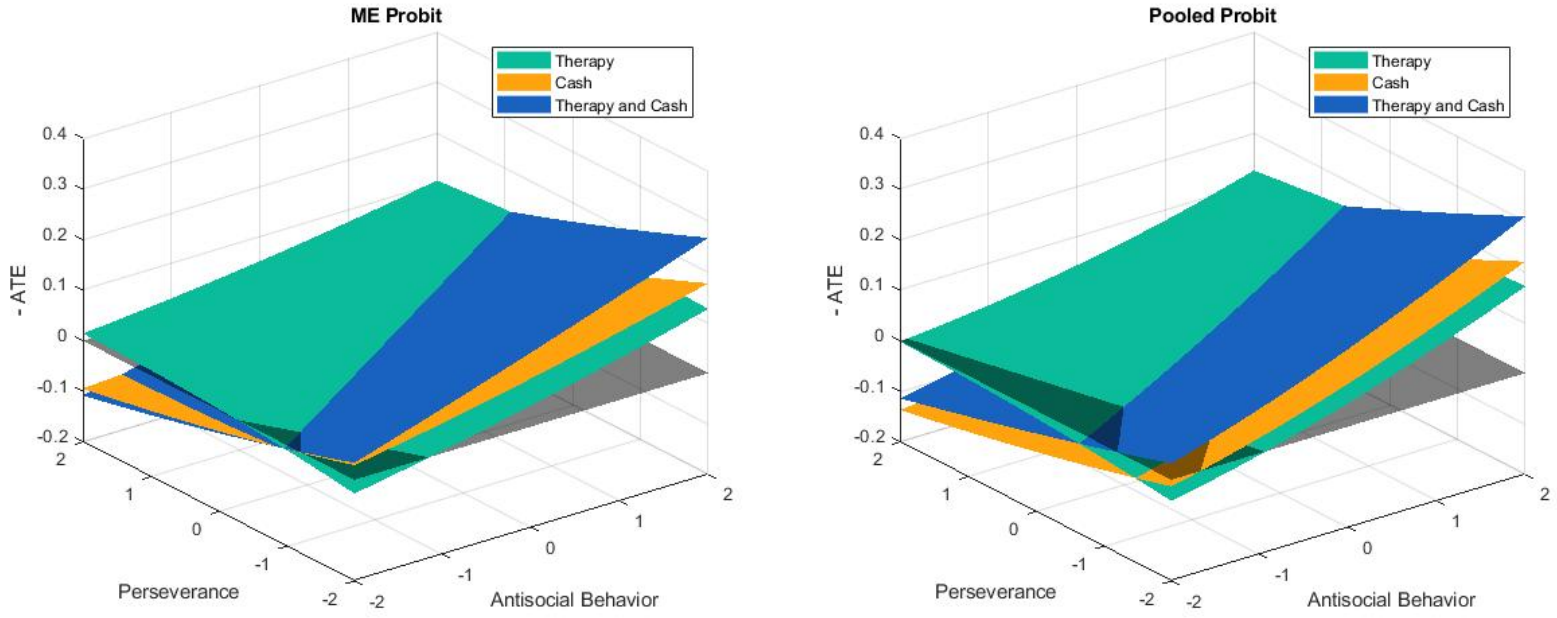
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.18: **ASF Estimates for T=20 under DGP3**



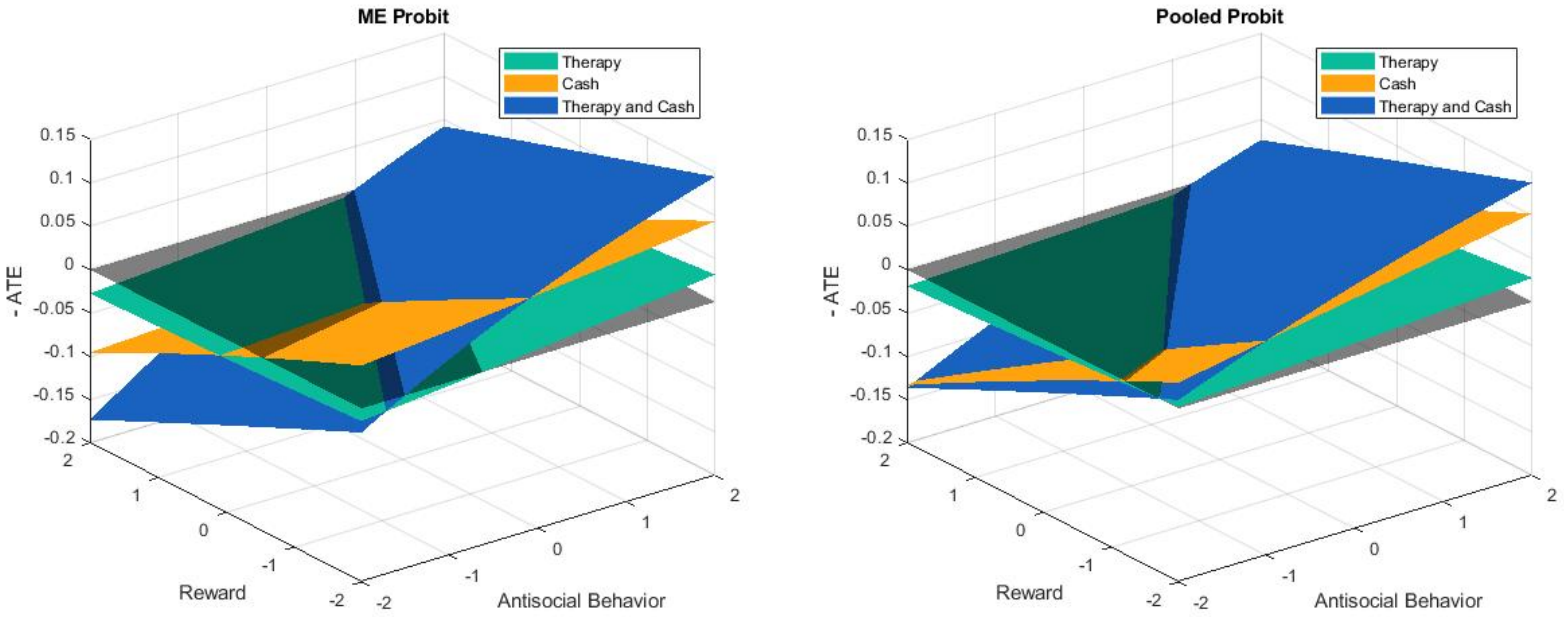
1,000 simulations of N=300. Increasing serial correlation as color darkens. Simulated 95% Confidence Intervals given with dotted lines.

Figure F.19: **ATE for Selling Drugs**



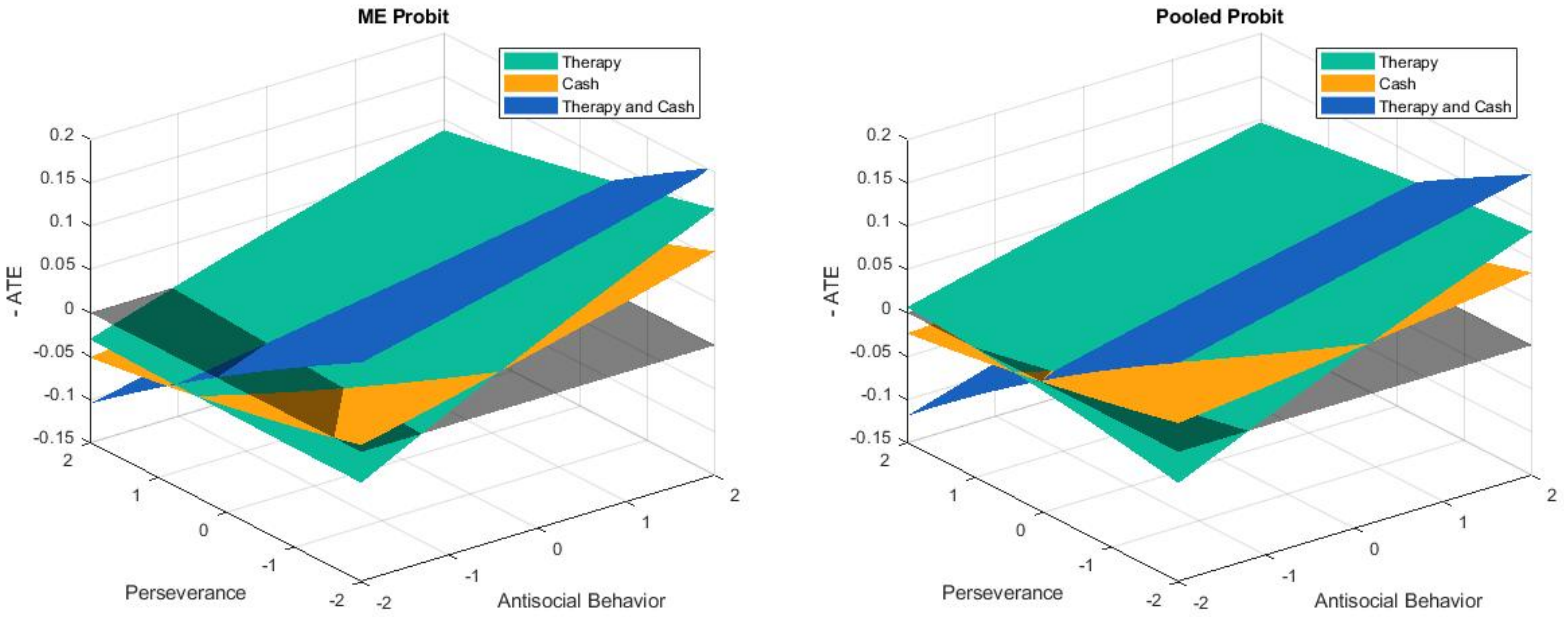
Vertical axis is the negative ATE such that the higher the ATE the better treatment it is (increasingly reduces the probability of the antisocial behavior outcomes). The transparent gray plane is flat at an ATE equal to 0, therefore any treatment effect above the plan is a desired outcome. The ATEs are calculated from a matrix of the characteristics of interest valued between [0,1] (recall that the characteristics are standardized to mean 0 and standard deviation of 1).

Figure F.20: **ATE for Being Arrested**



Vertical axis is the negative ATE such that the higher the ATE the better treatment it is (increasingly reduces the probability of the antisocial behavior outcomes). The transparent gray plane is flat at an ATE equal to 0, therefore any treatment effect above the plan is a desired outcome. The ATEs are calculated from a matrix of the characteristics of interest valued between [0,1] (recall that the characteristics are standardized to mean 0 and standard deviation of 1).

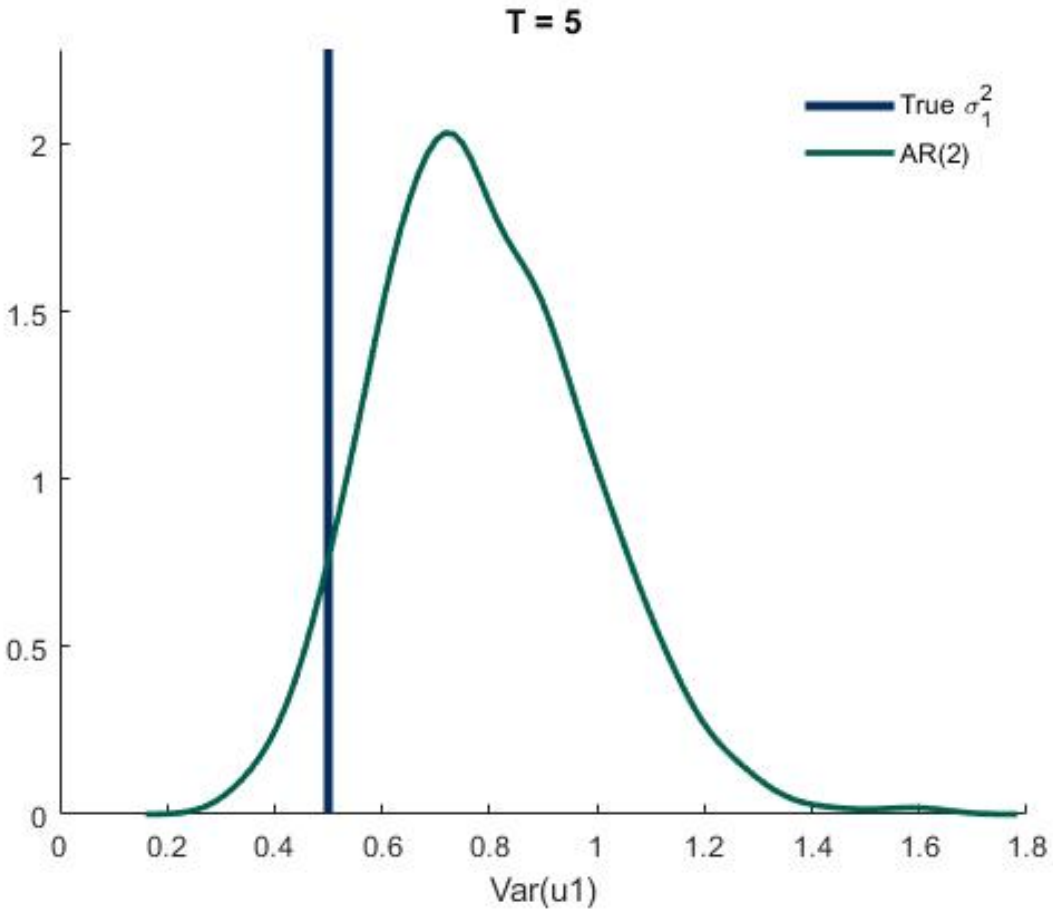
Figure F.21: **ATE for Engaging in Illicit Activities**



Vertical axis is the negative ATE such that the higher the ATE the better treatment it is (increasingly reduces the probability of the antisocial behavior outcomes). The transparent gray plane is flat at an ATE equal to 0, therefore any treatment effect above the plan is a desired outcome. The ATEs are calculated from a matrix of the characteristics of interest valued between [0,1] (recall that the characteristics are standardized to mean 0 and standard deviation of 1).

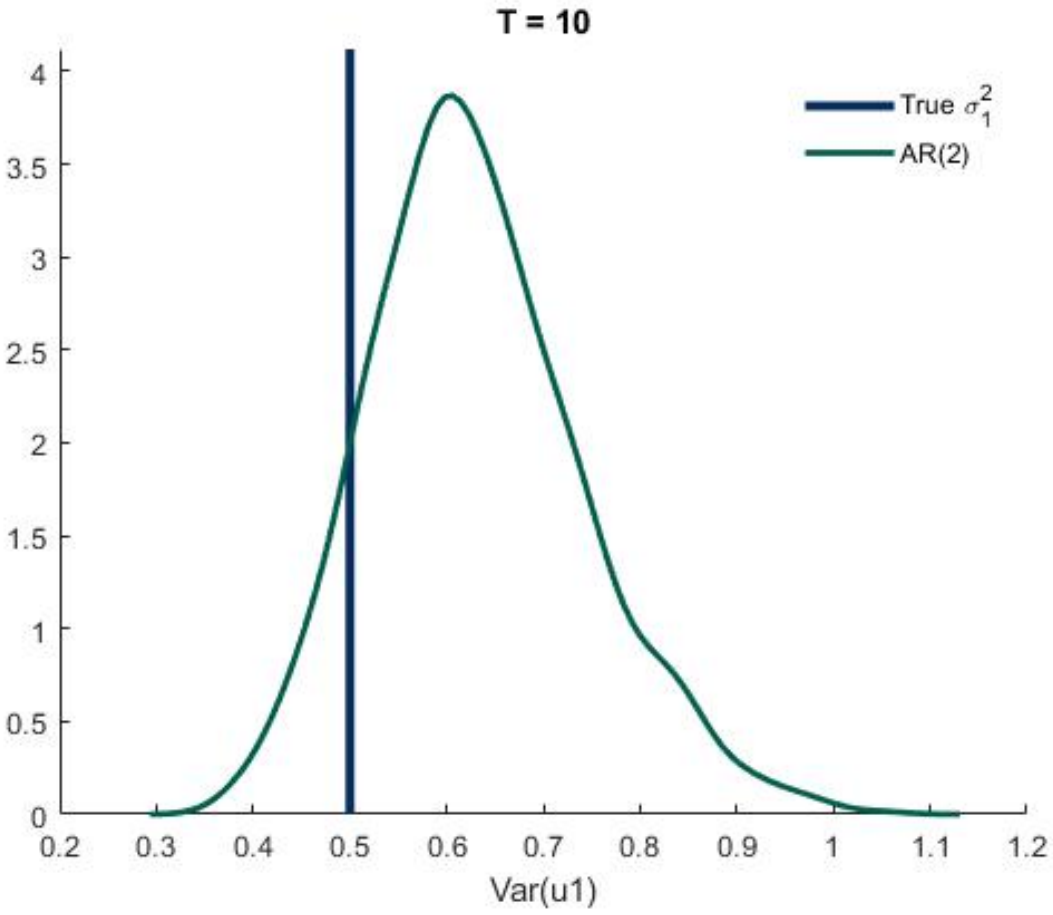
Discussion

Figure F.22: **Distribution of $\hat{\sigma}_1^2$ for T=5 under AR(2)**



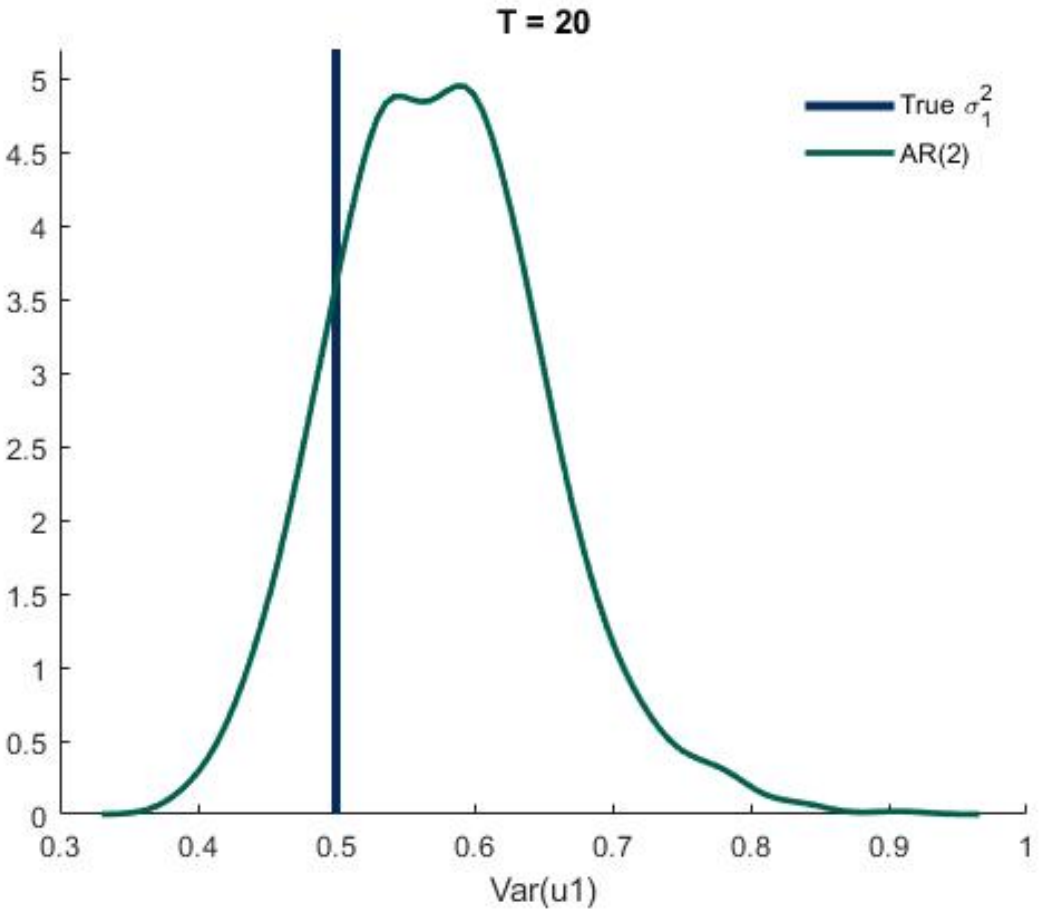
1,000 simulations of N=300.

Figure F.23: **Distribution of $\hat{\sigma}_1^2$ for T=10 under AR(2)**



1,000 simulations of N=300.

Figure F.24: **Distribution of $\hat{\sigma}_1^2$ for T=20 under AR(2)**



1,000 simulations of N=300.

APPENDIX G

Tables for Chapter 3

Simulation

Table G.1: **Estimation Times for DGP 1**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	0.696	0.656	1.581	1.845	0.708	0.668	1.617	1.852	0.678	0.651	1.950	2.407
	(0.11)	(0.10)	(0.14)	(0.25)	(0.12)	(0.13)	(0.15)	(0.14)	(0.11)	(0.11)	(0.28)	(9.03)
10	0.613	0.688	2.429	2.755	0.619	0.691	2.412	2.776	0.624	0.699	2.621	2.982
	(0.08)	(0.08)	(0.18)	(0.23)	(0.07)	(0.07)	(0.17)	(0.21)	(0.08)	(0.08)	(0.21)	(0.23)
20	0.676	0.873	4.277	4.837	0.681	0.879	4.251	4.755	0.687	0.882	4.471	4.692
	(0.06)	(0.08)	(0.36)	(0.39)	(0.06)	(0.07)	(0.34)	(0.34)	(0.07)	(0.08)	(0.42)	(0.40)

Average estimation time in seconds and standard deviations given in parenthesis. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.2: **Estimation Times for DGP 2**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	0.643	0.609	1.613	1.881	0.655	0.617	1.606	1.833	0.634	0.610	3.121	2.189
	(0.13)	(0.09)	(0.17)	(0.86)	(0.15)	(0.09)	(0.19)	(0.58)	(0.13)	(0.12)	(3.38)	(0.72)
10	0.614	0.616	2.199	2.647	0.633	0.633	2.213	2.624	0.643	0.638	2.579	2.844
	(0.09)	(0.07)	(0.23)	(0.25)	(0.10)	(0.07)	(0.23)	(0.24)	(0.11)	(0.08)	(0.31)	(0.67)
20	0.671	0.779	3.935	4.542	0.676	0.789	3.908	4.403	0.682	0.792	3.891	4.492
	(0.09)	(0.08)	(0.36)	(0.41)	(0.09)	(0.08)	(0.34)	(0.36)	(0.09)	(0.08)	(0.34)	(0.43)

Average estimation time in seconds and standard deviations given in parenthesis. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.3: **Estimation Times for DGP 3**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	0.169	0.208	1.025	1.053	0.306	0.302	1.142	1.104	0.445	0.405	1.610	1.371
	(0.03)	(0.20)	(0.25)	(0.25)	(0.05)	(0.06)	(0.11)	(0.20)	(0.08)	(0.08)	(0.81)	(0.33)
10	0.385	0.416	1.692	1.831	0.390	0.430	1.680	1.860	0.397	0.430	1.882	1.988
	(0.04)	(0.04)	(0.13)	(0.13)	(0.05)	(0.05)	(0.13)	(0.10)	(0.05)	(0.04)	(0.10)	(0.10)
20	0.411	0.523	2.970	3.108	0.416	0.527	2.928	3.043	0.428	0.539	2.954	3.174
	(0.03)	(0.03)	(0.25)	(0.21)	(0.03)	(0.03)	(0.22)	(0.19)	(0.03)	(0.03)	(0.21)	(0.20)

Average estimation time in seconds and standard deviations given in parenthesis. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.4: **Bias and Std Deviation of De-scaled ME Probit Estimates for DGP 1**

T		$\rho = 0$ (1)	$\rho = 0$ (2)	$\rho = 0.4$ (1)	$\rho = 0.4$ (2)	$\rho = 0.8$ (1)	$\rho = 0.8$ (2)
5	α	-0.8495	-0.0003	-1.0358	-0.0485	-1.8101	-0.2168
		(0.114)	(0.423)	(0.135)	(0.509)	(0.261)	(0.863)
	β_1	-0.9902	0.0698	-0.9320	0.2210	-0.6902	0.9404
		(0.080)	(0.436)	(0.092)	(0.454)	(0.140)	(0.666)
	β_2	0.0176	0.0073	0.1652	0.1439	0.7966	0.7320
		(0.078)	(0.075)	(0.094)	(0.088)	(0.189)	(0.173)
10	α	-0.8789	-0.0280	-0.9679	-0.0419	-1.3745	-0.1355
		(0.079)	(0.715)	(0.089)	(0.747)	(0.131)	(1.094)
	β_1	-0.8990	0.0439	-0.8702	0.1347	-0.7305	0.6394
		(0.056)	(0.577)	(0.058)	(0.623)	(0.076)	(0.813)
	β_2	0.0062	0.0007	0.0795	0.0734	0.4223	0.4120
		(0.047)	(0.046)	(0.051)	(0.051)	(0.077)	(0.076)
20	α	-0.8688	0.0067	-0.9107	-0.0910	-1.1243	-0.1522
		(0.059)	(1.120)	(0.065)	(1.243)	(0.084)	(1.705)
	β_1	-0.8687	0.0361	-0.8556	0.1208	-0.7811	0.3639
		(0.042)	(0.891)	(0.044)	(0.926)	(0.050)	(1.102)
	β_2	0.0020	0.0004	0.0360	0.0345	0.2212	0.2194
		(0.030)	(0.030)	(0.033)	(0.032)	(0.042)	(0.042)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25,$ and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.5: **Bias and Std Deviation of De-scaled ME Probit Estimates for DGP 2**

T		$\rho = 0$ (1)	$\rho = 0$ (2)	$\rho = 0.4$ (1)	$\rho = 0.4$ (2)	$\rho = 0.8$ (1)	$\rho = 0.8$ (2)
5	α	-0.7149	-0.0048	-0.9152	-0.0385	-1.6801	-0.2266
		(0.119)	(0.285)	(0.145)	(0.336)	(0.258)	(0.573)
	β_1	-1.2177	0.0516	-1.1591	0.2198	-0.9264	0.9672
		(0.100)	(0.311)	(0.113)	(0.365)	(0.173)	(0.538)
	β_2	0.0103	0.0107	0.1668	0.1518	0.8092	0.7461
		(0.082)	(0.079)	(0.098)	(0.093)	(0.182)	(0.165)
10	α	-0.8690	-0.0263	-0.9867	-0.0122	-1.4536	-0.1076
		(0.086)	(0.426)	(0.101)	(0.471)	(0.149)	(0.733)
	β_1	-0.9803	0.0166	-0.9423	0.1309	-0.7891	0.6271
		(0.071)	(0.375)	(0.074)	(0.421)	(0.093)	(0.562)
	β_2	0.0159	0.0058	0.0976	0.0845	0.4769	0.4544
		(0.049)	(0.048)	(0.054)	(0.053)	(0.086)	(0.083)
20	α	-0.8924	-0.0110	-0.9481	-0.0213	-1.1919	-0.0525
		(0.064)	(0.700)	(0.070)	(0.762)	(0.093)	(1.027)
	β_1	-0.9041	0.0354	-0.8816	0.0945	-0.7962	0.3712
		(0.047)	(0.562)	(0.052)	(0.591)	(0.060)	(0.706)
	β_2	0.0062	0.0020	0.0492	0.0446	0.2514	0.2464
		(0.032)	(0.032)	(0.033)	(0.033)	(0.044)	(0.043)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25,$ and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.6: **Bias and Std Deviation of De-scaled ME Probit Estimates for DGP 3**

T		$\rho = 0$ (1)	$\rho = 0$ (2)	$\rho = 0.4$ (1)	$\rho = 0.4$ (2)	$\rho = 0.8$ (1)	$\rho = 0.8$ (2)
5	α	-0.6468	-0.2268	-0.8121	-0.3008	-1.4530	-0.6283
		(0.116)	(0.451)	(0.145)	(0.523)	(0.256)	(0.894)
	β_1	-0.6229	-0.0972	-0.5174	0.0199	-0.0816	0.6856
		(0.099)	(0.474)	(0.122)	(0.522)	(0.204)	(0.752)
	β_2	-0.0332	0.0095	0.1248	0.1484	0.7717	0.7304
		(0.084)	(0.080)	(0.103)	(0.093)	(0.216)	(0.193)
10	α	-0.5603	-0.4977	-0.6190	-0.5812	-0.9248	-0.9415
		(0.077)	(0.710)	(0.084)	(0.792)	(0.127)	(1.142)
	β_1	-0.5215	-0.4936	-0.4707	-0.4200	-0.2153	-0.1168
		(0.067)	(0.640)	(0.074)	(0.680)	(0.098)	(0.846)
	β_2	-0.0094	0.0013	0.0626	0.0705	0.4171	0.4133
		(0.051)	(0.050)	(0.054)	(0.052)	(0.085)	(0.082)
20	α	-0.4860	-0.5450	-0.5140	-0.6948	-0.6467	-0.7962
		(0.061)	(1.182)	(0.064)	(1.260)	(0.084)	(1.669)
	β_1	-0.4739	-0.6094	-0.4437	-0.5018	-0.3001	-0.4328
		(0.050)	(0.958)	(0.052)	(0.970)	(0.061)	(1.226)
	β_2	-0.0004	0.0033	0.0333	0.0367	0.2138	0.2152
		(0.033)	(0.033)	(0.035)	(0.035)	(0.048)	(0.047)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25,$ and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.7: **Bias and Std Deviation of Scaled Coefficient Estimates for DGP 1**

		$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
		PHP		MEP		PHP		MEP		PHP		MEP	
T		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
	α_σ	-0.5288	0.0055	-0.6084	-0.0009	-0.5318	-0.0060	-0.6255	-0.0116	-0.5368	-0.0223	-0.6439	-0.0194
		(0.090)	(0.357)	(0.073)	(0.347)	(0.092)	(0.382)	(0.074)	(0.367)	(0.096)	(0.438)	(0.082)	(0.411)
5	$\beta_{1\sigma}$	-0.9152	0.0388	-0.8288	0.0606	-0.9164	0.0233	-0.8156	0.0435	-0.9140	0.0155	-0.7904	0.0237
		(0.086)	(0.392)	(0.057)	(0.354)	(0.088)	(0.375)	(0.058)	(0.323)	(0.088)	(0.401)	(0.054)	(0.309)
	$\beta_{2\sigma}$	-0.0751	0.0069	-0.0636	0.0091	-0.0726	0.0065	-0.0641	0.0119	-0.0704	0.0069	-0.0764	0.0090
		(0.085)	(0.087)	(0.054)	(0.058)	(0.082)	(0.087)	(0.053)	(0.056)	(0.084)	(0.085)	(0.057)	(0.058)
	α_σ	-0.6001	-0.0254	-0.6609	-0.0245	-0.6043	-0.0156	-0.6678	-0.0208	-0.6040	-0.0114	-0.6730	-0.0201
		(0.070)	(0.602)	(0.058)	(0.588)	(0.074)	(0.585)	(0.061)	(0.573)	(0.081)	(0.675)	(0.067)	(0.634)
10	$\beta_{1\sigma}$	-0.8329	0.0218	-0.7517	0.0422	-0.8318	0.0188	-0.7486	0.0424	-0.8310	0.0620	-0.7401	0.0792
		(0.055)	(0.517)	(0.042)	(0.476)	(0.056)	(0.530)	(0.041)	(0.479)	(0.056)	(0.554)	(0.041)	(0.470)
	$\beta_{2\sigma}$	-0.0388	-0.0013	-0.0453	0.0049	-0.0351	0.0027	-0.0434	0.0074	-0.0344	0.0042	-0.0483	0.0062
		(0.056)	(0.055)	(0.038)	(0.039)	(0.059)	(0.058)	(0.039)	(0.040)	(0.060)	(0.060)	(0.042)	(0.043)
	α_σ	-0.6442	-0.0124	-0.6812	0.0050	-0.6458	-0.0692	-0.6828	-0.0661	-0.6500	-0.0684	-0.6893	-0.0672
		(0.054)	(0.944)	(0.046)	(0.917)	(0.059)	(1.012)	(0.050)	(0.986)	(0.064)	(1.193)	(0.055)	(1.151)
20	$\beta_{1\sigma}$	-0.7695	0.0462	-0.7189	0.0350	-0.7706	0.0734	-0.7192	0.0661	-0.7697	0.0563	-0.7157	0.0691
		(0.043)	(0.805)	(0.033)	(0.732)	(0.041)	(0.842)	(0.034)	(0.734)	(0.042)	(0.884)	(0.033)	(0.743)
	$\beta_{2\sigma}$	-0.0174	0.0010	-0.0235	0.0046	-0.0171	0.0005	-0.0247	0.0035	-0.0140	0.0055	-0.0225	0.0075
		(0.040)	(0.040)	(0.027)	(0.027)	(0.040)	(0.040)	(0.028)	(0.027)	(0.043)	(0.042)	(0.030)	(0.030)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.8: **Bias and Std Deviation of Scaled Coefficient Estimates for DGP 2**

		$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
		PHP		MEP		PHP		MEP		PHP		MEP	
T		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
	α_σ	-0.2231	0.0042	-0.5189	-0.0034	-0.2203	0.0007	-0.5477	-0.0040	-0.2491	-0.0224	-0.5905	-0.0223
		(0.235)	(0.240)	(0.074)	(0.232)	(0.232)	(0.250)	(0.075)	(0.243)	(0.241)	(0.286)	(0.084)	(0.272)
5	$\beta_{1\sigma}$	-1.3919	0.0087	-0.9973	0.0403	-1.3901	0.0044	-0.9627	0.0374	-1.3667	0.0024	-0.8890	0.0357
		(0.241)	(0.288)	(0.075)	(0.253)	(0.244)	(0.303)	(0.073)	(0.263)	(0.247)	(0.304)	(0.067)	(0.248)
	$\beta_{2\sigma}$	-0.1331	-0.0015	-0.0577	0.0075	-0.1326	-0.0014	-0.0620	0.0124	-0.1207	0.0059	-0.0702	0.0156
		(0.140)	(0.088)	(0.060)	(0.064)	(0.143)	(0.088)	(0.060)	(0.063)	(0.140)	(0.090)	(0.063)	(0.062)
	α_σ	-0.5102	-0.0232	-0.6169	-0.0222	-0.5080	0.0036	-0.6304	0.0043	-0.5110	0.0011	-0.6458	-0.0008
		(0.065)	(0.362)	(0.057)	(0.349)	(0.069)	(0.373)	(0.063)	(0.359)	(0.080)	(0.442)	(0.068)	(0.419)
10	$\beta_{1\sigma}$	-0.9603	0.0025	-0.8227	0.0174	-0.9625	0.0160	-0.8131	0.0337	-0.9539	0.0319	-0.7907	0.0522
		(0.094)	(0.345)	(0.051)	(0.306)	(0.091)	(0.356)	(0.049)	(0.322)	(0.090)	(0.391)	(0.046)	(0.321)
	$\beta_{2\sigma}$	-0.0786	0.0039	-0.0705	0.0082	-0.0845	0.0024	-0.0753	0.0113	-0.0796	0.0047	-0.0791	0.0149
		(0.060)	(0.056)	(0.040)	(0.040)	(0.060)	(0.056)	(0.040)	(0.041)	(0.065)	(0.061)	(0.045)	(0.046)
	α_σ	-0.5803	-0.0156	-0.6620	-0.0094	-0.5802	-0.0190	-0.6670	-0.0100	-0.5803	0.0037	-0.6731	0.0044
		(0.055)	(0.585)	(0.047)	(0.574)	(0.055)	(0.626)	(0.050)	(0.600)	(0.060)	(0.714)	(0.055)	(0.681)
20	$\beta_{1\sigma}$	-0.8798	0.0228	-0.7584	0.0326	-0.8787	0.0358	-0.7526	0.0400	-0.8777	0.0294	-0.7445	0.0536
		(0.048)	(0.515)	(0.036)	(0.459)	(0.052)	(0.529)	(0.039)	(0.467)	(0.052)	(0.550)	(0.037)	(0.468)
	$\beta_{2\sigma}$	-0.0422	0.0010	-0.0534	0.0049	-0.0440	0.0002	-0.0535	0.0073	-0.0417	0.0039	-0.0549	0.0093
		(0.042)	(0.040)	(0.028)	(0.028)	(0.041)	(0.040)	(0.028)	(0.028)	(0.042)	(0.041)	(0.031)	(0.031)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.9: **Bias and Std Deviation of Scaled Coefficient Estimates for DGP 3**

		$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
		PHP		MEP		PHP		MEP		PHP		MEP	
T		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
	α_σ	-0.2446	-0.1737	-0.4033	-0.1846	-0.2453	-0.1748	-0.4163	-0.1907	-0.2485	-0.2025	-0.4216	-0.2142
		(0.090)	(0.378)	(0.071)	(0.367)	(0.090)	(0.395)	(0.073)	(0.374)	(0.096)	(0.454)	(0.079)	(0.426)
5	$\beta_{1\sigma}$	-0.5314	-0.0883	-0.5951	-0.0812	-0.5305	-0.1363	-0.5920	-0.1101	-0.5291	-0.1006	-0.5910	-0.1054
		(0.115)	(0.420)	(0.068)	(0.384)	(0.115)	(0.406)	(0.069)	(0.369)	(0.117)	(0.434)	(0.067)	(0.347)
	$\beta_{2\sigma}$	-0.2057	-0.0024	-0.1609	0.0068	-0.2075	-0.0058	-0.1585	0.0081	-0.2082	-0.0032	-0.1659	0.0023
		(0.079)	(0.090)	(0.054)	(0.063)	(0.076)	(0.085)	(0.053)	(0.060)	(0.076)	(0.084)	(0.053)	(0.059)
	α_σ	-0.2998	-0.3742	-0.3981	-0.4071	-0.2962	-0.3994	-0.3975	-0.4338	-0.3017	-0.4392	-0.4079	-0.4909
		(0.068)	(0.592)	(0.056)	(0.581)	(0.065)	(0.627)	(0.056)	(0.608)	(0.075)	(0.690)	(0.065)	(0.664)
10	$\beta_{1\sigma}$	-0.4438	-0.4378	-0.4789	-0.4021	-0.4441	-0.4221	-0.4806	-0.3842	-0.4407	-0.4056	-0.4814	-0.3591
		(0.075)	(0.549)	(0.052)	(0.522)	(0.078)	(0.575)	(0.054)	(0.522)	(0.077)	(0.568)	(0.052)	(0.493)
	$\beta_{2\sigma}$	-0.1064	-0.0033	-0.0802	0.0034	-0.1043	0.0003	-0.0805	0.0054	-0.1018	0.0052	-0.0781	0.0099
		(0.054)	(0.056)	(0.038)	(0.040)	(0.055)	(0.055)	(0.038)	(0.040)	(0.057)	(0.060)	(0.042)	(0.045)
	α_σ	-0.3181	-0.4102	-0.3739	-0.4487	-0.3192	-0.5041	-0.3754	-0.5452	-0.3152	-0.4641	-0.3739	-0.5001
		(0.053)	(1.006)	(0.048)	(0.969)	(0.054)	(1.022)	(0.048)	(1.000)	(0.059)	(1.141)	(0.054)	(1.126)
20	$\beta_{1\sigma}$	-0.3982	-0.5671	-0.4110	-0.4956	-0.3959	-0.5069	-0.4088	-0.4277	-0.3979	-0.5270	-0.4081	-0.4683
		(0.058)	(0.879)	(0.041)	(0.785)	(0.057)	(0.858)	(0.041)	(0.769)	(0.055)	(0.964)	(0.041)	(0.829)
	$\beta_{2\sigma}$	-0.0486	0.0019	-0.0314	0.0067	-0.0499	0.0010	-0.0325	0.0056	-0.0518	0.0004	-0.0339	0.0047
		(0.038)	(0.038)	(0.028)	(0.029)	(0.039)	(0.039)	(0.029)	(0.029)	(0.041)	(0.043)	(0.032)	(0.033)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.10: **Root Mean Square Error of $\hat{\beta}_{2\sigma}$ for Specification (2)**

		$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
	T	PHP	MEP	PHP	MEP	PHP	MEP
DGP 1	5	0.1552	0.1286	0.1415	0.1062	0.1611	0.0962
	10	0.2677	0.2284	0.2808	0.2311	0.3109	0.2275
	20	0.6496	0.5370	0.7146	0.5432	0.7839	0.5572
DGP 2	5	0.0832	0.0654	0.0920	0.0707	0.0926	0.0626
	10	0.1188	0.0937	0.1273	0.1049	0.1541	0.1056
	20	0.2662	0.2118	0.2811	0.2198	0.3029	0.2219
DGP 3	5	0.1845	0.1540	0.1833	0.1483	0.1985	0.1317
	10	0.4935	0.4338	0.5091	0.4200	0.4876	0.3717
	20	1.0950	0.8625	0.9926	0.7738	1.2071	0.9059

R=1,000, N=300

Table G.11: **Bias and Std Deviation of Variance Component σ_2^2 for Specification (2)**

	DGP 1			DGP 2			DGP 3		
T	$\rho = 0$	$\rho = 0.4$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.4$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.4$	$\rho = 0.8$
5	0.0050	0.0633	0.4829	-0.0165	0.1090	0.5318	0.0084	0.0458	0.3359
	(0.095)	(0.117)	(0.305)	(0.105)	(0.149)	(0.372)	(0.281)	(0.354)	(0.804)
10	-0.0023	0.0302	0.2278	-0.0064	0.0632	0.3325	-0.0506	-0.0079	0.2285
	(0.053)	(0.069)	(0.098)	(0.054)	(0.079)	(0.140)	(0.237)	(0.314)	(0.536)
20	-0.0035	0.0100	0.1113	-0.0067	0.0365	0.1913	-0.1004	-0.0695	0.0139
	(0.042)	(0.040)	(0.060)	(0.040)	(0.048)	(0.068)	(0.228)	(0.258)	(0.386)

R=1,000, N=300

Table G.12: **Bias and Std Deviation (×10) of APE Estimates for DGP 1**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	-0.6680	0.0027	-0.3953	0.0025	-0.6715	-0.0037	-0.3593	-0.0043	-0.6714	-0.0082	-0.2852	-0.0055
	(0.123)	(0.128)	(0.123)	(0.125)	(0.133)	(0.131)	(0.126)	(0.126)	(0.132)	(0.126)	(0.115)	(0.118)
10	-0.4028	0.0011	-0.1537	0.0009	-0.3999	-0.0027	-0.1408	-0.0010	-0.3983	-0.0021	-0.1127	-0.0007
	(0.102)	(0.098)	(0.096)	(0.096)	(0.102)	(0.096)	(0.092)	(0.092)	(0.101)	(0.093)	(0.089)	(0.090)
20	-0.2163	-0.0012	-0.0600	-0.0004	-0.2145	0.0002	-0.0561	0.0004	-0.2128	0.0011	-0.0454	0.0028
	(0.083)	(0.078)	(0.075)	(0.075)	(0.083)	(0.078)	(0.077)	(0.077)	(0.084)	(0.076)	(0.073)	(0.073)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is 0.0732. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.13: **Bias and Std Deviation (×10) of APE Estimates for DGP 2**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	-1.7912	0.0030	-0.8927	0.0027	-1.7813	-0.0017	-0.8139	0.0001	-1.7429	-0.0014	-0.6664	-0.0003
	(0.433)	(0.148)	(0.144)	(0.143)	(0.438)	(0.149)	(0.139)	(0.143)	(0.442)	(0.143)	(0.136)	(0.134)
10	-0.7884	-0.0013	-0.3996	-0.0012	-0.7904	-0.0019	-0.3735	0.0003	-0.7830	-0.0016	-0.3161	-0.0006
	(0.132)	(0.110)	(0.111)	(0.107)	(0.136)	(0.108)	(0.105)	(0.106)	(0.138)	(0.103)	(0.101)	(0.099)
20	-0.5333	-0.0022	-0.1712	-0.0014	-0.5307	0.0013	-0.1624	0.0027	-0.5338	-0.0011	-0.1436	0.0025
	(0.084)	(0.082)	(0.080)	(0.080)	(0.091)	(0.088)	(0.085)	(0.086)	(0.091)	(0.083)	(0.080)	(0.079)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is 0.0731. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.14: **Bias and Std Deviation (×10) of APE Estimates for DGP 3**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	-0.6549	-0.0037	-0.3342	-0.0092	-0.6575	-0.0089	-0.3137	-0.0157	-0.6564	-0.0037	-0.2646	-0.0139
	(0.136)	(0.122)	(0.119)	(0.119)	(0.141)	(0.127)	(0.120)	(0.121)	(0.143)	(0.125)	(0.111)	(0.115)
10	-0.3192	-0.0010	-0.1419	-0.0040	-0.3273	-0.0053	-0.1393	-0.0082	-0.3266	-0.0021	-0.1181	-0.0046
	(0.110)	(0.089)	(0.089)	(0.087)	(0.109)	(0.091)	(0.088)	(0.088)	(0.110)	(0.088)	(0.085)	(0.084)
20	-0.1452	-0.0039	-0.0610	-0.0032	-0.1425	-0.0013	-0.0559	-0.0003	-0.1414	-0.0010	-0.0491	-0.0003
	(0.079)	(0.071)	(0.069)	(0.068)	(0.078)	(0.070)	(0.067)	(0.0680)	(0.078)	(0.068)	(0.065)	(0.065)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is .1201. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.15: **Comparison of APE and PEA**

	DGP 1		DGP 2		DGP 3	
T	APE	PEA	APE	PEA	APE	PEA
5	0.0595	0.0719	0.0487	0.0572	0.1005	0.1147
	(0.050)	(0.056)	(0.055)	(0.065)	(0.076)	(0.090)
10	0.0640	0.0797	0.0531	0.0635	0.0986	0.1150
	(0.047)	(0.050)	(0.052)	(0.061)	(0.083)	(0.099)
20	0.0692	0.0879	0.0565	0.0682	0.0913	0.1071
	(0.045)	(0.047)	(0.052)	(0.059)	(0.086)	(0.106)

APE and PEA values using simulated data of sample size 1,000 and calculated using the true parameter values. Standard deviations of the distribution of the Partial effects given in parenthesis.

Application

Table G.16: **Select Baseline Summary Statistics**

Baseline Covariate	Therapy	Cash	Both	Control	Total
Age	25.16	25.69	25.45	25.37	25.41
	(4.82)	(5.01)	(5.09)	(4.65)	(4.88)
Married or partnered	0.152	0.133	0.198	0.143	0.155
	(0.36)	(0.40)	(0.40)	(0.35)	(0.36)
Number of children <15 in household	2.031	2.058	2.528	1.865	2.100
	(3.10)	(3.09)	(3.50)	(3.00)	(3.17)
Years of Schooling	7.576	7.832	7.766	7.647	7.702
	(3.38)	(3.29)	(3.27)	(3.06)	(3.25)
Has any disabilities	0.063	0.062	0.061	0.095	0.071
	(0.24)	(0.24)	(0.24)	(0.29)	(0.26)
Ex-Combatant	0.375	0.394	0.315	0.389	0.370
	(0.48)	(0.49)	(0.47)	(0.49)	(0.48)
Currently sleeping on street	0.228	0.252	0.244	0.258	0.246
	(0.42)	(0.44)	(0.43)	(0.44)	(0.43)
Saving Stock (US$)	33.92	28.16	41.11	27.37	32.21
	(70.16)	(63.35)	(79.37)	(47.58)	(65.39)
Hrs/week, illicit activites	15.56	12.11	14.20	13.68	13.87
	(32.54)	(23.74)	(26.08)	(27.12)	(27.60)
Hrs/week, agriculture	0.565	0.128	0.197	0.604	0.385
	(4.61)	(0.71)	(1.35)	(5.71)	(3.87)
Hrs/week, low-skill wage labor	18.44	17.88	19.09	18.97	18.59
	(30.63)	(26.91)	(28.44)	(27.22)	(28.29)
Hrs/week, low-skill business	14.19	8.657	9.515	10.26	10.67
	(28.68)	(20.92)	(20.80)	(22.41)	(23.55)
Hrs/week, illicit high skill work	1.813	1.795	0.947	1.054	1.406
	(8.15)	(8.08)	(5.10)	(6.32)	(7.07)

899 individuals, BJS provides tests of balance for select baseline covariates in Table 1

Table G.16 (cont'd)

Baseline Covariate	Therapy	Cash	Both	Control	Total
Sells Drugs	0.223	0.177	0.198	0.194	0.198
	(0.42)	(0.38)	(0.40)	(0.40)	(0.40)
Uses marijuana daily	0.464	0.465	0.416	0.484	0.459
	(0.50)	(0.50)	(0.49)	(0.50)	(0.50)
Indicator for usually Takes hard drugs	0.272	0.279	0.269	0.246	0.266
	(0.45)	(0.45)	(0.44)	(0.43)	(0.44)
Uses hard drugs daily	0.134	0.177	0.157	0.115	0.145
	(0.34)	(0.38)	(0.36)	(0.32)	(0.35)
Committed theft, past 2 wks	0.576	0.540	0.533	0.579	0.558
	(0.49)	(0.50)	(0.50)	(0.49)	(0.50)
Antisocial behavior index	-0.066	0.078	-0.036	0.035	0.005
	(0.962)	(1.090)	(1.048)	(1.084)	(1.050)
Perseverance Index	-0.026	-0.036	-0.009	0.008	-0.015
	(1.02)	(1.12)	(1.09)	(1.05)	(1.07)
Reward Responsiveness	-0.066	0.111	0.080	0.020	0.035
	(1.05)	(1.08)	(1.07)	(1.03)	(1.06)
Impulsiveness Index	-0.085	0.019	-0.004	0.109	0.013
	(1.05)	(1.09)	(1.07)	(1.04)	(1.07)

899 individuals, BJS provides tests of balance for select baseline covariates in Table 1

Table G.17: **Preliminary OLS Estimates**

	Sells Drugs				Arrested			
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Treatment								
Therapy	-0.0558	(0.023)	-0.0559	(0.023)	-0.0010	(0.018)	-0.0003	(0.019)
Cash	-0.0087	(0.024)	-0.0068	(0.024)	0.0044	(0.019)	0.0125	(0.019)
Both	-0.0724	(0.022)	-0.0642	(0.023)	-0.0195	(0.019)	-0.0153	(0.019)
Interact with Therapy								
Bad Behavior	-0.0522	(0.024)	-0.0541	(0.024)	-0.0099	(0.021)	-0.0064	(0.021)
Perseverance	0.0052	(0.021)	0.0007	(0.021)	-0.0181	(0.017)	-0.0199	(0.017)
Reward	-0.0032	(0.025)	-0.0065	(0.025)	0.0049	(0.016)	0.0007	(0.016)
Impulsiveness	-0.0179	(0.021)	-0.0198	(0.021)	-0.0052	(0.018)	-0.0061	(0.019)
Interact with Cash								
Bad Behavior	-0.0280	(0.026)	-0.0201	(0.027)	-0.0021	(0.023)	0.0097	(0.022)
Perserverence	0.0377	(0.021)	0.0387	(0.021)	-0.0238	(0.017)	-0.0173	(0.017)
Reward	0.0068	(0.021)	0.0032	(0.022)	0.0305	(0.017)	0.0251	(0.018)
Impulsiveness	0.0292	(0.022)	0.0155	(0.023)	-0.0012	(0.020)	-0.0121	(0.020)
Interact with Both								
Bad Behavior	-0.0708	(0.023)	-0.0729	(0.024)	-0.0480	(0.023)	-0.0561	(0.023)
Perserverence	0.0380	(0.019)	0.0260	(0.020)	0.0166	(0.017)	0.0167	(0.018)
Reward	-0.0080	(0.019)	-0.0145	(0.020)	0.0244	(0.017)	0.0187	(0.017)
Impulsiveness	0.0116	(0.019)	0.0007	(0.020)	0.0038	(0.018)	0.0027	(0.019)
Includes Block FE	No		Yes		No		Yes	
Number of Individuals	890		890		890		890	
Number of Observations	3,312		3,312		3,312		3,312	

Standard errors for both estimators are robust and clustered at the individual level. Treatment is randomly assigned within Blocks.

Table G.17 (cont'd)

	Illicit Activity			
	Coeff	SE	Coeff	SE
Treatment				
Therapy	-0.0575	(0.021)	-0.0565	(0.022)
Cash	-0.0204	(0.021)	-0.0163	(0.021)
Both	-0.0622	(0.021)	-0.0570	(0.022)
Interact with Therapy				
Bad Behavior	-0.0642	(0.024)	-0.0678	(0.024)
Perseverance	0.0094	(0.019)	0.0043	(0.019)
Reward	-0.0010	(0.023)	-0.0043	(0.023)
Impulsiveness	0.0039	(0.020)	0.0064	(0.019)
Interact with Cash				
Bad Behavior	-0.0271	(0.023)	-0.0270	(0.024)
Perseverance	0.0253	(0.018)	0.0253	(0.018)
Reward	0.0043	(0.018)	-0.0028	(0.019)
Impulsiveness	0.0175	(0.019)	0.0091	(0.019)
Interact with Both				
Bad Behavior	-0.0567	(0.024)	-0.0671	(0.025)
Perseverance	0.0495	(0.020)	0.0339	(0.020)
Reward	-0.0092	(0.018)	-0.0129	(0.019)
Impulsiveness	0.0028	(0.019)	-0.0031	(0.019)
Includes Block FE	No		Yes	
Number of Individuals	890		890	
Number of Observations	3,328		3,328	

Standard errors for both estimators are robust and clustered at the individual level. Treatment is randomly assigned within Blocks.

Table G.18: **Scaled Probit Coefficient Estimates for Selling Drugs**

	ME Probit				Pooled Probit			
	RE		CRE		RE		CRE	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Treatment								
Therapy	-0.760	(0.24)	-0.740	(0.26)	-0.679	(0.55)	-0.233	(0.49)
Cash	-0.086	(0.12)	-0.080	(0.13)	0.303	(0.21)	0.494	(0.20)
Both	-0.392	(0.12)	-0.337	(0.13)	-0.513	(0.55)	0.104	(0.45)
Therapy ×								
Antisocial			-0.155	(0.16)			-0.198	(0.13)
Perseverance			-0.140	(0.16)			-0.041	(0.13)
Reward			0.127	(0.18)			0.046	(0.15)
Impulsiveness			-0.176	(0.16)			-0.081	(0.13)
Cash ×								
Antisocial			-0.146	(0.12)			-0.223	(0.10)
Perseverance			0.213	(0.12)			0.191	(0.10)
Reward			0.106	(0.12)			0.014	(0.09)
Impulsiveness			0.114	(0.12)			0.060	(0.08)
Both ×								
Antisocial			-0.307	(0.13)			-0.279	(0.11)
Perseverance			0.237	(0.13)			0.229	(0.11)
Reward			-0.055	(0.12)			-0.073	(0.10)
Impulsiveness			0.081	(0.12)			0.074	(0.09)
Var. Components								
Therapy	1.621	(0.92)	1.550	(0.89)				
Cash	0.000	(0.00)	0.000	(0.00)				
Both	0.000	(0.00)	0.000	(0.00)				
Intercept	0.838	(0.19)	0.786	(0.18)				
Het. Coefficients								
Therapy					0.256	(0.33)	-0.046	(0.37)
Cash					-0.546	(0.33)	-0.819	(0.37)
Both					0.078	(0.37)	-0.431	(0.48)
Time (in Seconds)	11369.869		11973.093		1.029		2.208	

3,312 total observations for 890 individuals, in which dummies for the number of time observations for each individual are included to address the unbalanced panel. Standard errors for both estimators are robust and clustered at the individual level. ME Probit coefficient estimates are scaled by $(1/\sqrt{1+\sigma_a^2})$ and standard errors are calculated using the delta method.

Table G.19: **Scaled Probit Coefficient Estimates for being Arrested**

	ME Probit				Pooled Probit			
	RE		CRE		RE		CRE	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Treatment								
Therapy	-0.020	(0.13)	-0.019	(0.14)	-0.013	(0.32)	-0.060	(0.47)
Cash	-0.031	(0.15)	-0.098	(0.16)	0.397	(0.21)	0.335	(0.28)
Both	-0.305	(0.17)	-0.222	(0.17)	-1.185	(0.87)	-0.730	(0.87)
Therapy ×								
Antisocial			-0.055	(0.10)			-0.028	(0.12)
Perseverance			-0.099	(0.10)			-0.097	(0.12)
Reward			0.027	(0.10)			0.032	(0.10)
Impulsiveness			-0.039	(0.10)			-0.045	(0.10)
Cash ×								
Antisocial			-0.046	(0.11)			-0.088	(0.10)
Perseverance			-0.182	(0.11)			-0.109	(0.10)
Reward			0.226	(0.11)			0.180	(0.10)
Impulsiveness			-0.021	(0.10)			-0.008	(0.09)
Both ×								
Antisocial			-0.308	(0.13)			-0.286	(0.16)
Perseverance			0.072	(0.11)			0.072	(0.14)
Reward			0.187	(0.11)			0.229	(0.16)
Impulsiveness			0.040	(0.11)			0.052	(0.13)
Var. Components								
Therapy	0.062	(0.20)	0.068	(0.21)				
Cash	0.177	(0.26)	0.190	(0.25)				
Both	0.522	(0.31)	0.334	(0.27)				
Intercept	0.156	(0.14)	0.149	(0.15)				
Het. Coefficients								
Therapy					0.006	(0.28)	0.047	(0.39)
Cash					-0.455	(0.30)	-0.403	(0.36)
Both					0.646	(0.37)	0.431	(0.46)
Time (in Seconds)	818.475		827.293		4.746		5.988	

3,302 total observations for 880 individuals, in which dummies for the number of time observations for each individual are included to address the unbalanced panel. Standard errors for both estimators are robust and clustered at the individual level. ME Probit coefficient estimates are scaled by $(1/\sqrt{1+\sigma_a^2})$ and standard errors are calculated using the delta method.

Table G.20: **Scaled Probit Coefficient Estimates for Illicit Activity**

	ME Probit				Pooled Probit			
	RE		CRE		RE		CRE	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Treatment								
Therapy	-0.972	(0.31)	-0.877	(0.30)	-2.003	(1.65)	-1.242	(1.36)
Cash	-0.101	(0.12)	-0.095	(0.12)	-0.498	(0.46)	-0.356	(0.55)
Both	-0.704	(0.27)	-0.538	(0.25)	-1.320	(1.72)	-0.267	(0.90)
Therapy ×								
Antisocial			-0.309	(0.17)			-0.288	(0.19)
Perseverance			-0.025	(0.17)			-0.147	(0.23)
Reward			0.137	(0.18)			0.153	(0.27)
Impulsiveness			0.002	(0.16)			-0.028	(0.20)
Cash ×								
Antisocial			-0.118	(0.11)			-0.051	(0.17)
Perseverance			0.105	(0.11)			0.084	(0.13)
Reward			0.121	(0.11)			0.117	(0.14)
Impulsiveness			0.007	(0.11)			0.035	(0.12)
Both ×								
Antisocial			-0.157	(0.14)			-0.167	(0.15)
Perseverance			0.382	(0.15)			0.342	(0.17)
Reward			-0.087	(0.13)			-0.081	(0.12)
Impulsiveness			-0.017	(0.13)			-0.013	(0.13)
Var. Components								
Therapy	1.988	(0.98)	1.761	(0.89)				
Cash	0.000	(0.00)	0.000	(0.00)				
Both	1.018	(0.66)	0.568	(0.55)				
Intercept	0.422	(0.15)	0.461	(0.17)				
Het. Coefficients								
Therapy					0.830	(0.54)	0.542	(0.58)
Cash					0.327	(0.30)	0.205	(0.37)
Both					0.557	(0.74)	-0.075	(0.71)
Time (in Seconds)	4642.346		6735.654		1.063		1.872	

3,320 total observations for 882 individuals, in which dummies for the number of time observations for each individual are included to address the unbalanced panel. SStandard errors for both estimators are robust and clustered at the individual level. ME Probit coefficient estimates are scaled by $(1/\sqrt{1+\sigma_a^2})$ and standard errors are calculated using the delta method.

Table G.21: **ATE Estimates**

			Therapy Only		Cash Only		Both Cash and Therapy	
			Coeff	SE	Coeff	SE	Coeff	SE
Sells Drugs	OLS	(1)	-0.0569	(0.023)	-0.0074	(0.024)	-0.0704	(0.023)
		(2)	-0.0573	(0.023)	-0.0095	(0.024)	-0.0747	(0.023)
	ME	(1)	-0.0621	(0.033)	-0.0162	(0.023)	-0.0653	(0.021)
		(2)	-0.0648	(0.033)	-0.0207	(0.023)	-0.0764	(0.021)
	PHP	(1)	-0.0638	(0.019)	-0.0202	(0.020)	-0.0686	(0.019)
		(2)	-0.0642	(0.031)	0.0065	(0.059)	-0.0495	(0.055)
Arrested	OLS	(1)	-0.0006	(0.018)	0.0082	(0.019)	-0.0187	(0.019)
		(2)	-0.0005	(0.018)	0.0059	(0.019)	-0.0201	(0.019)
	ME	(1)	0.0011	(0.024)	0.0077	(0.025)	-0.0140	(0.027)
		(2)	0.0021	(0.025)	0.0072	(0.024)	-0.0163	(0.026)
	PHP	(1)	-0.0012	(0.018)	-0.0035	(0.019)	-0.0363	(0.020)
		(2)	0.0002	(0.019)	-0.0001	(0.020)	-0.0298	(0.020)
Illicit Activity	OLS	(1)	-0.0563	(0.021)	-0.0180	(0.021)	-0.0600	(0.021)
		(2)	-0.0587	(0.022)	-0.0211	(0.021)	-0.0649	(0.022)
	ME	(1)	-0.0536	(0.034)	-0.0174	(0.020)	-0.0549	(0.029)
		(2)	-0.0597	(0.034)	-0.0219	(0.021)	-0.0639	(0.029)
	PHP	(1)	-0.0238	(0.034)	0.0049	(0.027)	-0.0246	(0.034)
		(2)	-0.0473	(0.026)	-0.0112	(0.021)	-0.0533	(0.018)

Specification (1) assumes a RE structure such that the random treatment effects are not heterogeneous in terms of the individual characteristics. Specification (2) implements a flexible CRE specification that allows the treatment effects to be heterogeneous in individual characteristics.

Table G.22: **Bias and Std Deviation of Scaled Coefficient Estimates under AR(2)**

T		PHP (1)	PHP (2)	MEP (1)	MEP (2)
5	α_σ	-0.5520	-0.0159	-0.6439	-0.0193
		(0.091)	(0.388)	(0.076)	(0.375)
	$\beta_{1\sigma}$	-0.9095	0.0398	-0.8112	0.0659
		(0.087)	(0.397)	(0.057)	(0.343)
	$\beta_{2\sigma}$	-0.0542	0.0284	-0.0441	0.0358
		(0.083)	(0.088)	(0.056)	(0.059)
10	α_σ	-0.6284	-0.0580	-0.6898	-0.0647
		(0.074)	(0.592)	(0.061)	(0.573)
	$\beta_{1\sigma}$	-0.8229	0.0510	-0.7401	0.0695
		(0.055)	(0.531)	(0.041)	(0.484)
	$\beta_{2\sigma}$	-0.0115	0.0297	-0.0221	0.0321
		(0.059)	(0.058)	(0.039)	(0.040)
20	α_σ	-0.6678	-0.0247	-0.7055	-0.0063
		(0.060)	(1.031)	(0.053)	(1.002)
	$\beta_{1\sigma}$	-0.7616	0.1005	-0.7095	0.0918
		(0.043)	(0.821)	(0.034)	(0.718)
	$\beta_{2\sigma}$	0.0057	0.0255	-0.0016	0.0285
		(0.043)	(0.043)	(0.030)	(0.030)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.23: **Bias and Std Deviation (×10) of APE Estimates under AR(2)**

	PHP		MEP	
T	(1)	(2)	(1)	(2)
5	-0.6657	0.0024	-0.3645	0.0042
	(0.1284)	(0.1274)	(0.1238)	(0.1228)
10	-0.3966	0.0044	-0.1409	0.0058
	(0.0987)	(0.0936)	(0.0898)	(0.0914)
20	-0.2096	0.0059	-0.0510	0.0072
	(0.0820)	(0.0768)	(0.0743)	(0.0750)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is 0.0735. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.24: **Failure Count under no Random Coefficients**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	0	6	525	448	0	4	6	8	0	7	0	1
10	0	0	559	567	0	0	0	4	0	0	0	0
20	0	0	652	721	0	0	4	4	0	0	0	0

Specification (1) assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.25: **Estimation Times under no Random Coefficients**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	0.301	0.464	6.826	10.056	0.389	0.525	3.681	5.442	0.478	0.613	4.269	5.592
	(0.06)	(0.13)	(3.90)	(5.06)	(0.08)	(0.15)	(2.41)	(3.15)	(0.08)	(0.26)	(1.78)	(2.00)
10	0.538	0.695	7.991	11.219	0.533	0.686	5.078	7.038	0.532	0.692	7.455	9.362
	(0.10)	(0.14)	(4.69)	(6.23)	(0.08)	(0.14)	(3.39)	(4.22)	(0.08)	(0.13)	(3.18)	(3.78)
20	0.605	0.845	13.688	18.954	0.618	0.863	8.532	11.741	0.629	0.878	9.070	10.434
	(0.07)	(0.15)	(8.29)	(10.54)	(0.08)	(0.16)	(6.13)	(7.63)	(0.08)	(0.16)	(6.56)	(8.16)

Average estimation time in seconds and standard deviations given in parenthesis. Specification (1) assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.26: **Bias and Std Deviation (×10) of APE Estimates under no Random Coefficients**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	-0.0023	-0.0005	-0.0001	0.0016	0.0005	0.0052	0.0013	0.0057	0.0045	0.0044	0.0018	0.0024
	(0.084)	(0.092)	(0.081)	(0.088)	(0.093)	(0.099)	(0.089)	(0.095)	(0.094)	(0.098)	(0.087)	(0.090)
10	0.0016	0.0006	0.0031	0.0029	-0.0005	-0.0005	0.0005	0.0013	0.0017	0.0025	0.0008	0.0018
	(0.058)	(0.060)	(0.057)	(0.059)	(0.059)	(0.063)	(0.059)	(0.061)	(0.0672)	(0.068)	(0.065)	(0.066)
20	-0.0035	-0.0036	-0.0027	-0.0025	0.0018	0.0022	0.0022	0.0028	-0.0008	-0.0005	-0.0008	-0.0004
	(0.040)	(0.040)	(0.040)	(0.041)	(0.044)	(0.045)	(0.044)	(0.044)	(0.0438)	(0.044)	(0.043)	(0.044)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is 0.1511. Specification (1) assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.27: **Variance Component σ_1^2 Estimates under no Random Coefficients**

	$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
T	(1)	(2)	(1)	(2)	(1)	(2)
5	0.0556	0.0463	0.2944	0.2771	2.1492	2.1260
	(0.064)	(0.059)	(0.154)	(0.153)	(0.916)	(0.944)
10	0.0232	0.0186	0.1456	0.1366	0.9861	0.9650
	(0.026)	(0.024)	(0.060)	(0.060)	(0.205)	(0.203)
20	0.0117	0.0092	0.0688	0.0642	0.4779	0.4667
	(0.012)	(0.011)	(0.026)	(0.026)	(0.077)	(0.076)

Specification (1) assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.28: **Variance Component σ_2^2 Estimates under no Random Coefficients**

	$\rho = 0$		$\rho = 0.4$		$\rho = 0.8$	
T	(1)	(2)	(1)	(2)	(1)	(2)
5	0.0164	0.0107	0.0383	0.0268	0.0808	0.0524
	(0.041)	(0.033)	(0.063)	(0.055)	(0.188)	(0.160)
10	0.0106	0.0076	0.0245	0.0178	0.0337	0.0260
	(0.024)	(0.021)	(0.037)	(0.033)	(0.072)	(0.063)
20	0.0065	0.0047	0.0134	0.0102	0.0324	0.0282
	(0.013)	(0.011)	(0.019)	(0.017)	(0.039)	(0.035)

Specification (1) assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.29: **Rejection Rate of LR Test for Random Coefficients**

T	$\rho = 0$ (1)	$\rho = 0$ (2)	$\rho = 0.4$ (1)	$\rho = 0.4$ (2)	$\rho = 0.8$ (1)	$\rho = 0.8$ (2)
5	0.054	0.033	0.723	0.638	1	1
10	0.047	0.026	0.858	0.799	1	1
20	0.055	0.027	0.881	0.841	1	1

True value of σ_1^2 is 0.5. Specification (1) assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages.

Table G.30: **Bias and Std Deviation of De-scaled ME Logit Estimate under a Conditional Logistic AR(1) Process**

T		$\rho = 0$ (1)	$\rho = 0$ (2)	$\rho = 0.4$ (1)	$\rho = 0.4$ (2)	$\rho = 0.8$ (1)	$\rho = 0.8$ (2)
5	α	-0.7919	-0.0614	-1.0504	-0.0641	-1.9744	-0.2474
		(0.129)	(0.537)	(0.164)	(0.701)	(0.310)	(1.202)
	β_1	-1.0646	0.0599	-0.9749	0.2782	-0.6782	1.1248
		(0.100)	(0.523)	(0.113)	(0.586)	(0.159)	(0.895)
	β_2	-0.0034	0.0030	0.1738	0.1705	0.8856	0.8505
		(0.090)	(0.091)	(0.102)	(0.099)	(0.192)	(0.187)
10	α	-0.8592	-0.0628	-0.9699	0.0248	-1.4747	-0.1786
		(0.089)	(0.822)	(0.108)	(1.013)	(0.173)	(1.599)
	β_1	-0.9314	0.0007	-0.8845	0.1203	-0.7167	0.6357
		(0.066)	(0.688)	(0.070)	(0.689)	(0.087)	(0.980)
	β_2	0.0034	0.0030	0.0867	0.0854	0.5012	0.4960
		(0.061)	(0.061)	(0.065)	(0.065)	(0.093)	(0.093)
20	α	-0.8660	0.0009	-0.9119	-0.0367	-1.1741	-0.1120
		(0.067)	(1.397)	(0.077)	(1.486)	(0.113)	(2.326)
	β_1	-0.8784	0.0217	-0.8596	0.0820	-0.7661	0.3829
		(0.049)	(1.045)	(0.051)	(1.094)	(0.061)	(1.324)
	β_2	0.0020	0.0019	0.0441	0.0437	0.2643	0.2633
		(0.040)	(0.039)	(0.042)	(0.042)	(0.054)	(0.054)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages. Conditional logistic AR(1) is generated according to equation (3.35)

Table G.31: **Bias and Std Deviation of De-scaled ME Logit Estimate under a Marginal Logistic AR(1) Process**

T		$\rho = 0$ (1)	$\rho = 0$ (2)	$\rho = 0.4$ (1)	$\rho = 0.4$ (2)	$\rho = 0.8$ (1)	$\rho = 0.8$ (2)
5	α	-0.7962	-0.0541	-1.0212	-0.0769	-1.8377	-0.2540
		(0.129)	(0.538)	(0.167)	(0.690)	(0.276)	(1.259)
	β_1	-1.0646	0.0399	-0.9827	0.2104	-0.7226	0.9752
		(0.098)	(0.508)	(0.104)	(0.578)	(0.153)	(0.818)
	β_2	0.0057	0.0119	0.1510	0.1480	0.7812	0.7508
		(0.086)	(0.087)	(0.106)	(0.106)	(0.177)	(0.171)
10	α	-0.8561	-0.0393	-0.9424	-0.1014	-1.3900	-0.1505
		(0.095)	(0.818)	(0.106)	(0.995)	(0.168)	(1.688)
	β_1	-0.9296	0.0449	-0.8944	0.1266	-0.7464	0.5626
		(0.063)	(0.659)	(0.069)	(0.750)	(0.088)	(0.961)
	β_2	0.0017	0.0010	0.0633	0.0618	0.4205	0.4153
		(0.059)	(0.059)	(0.061)	(0.061)	(0.086)	(0.085)
20	α	-0.8632	0.0045	-0.8933	-0.0981	-1.0903	-0.0510
		(0.070)	(1.369)	(0.079)	(1.510)	(0.109)	(2.344)
	β_1	-0.8780	0.0646	-0.8695	0.0737	-0.7932	0.3125
		(0.050)	(1.054)	(0.052)	(1.093)	(0.057)	(1.285)
	β_2	0.0004	0.0003	0.0234	0.0232	0.1897	0.1891
		(0.039)	(0.040)	(0.039)	(0.040)	(0.049)	(0.049)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages. Marginal logistic AR(1) is generated according to equation (3.36)

Table G.32: **Bias and Std Deviation of Scaled Coefficient Estimates under a Conditional Logistic AR(1) Process**

		$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
		PHP		MEL		PHP		MEL		PHP		MEL	
T		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
	α_σ	-0.3808	-0.0389	-0.3876	-0.0321	-0.3824	-0.0172	-0.4311	-0.0135	-0.3783	-0.0183	-0.4683	-0.0176
		(0.078)	(0.305)	(0.060)	(0.277)	(0.080)	(0.340)	(0.064)	(0.318)	(0.086)	(0.383)	(0.072)	(0.352)
5	$\beta_{1\sigma}$	-0.5884	0.0664	-0.5505	0.0321	-0.5873	0.0750	-0.5239	0.0512	-0.5858	0.0583	-0.4887	0.0528
		(0.081)	(0.323)	(0.049)	(0.270)	(0.081)	(0.317)	(0.048)	(0.265)	(0.074)	(0.339)	(0.042)	(0.257)
	$\beta_{2\sigma}$	0.0035	0.0333	-0.0197	0.0026	0.0024	0.0374	-0.0081	0.0176	0.0002	0.0382	-0.0073	0.0282
		(0.074)	(0.074)	(0.044)	(0.047)	(0.074)	(0.072)	(0.042)	(0.043)	(0.077)	(0.073)	(0.044)	(0.046)
	α_σ	-0.4309	-0.0447	-0.4267	-0.0321	-0.4268	0.0117	-0.4410	0.0198	-0.4238	-0.0292	-0.4696	-0.0269
		(0.058)	(0.460)	(0.043)	(0.423)	(0.064)	(0.531)	(0.049)	(0.489)	(0.071)	(0.629)	(0.059)	(0.579)
10	$\beta_{1\sigma}$	-0.5172	0.0429	-0.4826	0.0018	-0.5162	0.0557	-0.4715	0.0190	-0.5210	0.0459	-0.4572	0.0414
		(0.045)	(0.408)	(0.033)	(0.353)	(0.046)	(0.389)	(0.032)	(0.332)	(0.044)	(0.458)	(0.030)	(0.356)
	$\beta_{2\sigma}$	0.0270	0.0386	-0.0114	0.0029	0.0264	0.0384	-0.0064	0.0101	0.0264	0.0404	0.0069	0.0283
		(0.050)	(0.051)	(0.030)	(0.031)	(0.050)	(0.050)	(0.029)	(0.030)	(0.049)	(0.049)	(0.031)	(0.032)
	α_σ	-0.4536	-0.0053	-0.4367	0.0003	-0.4502	-0.0185	-0.4409	-0.0146	-0.4511	-0.0260	-0.4619	-0.0254
		(0.043)	(0.777)	(0.034)	(0.720)	(0.045)	(0.803)	(0.037)	(0.741)	(0.055)	(1.034)	(0.047)	(0.985)
20	$\beta_{1\sigma}$	-0.4716	0.0398	-0.4539	0.0125	-0.4721	0.0525	-0.4508	0.0221	-0.4737	0.0681	-0.4415	0.0484
		(0.033)	(0.603)	(0.025)	(0.538)	(0.034)	(0.620)	(0.025)	(0.546)	(0.036)	(0.654)	(0.025)	(0.559)
	$\beta_{2\sigma}$	0.0331	0.0371	-0.0062	0.0021	0.0333	0.0373	-0.0020	0.0066	0.0328	0.0375	0.0104	0.0209
		(0.032)	(0.032)	(0.020)	(0.020)	(0.031)	(0.032)	(0.020)	(0.020)	(0.034)	(0.033)	(0.023)	(0.023)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages. Conditional logistic AR(1) is generated according to equation (3.35)

Table G.33: **Bias and Std Deviation of Scaled Coefficient Estimates under a Marginal Logistic AR(1) Process**

		$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
		PHP		MEL		PHP		MEL		PHP		MEL	
T		(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
	α_σ	-0.3855	-0.0365	-0.3901	-0.0280	-0.3738	-0.0261	-0.4223	-0.0211	-0.3639	-0.0273	-0.4549	-0.0262
		(0.078)	(0.304)	(0.059)	(0.277)	(0.084)	(0.338)	(0.067)	(0.314)	(0.087)	(0.406)	(0.069)	(0.380)
5	$\beta_{1\sigma}$	-0.5856	0.0654	-0.5503	0.0226	-0.5854	0.0458	-0.5263	0.0244	-0.5894	0.0331	-0.4948	0.0333
		(0.079)	(0.307)	(0.048)	(0.262)	(0.078)	(0.310)	(0.045)	(0.263)	(0.079)	(0.325)	(0.042)	(0.246)
	$\beta_{2\sigma}$	0.0083	0.0428	-0.0149	0.0077	-0.0064	0.0292	-0.0147	0.0102	-0.0123	0.0219	-0.0157	0.0175
		(0.072)	(0.073)	(0.041)	(0.044)	(0.075)	(0.073)	(0.043)	(0.046)	(0.074)	(0.069)	(0.043)	(0.043)
	α_σ	-0.4298	-0.0279	-0.4249	-0.0199	-0.4183	-0.0521	-0.4311	-0.0414	-0.4062	-0.0410	-0.4524	-0.0209
		(0.060)	(0.457)	(0.046)	(0.421)	(0.059)	(0.521)	(0.047)	(0.483)	(0.069)	(0.655)	(0.058)	(0.623)
10	$\beta_{1\sigma}$	-0.5160	0.0626	-0.4818	0.0243	-0.5190	0.0652	-0.4752	0.0245	-0.5265	0.0626	-0.4638	0.0274
		(0.045)	(0.397)	(0.031)	(0.339)	(0.047)	(0.426)	(0.032)	(0.363)	(0.045)	(0.438)	(0.031)	(0.353)
	$\beta_{2\sigma}$	0.0271	0.0369	-0.0125	0.0016	0.0187	0.0294	-0.0147	0.0005	0.0066	0.0193	-0.0105	0.0088
		(0.048)	(0.048)	(0.029)	(0.029)	(0.046)	(0.046)	(0.027)	(0.028)	(0.049)	(0.047)	(0.029)	(0.030)
	α_σ	-0.4515	-0.0001	-0.4351	0.0024	-0.4428	-0.0584	-0.4336	-0.0458	-0.4230	-0.0117	-0.4350	-0.0006
		(0.044)	(0.760)	(0.034)	(0.705)	(0.046)	(0.818)	(0.038)	(0.754)	(0.054)	(1.054)	(0.046)	(1.004)
20	$\beta_{1\sigma}$	-0.4715	0.0620	-0.4538	0.0338	-0.4763	0.0514	-0.4550	0.0197	-0.4820	0.0491	-0.4501	0.0270
		(0.033)	(0.612)	(0.026)	(0.542)	(0.034)	(0.617)	(0.025)	(0.545)	(0.033)	(0.640)	(0.024)	(0.550)
	$\beta_{2\sigma}$	0.0323	0.0364	-0.0073	0.0009	0.0242	0.0284	-0.0106	-0.0022	0.0072	0.0117	-0.0136	-0.0040
		(0.032)	(0.032)	(0.019)	(0.020)	(0.031)	(0.030)	(0.019)	(0.019)	(0.033)	(0.033)	(0.021)	(0.022)

R=1,000, N=300. Standard Deviations are given in parenthesis. The true coefficient values are $\alpha = -0.25, \beta_1 = 1.25$, and $\beta_2 = 1$. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages. Marginal logistic AR(1) is generated according to equation (3.36)

Table G.34: **Bias and Std Deviation (×10) of APE Estimates under a Conditional Logistic AR(1) Process**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	-0.4665	0.1036	-0.2912	0.0033	-0.4681	0.0985	-0.2173	-0.0030	-0.4643	0.1040	-0.1036	0.0051
	(0.155)	(0.148)	(0.144)	(0.142)	(0.156)	(0.138)	(0.139)	(0.132)	(0.139)	(0.125)	(0.117)	(0.116)
10	-0.2824	0.0487	-0.1052	0.0003	-0.2836	0.0498	-0.0755	0.0017	-0.2919	0.0438	-0.0353	-0.0029
	(0.104)	(0.101)	(0.101)	(0.099)	(0.104)	(0.098)	(0.098)	(0.097)	(0.100)	(0.089)	(0.088)	(0.088)
20	-0.1498	0.0272	-0.0322	0.0017	-0.1508	0.0241	-0.0258	0.0000	-0.1520	0.0230	-0.0096	0.0000
	(0.080)	(0.075)	(0.078)	(0.077)	(0.080)	(0.076)	(0.077)	(0.077)	(0.084)	(0.074)	(0.076)	(0.075)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is 0.0585. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages. Conditional logistic AR(1) is generated according to equation (3.35)

Table G.35: **Bias and Std Deviation (×10) of APE Estimates under a Marginal Logistic AR(1) Process**

	$\rho = 0$				$\rho = 0.4$				$\rho = 0.8$			
	PHP		MEP		PHP		MEP		PHP		MEP	
T	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
5	-0.4652	0.0975	-0.2982	-0.0028	-0.4634	0.0947	-0.2191	-0.0053	-0.4708	0.0896	-0.1200	-0.0057
	(0.151)	(0.140)	(0.144)	(0.136)	(0.148)	(0.135)	(0.132)	(0.130)	(0.150)	(0.126)	(0.118)	(0.117)
10	-0.2827	0.0498	-0.1051	-0.0002	-0.2878	0.0404	-0.0839	-0.0052	-0.2913	0.0365	-0.0442	-0.0109
	(0.098)	(0.099)	(0.099)	(0.098)	(0.105)	(0.100)	(0.098)	(0.097)	(0.105)	(0.094)	(0.090)	(0.090)
20	-0.1498	0.0264	-0.0314	0.0023	-0.1551	0.0196	-0.0326	-0.0055	-0.1629	0.0096	-0.0213	-0.0109
	(0.081)	(0.077)	(0.077)	(0.077)	(0.082)	(0.076)	(0.078)	(0.077)	(0.078)	(0.073)	(0.072)	(0.071)

R=1,000, N=300. Standard deviations are given in parenthesis. Both bias and standard deviations are multiplied by 10. True APE value is 0.0582. Specification (1) incorrectly assumes that the random effects a_i and b_i are uncorrelated with the x's while specification (2) assumes that a_i and b_i are correlated with the x's through their time averages. Marginal logistic AR(1) is generated according to equation (3.36)

BIBLIOGRAPHY

BIBLIOGRAPHY

Ackerberg, D., X. Chen, and J. Hahn (2012): 'A practical asymptotic variance estimator for two-step semiparametric estimators,' *Review of Economics and Statistics*, 94(2), 481–498.

Ai, C., and X. Chen (2003): 'Efficient estimation of models with conditional moment restrictions containing unknown functions,' *Econometrica*, 71(6), 1795–1843.

Akin, J. S., D. K. Guilkey, and R. Sickles (1979): 'A random coefficient probit model with an application to a study of migration,' *Journal of Econometrics*, 11(2), 233 – 246.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): 'How Much Should We Trust Differences-In-Differences Estimates?*,' *The Quarterly Journal of Economics*, 119(1), 249–275.

Blattman, C., J. C. Jamison, and M. Sheridan (2017): 'Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia,' *American Economic Review*, 107(4), 1165–1206.

Blundell, R., and R. L. Matzkin (2014): 'Control functions in nonseparable simultaneous equations models,' *Quantitative Economics*, 5(2), 271–295.

Blundell, R., and J. L. Powell (2003): 'Endogeneity in nonparametric and semiparametric regression models,' *Econometric society monographs*, 36, 312–357.

Blundell, R. W., and J. L. Powell (2004): 'Endogeneity in Semiparametric Binary Response Models,' *The Review of Economic Studies*, 71(3), 655–679.

Chen, X. (2007): 'Large sample sieve estimation of semi-nonparametric models,' *Handbook of econometrics*, 6, 5549–5632.

Chen, X., V. Chernozhukov, S. Lee, and W. K. Newey (2014): 'Local identification of nonparametric and semiparametric models,' *Econometrica*, 82(2), 785–809.

Chen, X., O. Linton, and I. Van Keilegom (2003): 'Estimation of semiparametric models when the criterion function is not smooth,' *Econometrica*, 71(5), 1591–1608.

D'Haultfœuille, X., and P. Février (2015): 'Identification of nonseparable triangular models with discrete instruments,' *Econometrica*, 83(3), 1199–1210.

Dong, Y., and A. Lewbel (2015): 'A simple estimator for binary choice models with endogenous regressors,' *Econometric Reviews*, 34(1-2), 82–105.

Duckworth, A. L., and P. D. Quinn (2009): 'Development and validation of the Short Grit Scale (GRIT–S),' *Journal of personality assessment*, 91(2), 166–174.

Escanciano, J. C., D. Jacho-Chávez, and A. Lewbel (2016): 'Identification and estimation of semiparametric two-step models,' *Quantitative Economics*, 7(2), 561–589.

Fernández-Val, I. (2009): 'Fixed effects estimation of structural parameters and marginal effects in panel probit models,' *Journal of Econometrics*, 150(1), 71–85.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects,' *Econometrica*, 76(5), 1191–1206.

Gandhi, A., K. I. Kim, and A. Petrin (2013): 'Identification and Estimation in Discrete Choice Demand Models when Endogenous Variables Interact with the Error,' .

Greene, W. (2004): 'The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects,' *The Econometrics Journal*, 7(1), 98–119.

Greene, W. (2011): *Econometric Analysis*. Pearson Education.

Hahn, J., Z. Liao, and G. Ridder (2018): 'Nonparametric two-step sieve M estimation and inference,' *Econometric Theory*, pp. 1–44.

Hahn, J., and G. Ridder (2011): 'Conditional moment restrictions and triangular simultaneous equations,' *Review of Economics and Statistics*, 93(2), 683–689.

Hall, P., J. L. Horowitz, et al. (2005): 'Nonparametric methods for inference in the presence of instrumental variables,' *The Annals of Statistics*, 33(6), 2904–2929.

Harvey, A. C. (1976): 'Estimating regression models with multiplicative heteroscedasticity,' *Econometrica*, Vol. 44(3), 461–465.

Hausman, J. A., and D. A. Wise (1978): 'A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences,' *Econometrica*, 46(2), 403–426.

Hoderlein, S., H. Holzmann, M. Kasy, and A. Meister (2016): 'Erratum Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment,' *The Review of Economic Studies, forthcoming*.

Hoderlein, S., H. Holzmann, and A. Meister (2017): 'The triangular model with random coefficients,' *Journal of Econometrics*, 201(1), 144–169.

Hong, H., and E. Tamer (2003): 'Endogenous binary choice model with median restrictions,' *Economics Letters*, 80(2), 219–225.

Horowitz, J. L. (1992): 'A smoothed maximum score estimator for the binary response model,' *Econometrica: journal of the Econometric Society*, pp. 505–531.

Ichimura, H., and L.-F. Lee (1991): 'Semiparametric least squares estimation of multiple index models: single equation estimation,' in *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge*, pp. 3–49. Cambridge University Press.

Imbens, G. W., and W. K. Newey (2009): 'Identification and estimation of triangular simultaneous equations models without additivity,' *Econometrica*, 77(5), 1481–1512.

Kasy, M. (2011): 'Identification in triangular systems using control functions,' *Econometric Theory*, 27(3), 663–671.

——— (2014): 'Instrumental variables with unrestricted heterogeneity and continuous treatment,' *The Review of Economic Studies*, 81(4), 1614–1636.

Keane, M. P. (1994): 'A Computationally Practical Simulation Estimator for Panel Data,' *Econometrica*, 62(1), 95–116.

Khan, S. (2013): 'Distribution free estimation of heteroskedastic binary response models using Probit/Logit criterion functions,' *Journal of Econometrics*, Vol. 172(1), 168 – 182.

Kim, K. i., and A. Petrin (2017): 'A New Control Function Approach for Non-Parametric Regressions with Endogenous Variables,' .

Klein, R., and F. Vella (2009): 'A semiparametric model for binary response and continuous outcomes under index heteroscedasticity,' *Journal of Applied Econometrics*, Vol. 24(5), 735–762.

Krief, J. M. (2014): 'An integrated kernel-weighted smoothed maximum score estimator for the partially linear binary response model,' *Econometric Theory*, 30(3), 647–675.

Lewbel, A. (2000): 'Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables,' *Journal of Econometrics*, 97(1), 145–177.

——— (forthcoming): 'The identification zoo–meanings of identification in econometrics,' *Journal of Economic Literature*.

Lewbel, A., Y. Dong, and T. T. Yang (2012): 'Comparing features of convenient estimators for binary choice models with endogenous regressors,' *Canadian Journal of Economics/Revue canadienne d'économique*, 45(3), 809–829.

Lin, W., and J. M. Wooldridge (2015): 'On different approaches to obtaining partial effects in binary response models with endogenous regressors,' *Economics Letters*, 134, 58 – 61.

Manski, C. F. (1985): 'Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator,' *Journal of Econometrics*, 27(3), 313–333.

Manski, C. F. (1988): 'Identification of Binary Response Models,' *Journal of the American Statistical Association*, 83(403), 729–738.

McCulloch, C. E., and J. M. Neuhaus (2001): *Generalized linear mixed models*. Wiley Online Library.

Newey, W. K. (1994): 'The asymptotic variance of semiparametric estimators,' *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

——— (2013): 'Nonparametric instrumental variables estimation,' *American Economic Review*, 103(3), 550–56.

Newey, W. K., and D. McFadden (1994): 'Chapter 36: Large sample estimation and hypothesis testing,' *Handbook of Econometrics*, Vol. 4, 2111 – 2245.

Newey, W. K., and J. L. Powell (2003): 'Instrumental variable estimation of nonparametric models,' *Econometrica*, 71(5), 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999): 'Nonparametric estimation of triangular simultaneous equations models,' *Econometrica*, 67(3), 565–603.

Petrin, A., and K. Train (2010): 'A Control Function Approach to Endogeneity in Consumer Choice Models,' *Journal of Marketing Research*, 47(1), 3–13.

Pinkse, J. (2000): 'Nonparametric Two-Step Regression Estimation When Regressors and Error Are Dependent,' *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28(2), 289–300.

Rivers, D., and Q. H. Vuong (1988): 'Limited information estimators and exogeneity tests for simultaneous probit models,' *Journal of Econometrics*, 39(3), 347 – 366.

Rothe, C. (2009): 'Semiparametric estimation of binary response models with endogenous regressors,' *Journal of Econometrics*, 153(1), 51–64.

Rothenberg, T. J. (1971): 'Identification in parametric models,' *Econometrica: Journal of the Econometric Society*, pp. 577–591.

Sim, C. H. (1993): 'First-Order Autoregressive Logistic Processes,' *Journal of Applied Probability*, 30(2), 467–470.

Smith, R. J., and R. W. Blundell (1986): 'An exogeneity test for a simultaneous equation Tobit model with an application to labor supply,' *Econometrica: Journal of the Econometric Society*, pp. 679–685.

Song, W. (2016): 'A Semiparametric Estimator for Binary Response Models with Endogenous Regressors,' .

Stock, J. H., J. H. Wright, and M. Yogo (2002): 'A survey of weak instruments and weak identification in generalized method of moments,' *Journal of Business & Economic Statistics*, 20(4), 518–529.

Su, L., and A. Ullah (2008): 'Local polynomial estimation of nonparametric simultaneous equations models,' *Journal of Econometrics*, 144(1), 193–218.

Swamy, P., and G. S. Tavlas (1995): 'Random Coefficient Models: Theory and Applications.,' *Journal of Economic Surveys*, 9(2), 165.

Swamy, P. A. V. B. (1970): 'Efficient Inference in a Random Coefficient Regression Model,' *Econometrica*, 38(2), 311–323.

Torgovitsky, A. (2015): 'Identification of nonseparable models using instruments with small support,' *Econometrica*, 83(3), 1185–1197.

Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data*, Econometric Analysis of Cross Section and Panel Data. MIT Press.

Wooldridge, J. M. (2005): 'Unobserved heterogeneity and estimation of average partial effects,' *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 27–55.

——— (2018): 'Correlated Random Effects Models with Unbalanced Panels,' *Journal of Econometrics*.

Wooldridge, J. M., and Y. Zhu (Forthcoming): 'Inference in Approximately Sparse Correlated Random Effects Probit Models,' *Journal of Business and Economic Statistics*.