AVOIDING BAD CONTROL IN REGRESSION FOR PARTIALLY QUALITATIVE OUTCOMES, AND CORRECTING FOR ENDOGENEITY BIAS IN TWO-PART MODELS: CAUSAL INFERENCE FROM THE POTENTIAL OUTCOMES PERSPECTIVE

Daniel Abebe Asfaw

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Economics,
Indiana University

May 2021

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

Joseph Terza, Ph.D, Chair

Mark Ottoni-Wilhelm, Ph.D

January 28, 2021

Vidhura Tennekoon, Ph.D

Fei Tan, Ph.D

Dedication

To my late mother Fantaye Birhanu Tone,

To my father Abebe Asfaw Anbeso,

To my wife Kidist Ibrie Yasin and

To my daughter Maranatah Daniel Abebe

Acknowledgement

I am immensely grateful to Dr Joseph Terza, my advisor and the chair of my dissertation committee. I thank him very much for his excellent guidance, insights and close supervision from the inception to the completion stages of this dissertation. His persistent encouragement has kept my motivation high during the entire research process and his extreme patience has given me the opportunity to master essential research skills. Dr Terza has practically shown me the qualities of a leading scholar. Truly, I feel privileged for getting the opportunity to be trained and supervised by him.

I also want to extend my sincere appreciation to my dissertation committee members: Dr Mark Ottoni-Wilhelm, Dr Vidhura Tennekoon, and Dr Fei Tan for their insightful and constructive comments which has improved the dissertation. I am grateful to Dr Steven Russell, the chair of our Department for his tremendous help throughout my stay in the PhD program. My appreciation also goes to Dr Ann Royalty and Dr Wendy Morrison, the former and the current directors of the PhD program, respectively, for their outstanding leadership. I have also benefited from excellent instruction from the faculty of our department named above, and others including Dr Sumedha Gupta, Dr Henry Mak, Dr Jaesoo Kim, Dr Subir Chakrabarti and Dr Peter Rangazas. I also want to thank my mentor Emeritus Professor David Bivin who generously provided me with personal and professional support. I am also indebted to Dana Ward and Terri Crews who have always been there to help me with any needed administrative issues.

My special thanks go to my lovely wife Kidist Ibrie Yasin and our precious daughter Maranatah. They have always been a source of inspiration for me to achieve great things in life. I thank my wife for her love, persistent encouragement, prayer and for all the

sacrifices she has made to realize my success. I also thank my daughter for filling my life with joy and for her sweet prayers. Last but not least, I am extremely grateful to my family who reside in different parts of the world for their love, prayer, and support.

Daniel Abebe Asfaw

AVOIDING BAD CONTROL IN REGRESSION FOR PARTIALLY QUALITATIVE OUTCOMES, AND CORRECTING FOR ENDOGENEITY BIAS IN TWO-PART MODELS: CAUSAL INFERENCE FROM THE POTENTIAL OUTCOMES PERSPECTIVE

The general potential outcomes framework (GPOF) is an essential structure that facilitates clear and coherent specification, identification, and estimation of causal effects. This dissertation utilizes and extends the GPOF, to specify, identify, and estimate causally interpretable (CI) effect parameter (EP) for an outcome of interest that manifests as either a value in a specified subset of the real line or a qualitative event -- a *partially qualitative outcome* (PQO). The limitations of the conventional GPOF for casting a regression model for a PQO is discussed. The GPOF is only capable of delivering an EP that is subject to a bias due to bad control. The dissertation proposes an outcome measure that maintains all of the essential features of a PQO that is entirely real-valued and is not subject to the bad control critique; the *P-weighted outcome* – the outcome weighted by the probability that it manifests as a quantitative (real) value. I detail a regression-based estimation method for such EP and, using simulated data, demonstrate its implementation and validate its consistency for the targeted EP. The practicality of the proposed approach is demonstrated by estimating the causal effect of a fully effective policy that bans pregnant women from smoking during pregnancy on a new measure of birth weight. The dissertation also proposes a Generalized Control Function (GCF) approach for modeling and estimating a CI parameter in the context of a fully parametric two-part model (2PM) for a continuous outcome in which the causal variable of interest is continuous and endogenous. The proposed approach is cast within the GPOF. Given a fully parametric specification for the causal variable and under regular Instrumental Variables (IV) assumptions, the approach is shown to satisfy the conditional independence assumption that is often difficult to hold under alternative approaches. Using simulated data, a full information maximum likelihood (FIML) estimator is derived for estimating the “deep” parameters of the model. The Average Incremental Effect (AIE) estimator based on these deep parameter estimates is shown to outperform other conventional estimators. I apply the method for estimating the medical care cost of obesity in youth in the US.

Joseph V. Terza, Ph.D., Chair

Table of Contents

List of Tables

List of Abbreviations

AIE	Average Incremental Effect
APB	Absolute Percentage Bias
BMI	Body Mass Index
CDF	Cumulative Density Function
CI	Causally Interpretable
CIND	Conditional Independence
CPOM	Conditional Potential Outcomes Model
DGP	Data Generating Process
EP	Effect Parameter
EST AIE	Estimated AIE
FIML	Full Information Maximum Likelihood
FIML-EXOG	Full Information Maximum Likelihood Estimator with Exogenous Causal Variable
FP2PM	Fully Parametric Two-Part Model
GCF	Generalized Control Function
GCF-FIML	Generalized Control Function-Full Information Maximum Likelihood
GG	Generalized Gamma
IV	Instrumental Variables
MLE	Maximum Likelihood Estimator
NSD	No-Structural Difference
NSFG	National Survey of Family Growth
OLS	Ordinary Least Squares

PCO	P-weighted Conditional Outcome
PDF	Probability Density Function
PQO	Partially Qualitative Outcomes
TBR	Terza, Basu and Rathouz
2PM	Two-Part Models
2SLS	Two-Stage Least Squares
2SRI	Two-Stage Residual Inclusion
2SPS	Two-Stage Predictor Substitution
2SGCF	Two-Stage Generalized Control Function

Chapter 1

Introduction, Background and Significance, Summary

Causal Inference is at the heart of nearly all empirical economic research. Essential for conducting valid causal inference are rigorous specification and accurate estimation of parameters that characterize causal relationships of interest. In this dissertation, two regression-based approaches are developed. The first one is designed for specification and estimation of causally interpretable (CI) parameter for Partially Qualitative Outcomes (PQO) – outcomes that manifest either as a value on the real line or as a qualitative event. Birth weight is an example of a PQO because it is observed only when a pregnancy ends in a live birth; otherwise, the outcome would be non-live birth. The second approach is developed for specification, identification, and estimation of CI parameters in Two-Part Model (2PM) context for continuous nonnegative outcomes where the causal variable of interest is continuous and endogenous. To ensure causal interpretability of the targeted parameters and their estimates, both approaches are developed within the General Potential Outcomes framework (GPOF).

In the conventional Conditional Potential Outcomes Model (CPOM), an essential model within the GPOF, outcomes are assumed to manifest either *exclusively* as a value on the real line or *exclusively* as a qualitative event. Casting a regression model for a PQO using the CPOM is difficult because to satisfy the aforementioned assumption, one needs to ignore either the quantitative or the qualitative component of the PQO. While focusing only on the qualitative component will change the causal inference objective, ignoring it would cause a bias due to bad control. The dissertation proposes an outcome measure that maintains all of the essential features of a PQO that is entirely real-valued and is not subject to the bad control critique; the *P-weighted outcome* – the outcome weighted by the probability that it manifests as a quantitative (real) value. A regression-based estimation method for such effect parameters is detailed and using simulated data, I demonstrate its implementation and validate its consistency for the targeted effect parameter. To demonstrate the practicality of the proposed approach, I apply the model and method to estimate the causal effect of a fully effective policy that bans pregnant women from smoking during pregnancy on a new measure of birth weight.

The 2PM is one of the most widely applied empirical modeling and estimation framework in empirical health economics. In this dissertation, I extend the generic fully parametric 2PM (FP2PM) framework developed in Hao and Terza (2018) to accommodate cases in which the causal variable of interest is endogenous. The proposed approach considers continuous outcome and continuous endogenous variable. By casting the parameter of interest within the GPOF, the proposed approach provides a consistent definition of endogeneity. In particular, I propose a generalized control function (GCF) approach to identify, estimate and draw inference for an average incremental effect (AIE) of a one-unit change in the causal variable of interest. Under regular Instrumental Variables (IV) assumptions, the GCF is shown to satisfy the conditional independence assumption that is often difficult to hold in alternative approaches. Within a FP2PM framework, the GCF specification implies a full information maximum likelihood (FIML) model whose parameters are estimated by FIML method. I call this estimator for the "deep" parameters the Generalized Control Function-Full Information Maximum Likelihood (GCF-FIML) estimator. The GCF-FIML is able to identify causal effects that vary across units in the population based on unobservable characteristics. Using these "deep" parameters, the AIE (the main object of interest) is specified and estimated. The GCF estimator within the FP2PM allows to test two important null hypotheses: “no 2PM is needed” and “the causal variable is exogenous”. In a simulation study, the proposed GCF-FIML based estimator for the targeted parameter is shown to outperform conventional estimators that are used in empirical research. Finally, using data from the Medical Expenditure Panel Survey, the model and method are illustrated by estimating the causal effect of a unit increase in Body Mass Index (BMI) and of moving an average obese individual to an average normal weight BMI on health care cost in youth in the US. A comparison of the result based on the proposed GCF approach to the two-stage predictor substitution estimator used in Biener et al (2020) reveals that the latter significantly overestimates the effect of a change in BMI/obesity on medical care cost.

The remaining part of the dissertation is organized as follows. In chapter 2, specification, identification, and estimation of CI parameter within the GPOF is discussed. I start by specifying an AIE based on relevant counterfactuals. Then a regression model is detailed that can be used to estimate the AIE using observable (factual) data from an appropriately specified Data Generating Process (DGP). The conditions under which such substitution is legitimate is also detailed. This discussion is extended to define endogeneity of the causal variable of interest and mention a general point on how to correct a bias caused by endogeneity. In the last section, an important implicit assumption within the GPOF is discussed that limits its applicability to a special kind of outcomes. This limitation is then addressed in chapter 3 in the context of a PQO. By presenting a detailed overview of the GPOF, chapter 2 lays the ground for the approaches developed in chapters 3 and 4.

Chapter 3 begins by providing definition of a PQO and further elaborating the limitation of the conventional GPOF. Then, a new outcome measure is defined by extending the GPOF that is capable of casting a regression model that maintains all the essential features of a PQO and enables estimation of CI parameter. I present the extended GPOF along with the corresponding regression modeling and estimation method. Then, a simulation study is presented that demonstrates the implementation of the proposed regression model and validate its consistency for estimating the specified AIE. This is followed by an empirical application of the proposed approach.

In chapter 4, I first specify an AIE in the context of a FP2PM for a continuous outcome in which the causal variable of interest is continuous and endogenous. Then, the proposed identification approach is detailed and a FIML model is presented followed by a discussion on estimation of the "deep" parameters of the model. A section is devoted to present statistical tests for two important null hypotheses. A simulation study follows in which the implementation of the proposed GCF-FIML estimator is demonstrated and its consistency for estimating an AIE is evaluated. Therein, the performance of the GCF-FIML estimator is compared with alternative approaches. Then an empirical application is presented where I demonstrate the implementation of the proposed approach and compare the estimated AIE to those obtained by using alternative estimators. Chapter 5 summarizes and concludes the dissertation.

Chapter 2

Causal Inference Within and Outside the Conventional General Potential Outcomes Framework

Assessing causal relationships of interest based on relevant counterfactuals is at the heart of nearly all empirical economic research. Essential to such assessments are rigorous specification and accurate estimation of parameters that describe the relationship between *a presumed causal variable of interest*, $\mathbf{X}$, whose value is to be set and altered in the context of relevant counterfactual, and a designated *outcome of interest*, $\mathbf{Y}$.[1] Relationships of this type are typically characterized by an effect parameter (EP) and estimation of the EP is the objective of the empirical analysis. The general potential outcomes framework (GPOF) provides a means to coherently define the EP in such a way that it is causally interpretable (CI).[2]

This chapter presents the GPOF in the context of estimating an average incremental effect (AIE) as an example of an EP that is considered in many empirical contexts. Within the GPOF, I detail specification of the AIE based on a conditional mean function implied by a conditional probability density function (pdf) for the $\mathbf{Y}$ given an exogenously set values of the $\mathbf{X}$. Then, the conditional potential outcomes model (CPOM) – a model that facilitates a regression-based approach for estimation of the EP within the GPOF – is discussed. Therein, I also outline the conditions under which the stated EP can be identified

---

[1] Henceforth, $\mathbf{X}$ and $\mathbf{Y}$ are to be taken as global replacements for the phrases "presumed causal variable of interest" and "outcome of interest," respectively.

[2] The GPOF is an extension of the potential outcomes framework of Rubin (1974) to non-binary $\mathbf{X}$ and nonlinear $\mathbf{Y}$.

and is estimable by using an observable version of the data. This will be followed by a discussion on a consistent definition of endogeneity within the GPOF that will be employed in chapter 4 where a regression-based modeling and an estimation approach is proposed to correcting endogeneity in 2PM. The chapter closes by discussing an implicit assumption in the GPOF about the way the $\mathbf{Y}$ manifests that, if not satisfied, requires an extension of the framework to casting a regression-based approach for estimating CI parameter. This point is further elaborated in chapter 3 where I discuss regression for PQO.

2.1 The GPOF: Introduction, Basic Concepts and Definitions

Many existing empirical studies in health economics and health service research commence their discussion of a causal inference problem from the data generating process (DGP) from which sample values are drawn. By focusing only on the DGP, the conventional approach fails to explicitly incorporate relevant counterfactuals. This in turn renders the approach to be deficient in recognizing the conditions under which EPs are identified and estimation results are CI (Terza, 2019a).

Casting a causal inference problem exclusively based on the DGP is even more problematic when the specific empirical context involves endogeneity. This is because the conventional DGP-based approach defines endogeneity in ambiguous and self-obviating way (Terza, 2019b). The GPOF, on the other hand, provides a framework that facilitates a clear and rigorous definition of the EP based on relevant counterfactuals. It also enables the analyst to define endogeneity in sustainable and unambiguous way that delivers the analyst a path on which to expand the DGP by adding appropriate structure to achieve identification of CI parameters.

In chapter 3, a special case is discussed where the conventional GPOF is limited for casting a regression model for an empirical context. In particular, I argue that for partially qualitative outcomes (PQO) – outcomes that manifest either as a value in a specified set of the real line or a qualitative event – it is not possible to define an EP that is real-valued and can be estimated without bias. The GPOF, however, can be extended by replying on its fundamental principle, viz. characterizing outcomes based on relevant counterfactuals. To set the stage for the proposed approach for PQO in chapter 3 and for correcting endogeneity in the context of 2PM in chapter 4, below I review the GPOF as detailed in Terza (2019a, 2019b).

Here the fundamental definitions and concepts that characterize the GPOF as detailed in Terza (2019a) is presented. I begin with definitions of the counterfactual and observable versions of the $\mathbf{X}$ and the $\mathbf{Y}$. In the GPOF, two versions of the $\mathbf{X}$ are distinguished as:

> $X^* \equiv$ the random variable representing the hypothetical (counterfactual) exogenously mandated version of the distribution of the $\mathbf{X}$ that might result from a policy intervention (X^* is, by design, independent of all other variates germane to specification, identification, and estimation of the EP of interest).

and

> $X \equiv$ the random variable representing the observable (factual) version of the distribution of the $\mathbf{X}$ (sampled values of the $\mathbf{X}$ are drawn from the distribution of X).

Likewise, two versions of the **Y** are distinguished as:

Y_{X^*} ≡ the random variable representing the distribution of the potential outcome, defined as the counterfactual distribution of values of the **Y** that would have manifested for a particular X^*.

and

Y ≡ the random variable representing the factual version of the distribution of the **Y** (the sampled values of the outcome are drawn from the distribution of Y).

Note that although X^* is a random variable, its character is different from X, Y and Y_{X^*}. X^* is a random variable in the sense that its value differs for each elementary unit in the population. Unlike X, Y and Y_{X^*}, the values of X^* is determinate and knowable in the context of relevant counterfactual.

Throughout the remainder of the discussion, I will explicitly and implicitly reference a hypothetical counterfactual (e.g., a prospective policy intervention) in which the X is exogenously changed from X^{pre} to X^{post} (from pre-intervention to post-intervention). Without loss of generality, I write $X^{post} = X^{pre} + \Delta$, where Δ represents the relevant distribution of counterfactually imposed increments to X^{pre} (e.g., as in a policy intervention). Note that, strictly speaking, X^{pre}, X^{post} and Δ are random variables because their possible values vary across the relevant population of individuals with differing probability densities, but these random variables differ in character from X and Y (which, of course, are also random variables). Unlike X and Y which are components of the DGP; X^{pre}, X^{post} and Δ are deterministic in the sense that, for any individual in the relevant population, their values are imposed by the policy maker and/or researcher as part of the relevant counterfactual. Note that, for this reason, X^{pre} and $X^{pre} + \Delta$, are independent of all other variates germane to the specification, identification, and estimation of the relevant EP. So is Δ. To X^{pre} and $X^{pre} + \Delta$ there correspond potential outcomes $Y_{X^{pre}}$ and $Y_{X^{pre} + \Delta}$, respectively. The relevant EP is based on the counterfactually defined entities X^{pre}, Δ, $Y_{X^{pre}}$ and $Y_{X^{pre} + \Delta}$.

2.2 Specifying the Effect Parameter of Interest in the GPOF

To facilitate the discussion on the proposed approaches in this dissertation, two empirical settings are considered. In chapter 3, estimation of the birth weight effect of a hypothetical intervention that effectively bans pregnant women from smoking during their pregnancy is considered. In chapter 4, the illustrative empirical example focuses on estimating the medical care cost of a hypothetical event that increases BMI of each youth in the US by 1 unit. Throughout this chapter except in the last section, the latter is used to illustrate specification, identification, and estimation of an EP in the GPOF. The components of the relevant counterfactuals are

$X^{pre} \equiv$ the random variable representing the pre-counterfactual distribution of BMI among the youth in the US.

$\Delta \equiv$ a one-unit increase to the pre-counterfactual level of BMI to each individual in the relevant population.

Formally, I seek to estimate the following *average incremental effect* (AIE)

$$\mathrm{AIE}(\Delta) = E[Y_{X^{pre} + \Delta}] - E[Y_{X^{pre}}] \tag{1}$$

where $\Delta = -1$, $Y_{X^{pre}}$ is the potential outcome (PO) corresponding to X^{pre}, [i.e., the medical care cost that corresponds to the pre-counterfactual distribution of BMI]. $Y_{X^{pre} + \Delta}$ is the PO corresponding to $X^{pre} + \Delta$ [i.e., the medical care cost that corresponds to the post-counterfactual distribution of BMI where each youth has one-unit higher BMI]. Thus, the AIE in (1) is the average incremental medical care cost effect of a one-unit increase in BMI across the entire youth population in the US.

Even though Y has an observable version (viz., Y), the EP [e.g., (1)] cannot be directly estimated because $Y_{X^{pre}}$ and $Y_{X^{pre} + \Delta}$ are counterfactual entities (the pre- and post-counterfactual POs) and, therefore, are at least partially unobservable (cannot be sampled). In fact, one may only have data on either $Y_{X^{pre}}$ or $Y_{X^{pre} + \Delta}$ but not both. In other words, it is not possible to observe the distribution of the medical care cost for the entire youth under two different BMI distributions at the same time. Therefore, in general, attempts to accurately (consistently) estimate the EP with observable data (X and Y) will be futile because the EP (which is inherently counterfactual) in no substantive way coincides with the observable data from X and Y (which is inherently factual).

2.3 Conditional Potential Outcome Model, and Identification and Estimation of an EP

Counterfactuals are at the heart of causal inference. In the previous sections, I discussed specification of the EP within the GPOF based on relevant counterfactuals. The problem is, however, that these counterfactuals are only partially observable i.e., although one may have data for the $\mathbf{X}$ and the $\mathbf{Y}$ for the entire population, it is virtually impossible to have data on all the relevant counterfactual outcomes for anyone in the population. Therefore, without a rigorous procedure that formalizes the conditions under which the counterfactuals are in congruity with the observed version of the relevant random variables, one cannot simply use the latter to estimate causally interpretable EP and make inference about it.

Terza (2019a) discusses regression-type modeling of the potential outcome that can be used to bridge the gap between the counterfactual object of interest (the EP) and the factual data (sampled from X and Y) to be used for estimation. Terza (2019a) refers to such modeling of the potential outcome as the conditional potential outcome model (CPOM). The CPOM provides a basis on which to build conditions under which one can substitute observable version of the **X** and the **Y** for their counterfactuals (more on this later). This in turn ensures identification of the specified EP.

The CPOM can be defined at any level of parametricity. Throughout the discussion, I define a fully parametric (FP) version of the CPOM. Conditioning on a vector of control variables V, a FP version of the CPOM can be specified for a continuous Y_{X^*} as

$$\mathrm{pdf}(Y_{X^*}|V) \equiv f(Y_{X^*}, X^*, V; \pi) \qquad (2)$$

where pdf(A|B) is the conditional pdf of A given B, f(.) is a known function whose value is determined by the scalars and vectors in the bracket. π is a vector of unknown regression parameters (henceforth I refer to π as the vector of *deep* parameters). Using the CPOM, it is possible to rewrite the EP [e.g., (1)] by exploiting the regression-like conditional mean function that (2) implies. It follows from (2) that

$$E[Y_{X^*}|V] = m(X^*, V; \pi) \qquad (3)$$

Note that, because (2) is known, the conditional mean function (3) also has known form. Using the law of iterated expectations and (3), (1) can be rewritten as

$$\mathrm{AIE}(\Delta) = E[m(X^{pre} + \Delta, V; \pi)] - E[m(X^{pre}, V; \pi)] \qquad (4)$$

It is clear from (4) that if we had a consistent estimator for the vector of deep parameters π (say, $\hat{\pi}$) and V were fully observable, then we would be able to consistently estimate (4) using its following sample analog[3]

$$\widehat{\mathrm{AIE}(\Delta)} = \sum_{i=1}^{n} \frac{1}{n}\{m(X_i^{pre} + \Delta_i, V_i; \hat{\pi}) - m(X_i^{pre}, V_i; \hat{\pi})\} \qquad (5)$$

where X_i^{pre} and Δ_i are the counterfactually imposed values of X^{pre} and Δ, respectively, for the ith member of the sample of size n (i = 1, ..., n) and V_i is the value of the vector of controls sampled for the ith observation. Before embarking on the consistent estimation of the deep parameters of the model, identification must be established at two levels (Terza, 2019a). Below I discuss these two levels of identification.

2.3.1 Non-parametric Conditions for Identification of the CPOM

First, aside from any particular parametric specification one must show that the CPOM is non-parametrically identified. In the context of the CPOM in (2), according to Terza (2019a), non-parametric identification of the EP [e.g. (4)] is established if

$$\mathrm{pdf}(Y|V, X) = f_{(Y_{X^*}|V)}(Y, V, X; \pi) \qquad (6)$$

In other words, the EP is identified if the conditional pdf of Y given V and X can be obtained by substituting X and Y for X^* and Y_{X^*} in (2), respectively. Terza (2019a) details conditions under which such substitution is legitimate. These conditions are: i) the

[3] The asymptotic standard error of (5) can be obtained using the approach in Terza (2016a, 2016b, 2016c, 2017).

conditional independence assumption (CIND) which requires that, conditional on V, Y_{X^*} be independent of X; ii) Conditional Outcome Invariance which holds if

$$\left(Y_{X^*=a} \mid V, X^*=a\right) = (Y \mid V, X=a)$$

where "a" is a value in the support of the conditional distribution. ii) implies that conditional on V it should not matter to the value of the outcome whether X = a is chosen by an agent or exogenously imposed by a policy maker. iii) Overlap which holds if

$$0 < P_{(X \mid V)}(x|v) < 1$$

where, $P_{(X \mid V)}(x|v)$ denotes the conditional pdf of X given V = v evaluated at X = x. Overlap requires that at each value of V, X has a nontrivial but uncertain probability that it equals x. In the case where X is binary overlap implies that at each value of V there are units who are and are not exposed to the relevant policy.

Among the above three conditions, the CIND is the most important, and the least likely to be true. Intuitively, CIND implies that all other variables that confound the pure causal relationship between the **X** and the **Y** are included in the vector of controls V. In other words, the CIND guarantees that for a given V, the distribution of the observed outcome, Y, for those units in the population with $X = X^{pre} + \Delta$ would have been the same as for units with $X = X^{pre}$ had Δ, have been applied to the latter group. Thus, given V, Δ was as good as randomly assigned.

2.3.2 Parametric Conditions for Identification of the CPOM

With non-parametric identification maintained, the second level of identification to be established is parametric identification.[4] This type of identification has been extensively covered in the literature and in most graduate level econometrics texts. The discussion of parametric identification is exclusively focused on the level of the DGP. The parameters of the DGP model are identified if the chosen functional forms for the relevant aspects of the DGP (e.g. conditional mean, higher-order conditional moments, conditional pmf/pdf, etc.) are such that full knowledge of the values of those aspects of the DGP would imply knowledge of the values of the relevant parameters.

Given that the CPOM is non-parametrically and parametrically identified, it follows that π can be consistently estimated as the maximum likelihood estimator (MLE) obtained as

$$\hat{\pi} = \text{argmax} \ \sum_{i=1}^{n} q(\breve{\pi}, Z_i) \qquad (7)$$

where $q(\breve{\pi}, Z_i) = \ln[f_{(Y_{X^*}|V)}(Y_i, X_i, V_i; \breve{\pi})]$ is the log likelihood of the ith unit in the sample and $Z_i=[Y_i \quad X_i \quad V_i]$ is the data vector for the ith sample.

2.4 Endogeneity in the GPOF

Endogeneity is one of the most common problem in empirical economic research that leads to inconsistent estimation of parameters of interest. Correcting for endogeneity bias requires a framework that facilitates a correct definition of endogeneity. The

[4] Note that if one cannot establish non-parametric identification as detailed in Terza (2019a) then subsequent discussion of parametric identification have no useful content from the perspective of causal inference.

conventional DGP-based approach is problematic in this respect as it provides ambiguous and self-obviating definition of endogeneity (Terza, 2019b). For instance, in a minimally parametric setting where only the first moment of the relevant random variables are specified, this approach defines endogeneity as the absence of correlation between the causal variable of interest and an additive error term given an arbitrarily set vector of controls. As discussed in Terza (2019b), this reduces the problem of endogeneity to misspecification of the conditional mean function. Moreover, the lack of specificity about the vector of controls in the conventional DGP-based approach renders the above definition of endogeneity ambiguous because for any vector of controls, there correspond a "true" parametric or nonparametric conditional mean function.

By taking into account of the counterfactual nature of the EP, the GPOF, on the other hand, provides a more consistent definition of endogeneity. In particular, the CPOM within the GPOF is unique because it is a known function in which a unique vector of essential controls, V, induces CIND between Y_{X^*} and X.[5] Endogeneity of the $\mathbf{X}$ in the GPOF is, thus, defined as a situation where V is only partially observable. In such cases, one can write $V = [X_o \quad X_u]$ where X_o is a partition of V comprising a vector of observable control variables and X_u is a scalar representing essential unobservable element of V.

In the context of estimating the AIE of a one-unit increase in BMI on medical care cost, BMI is likely endogenous because the vector V that induces CIND between observed BMI and potential medical care cost includes unobservable. For example, those youth with

[5] Terza (2019b) defines essential control as "a vector of variates comprising all, and only, confounders for Y_{X^*} and X."

lower BMI may have different level of health consciousness that is unobservable and also related with medical care cost.

In the presence of Endogeneity, additional structure should be built that would resolve the identification problem caused by partial unobservability of the unique vector V. This structure expands the DGP to include additional (instrumental) variables and (possibly) concomitant assumptions about the specifications of conditional moment of the expanded DGP (and those variables). In chapter 4, I present an approach that corrects endogeneity in 2PM context for continuous nonnegative outcomes for the specific case where the endogenous variable is continuous.

2.5 An Implicit Assumption in the Conventional GPOF

Implicit in the GPOF is that the $\mathbf{Y}$ is assumed to manifest either *exclusively* as a value in a specified subset on the real line or *exclusively* as a qualitative event. For example, in the empirical example where one studies the medical care cost of obesity, the $\mathbf{Y}$ can take only real values that are nonnegative. These kinds of empirical context can be analyzed within the conventional GPOF. But what if the $\mathbf{Y}$ in a given empirical context consists of a union of events that correspond to values on the real line and qualitative events? Suppose the interest is in estimating the AIE of a policy intervention that effectively bans all pregnant women from smoking during pregnancy on birth weight. In this case, birth weight is observed only for those pregnancies that end in a live birth. All other pregnancies that do not end in live birth have an observed outcome that is just non-live birth, not latent birth weight. As mentioned earlier, such outcomes – that manifest either as a value on the real line or as a qualitative event – are called Partially Qualitive Outcomes (PQO). In the next chapter, I discuss how the extant GPOF is limited in handling causal inference for PQO and extend it to accommodate such cases.

Chapter 3

Avoiding Bad Control Bias in Partially Qualitative Regression: Causal Inference by Extending the Conventional GPOF

Drawing causal inference within the conventional GPOF is predicated on the assumption that the outcome of interest manifests either *exclusively* as a value on the real line or *exclusively* as a qualitative event. This chapter presents an approach for specification, identification, and estimation of an EP for cases in which the values that $\mathbf{Y}$ takes is a union of two non-empty sets: a set containing values on the real line and a set of qualitative event(s). I call such outcomes as Partially Qualitative Outcomes (PQO). The chapter begins with a description of a PQO using the running example of estimating the birth weight effect of a fully effective policy that bans pregnant women from smoking during pregnancy. To shed light on its distinctive feature, the PQO is compared to outcomes that are typically modeled in the context of widely known corner solution models, namely the two-part model (2PM) and the sample selection model. Therein, I also discuss the limitation of the conventional GPOF for casting a regression model. Then the conventional framework is extended to encompass cases where the $\mathbf{Y}$ is a PQO. Within the expanded framework, a new outcome measure is proposed that allows casting a regression model that would maintain all of the essential features of a PQO and enables identification, and estimations of a causally interpretable EP for a PQO. The proposed outcome measure is referred to as the *P-weighted outcome* – the outcome weighted by the probability that it manifests as a quantitative (real) value. I discuss the practicality and usefulness of this new measure for specifying and identifying an EP that characterize the causal relationships between a policy variable of interest and the PQO. Then, a regression-based estimation method for such EP is detailed and using simulated data, the implementation of the method and the concomitant estimator of the EP is demonstrated, and its consistency is validated. Using data from the National Survey of Family Growth (NSFG), I apply the proposed model and method in estimating the AIE of a counterfactually mandated fully effective policy intervention that brings the smoking levels of all pregnant women down to zero on *natility-weighted birth weight* (a new measure of birth weight discussed in detail later).

3.1 Defining Partially Qualitative Outcomes

The suitability of the GPOF for casting regression model is predicated on the nature of the outcome under consideration. The conventional GPOF is suited to cast EP for cases in which the outcome of interest manifests either *exclusively* as a value on the real line or *exclusively* as a qualitative event that would be assigned a quantitative value for analysis. Examples in the former category include outcomes such as wage, BMI, health care expenditures and so on whereas examples for the latter category are outcomes indicating a person's subjective health status, whether she has health insurance, and so on. In this chapter I consider PQO – defined as outcomes that manifest either as a value on the real line or a qualitative event. For example, a newborn's quantitative health outcome is defined only if a pregnancy ends in a live birth. In other words, the $\mathbf{Y}$ from a pregnancy might manifest as a non-live birth or the value of a specific measure of the newborn health outcome of interest, such as birth weight. For the purpose of exposition and to fix ideas, as running example I consider specification, identification and estimation of the EP representing the causal effect of smoking during pregnancy on birth weight (henceforth I use S+B to refer to this example).

Since the US surgeon general report in 1964 that publicized the adverse relationship between maternal smoking and infant health, many studies have been conducted with the aim of investigating how changes in smoking during pregnancy affect pregnancy and infant health outcomes. A number of studies, for instance, document that smoking during pregnancy would lead to miscarriages and still births (Walsh, 1994; Ness et al., 1999; Mishra, Dobson, & Schofield, 2000; Pineles, Park, & Samet, 2014; Hyland et al., 2016). Others find that maternal smoking during pregnancy significantly reduces birth weight (Rosenzweig & Schultz, 1983; Evans & Ringel, 1999; Lumley et al., 2004; Lien & Evans, 2005; Abrevaya and Dahl, 2008). Despite using standard methodologies designed for estimating causal effects, these studies have limitations. On the one hand, those that focus on the effect of smoking on whether or not a pregnancy ends in a live birth are useful in that they produce arguably unbiased result, but they are less than comprehensive given the many other health related birth outcomes that are of great interest. On the other hand, studies that extend beyond the live birth question ignore the PQO nature of these other birth outcomes by focusing (conditioning) only on pregnancies that end in live births.[6] Such studies, including those in the S+B context, generally produce biased estimates of the causal effect of the $\mathbf{X}$ on the quantitative component of the $\mathbf{Y}$ because they ignore the likelihood that occurrence of the qualitative event is itself affected by the $\mathbf{X}$. This results in bias due to so-called "bad control". Bad control is a conditioning variable that is itself affected by the $\mathbf{X}$ (more on this later)[7].

[6] Literature from epidemiology suggests that only 60-70% of fertilized eggs results in live birth (Liew et al. 2015).

[7] Heckman and Navarro-Lozano (2004) and Wooldridge (2005) present analytical proof showing how the presence of a bad control in the conditioning vector of a regression-based

3.2 The Conventional GPOF and PQO: the dilemma

In general, PQO models demand special attention for two reasons.

1) As mentioned above, conditioning on a bad control causes bias. In order to avoid this problem, one must take account of the qualitative component of the model in the specification of the relevant potential outcome. In the S+B context, accurate estimation of the EP requires one to incorporate into the analysis the pregnancies that end in non-live birth outcomes in addition to pregnancies that end in live birth. To the best of my knowledge, there is no empirical study that analyzes the S+B while explicitly accounting for the effect of smoking during pregnancy on the event that the pregnancy may end in non-live birth.[8]

2) In such PQO contexts, as in all modeling contexts in applied econometrics, the analyst seeks to specify and estimate an EP (representing the causal effect of the $\mathbf{X}$ on the $\mathbf{Y}$) that is real-valued using observable real-valued data on the $\mathbf{X}$ and the $\mathbf{Y}$. In PQO models this is tricky because the only definition of the $\mathbf{Y}$ that is real-valued is one that is conditioned on occurrence of the qualitative event. In the S+B context, this would be birth weight

model leads to inconsistent estimation of an EP. The former demonstrates their analytical proof using simulation. Angrist and Pischke, (2008) also discussed how a bias arises due to the bad control problem.

[8] Bad controls are well recognized in the epidemiology literature in the context of child health. For example, the birth weight paradox, a phenomenon that maternal smoking is inversely associated with infant mortality among low-birth weight babies, is well documented (Wilcox, 1993, 2001; Hernandez-Diaz et al, 2006). There is also a small yet growing literature on the problem of conditioning on live-live birth and its consequence on estimated parameters for different exposures during pregnancy (Suarez et al. 2018; Liew et al. 2015; Lisonkova and Joseph, 2015). The studies, however, rely on simulation analysis with the aim of quantifying the bias due to bad control rather than coming up with an approach that can be used to directly estimate a consistent EP. Moreover, none of them specifically consider the S+B case.

conditional on live birth. As we have seen, however, this leads to the bad control problem.

To overcome this dilemma, an alternative definition of the $\mathbf{Y}$ is suggested that is real-valued but is not subject to this critique; viz., *the P-weighted conditional outcome (PCO)* – the outcome weighted by the probability that it manifests as a quantitative (real) value. In the S+B context, this is the birth weight conditional on live birth weighted by the probability of that event. This measure is called as *natility-weighted birth weight*.[9] Note that the $\mathbf{Y}$ defined in this way does not have a version that is directly observable (sampleable). As I will show, however, this is of no consequence for the practical implementation of the PCO because, despite the fact that it is unobservable (cannot be sampled) it can be used to not only specify the relevant EP but also to estimate it (and conduct inference about it) using the observable (though not entirely real-valued) data. To implement the PCO, a regression-based approach is proposed that involves multiplying the conditional probability that the qualitative event does not occur and the conditional mean of the quantitative component of the PQO for the appropriate sub-population. In the S+B context, this is the product of the probability that the pregnancy ends in a live birth and the conditional mean birth weight for those pregnancies that end in live birth.

3.3 PQO and Corner Solution Models

It should be noted that the empirical contexts to which the proposed PQO modeling approach applies differs distinctly from those for which the two-part and sample selection models are relevant. This section discusses the distinction between the PQO modeling and the two widely applied modeling frameworks.

[9] The word "*Natility*" is formed by combining two words: natal and probability.

3.3.1 Modeling for PQO Vs the Two-Part Model

The support of the outcome in the 2PM proposed by Cragg (1971) is comprised entirely of the nonnegative real values. In the conventional 2PM setup, the part of the model pertaining to manifestation (or not) of zero values for the outcome (the so-called *extensive margin*) differs systematically from the part of the model characterizing the manifestation of the non-zero continuous or count outcome values (the so-called *intensive margin*). As we have seen, the PQO framework likewise comprises two systematically different model components, but its distinguishing feature is that the PQO outcome is based on a sample space that includes a qualitative event (corresponding to the first component of the model) that is not real-valued. For this reason, the outcome for the first component of the PQO model has no quantitative meaning. Whereas, in the 2PM the zero values of the outcome manifested at the extensive margin have cardinal interpretation. Therefore, the PQO and 2PM are not applicable under the same empirical circumstances. For example, in the S+B illustration the 2PM is not applicable because zero birth weight is not a meaningful concept. For instance, it would not be appropriate to assign zero birth weights to pregnancies that did not result in a live birth.

3.3.2 Modeling for PQO Vs the Sample Selection Model

The sample selection model proposed by Heckman (1976, 1979) is designed for empirical contexts in which the objective is estimation of the parameters of an underlying and partially latent regression model of interest. The classic example is estimation of the parameters of the best wage offer regression. The problem is that best wage offers (the outcome variable for the regression) are not fully observable. Presumably, they are only observable as accepted wages for those whose best wage offers exceeded their reservation wages. Such restricted observability of the $\mathbf{Y}$ will result in bias if there are unobservables that are correlated with the best wage offer and the decision to accept a wage offer. The sample selection estimator is designed to correct for this bias.

In the PQO context that I consider here, there exists no such latent regression model of interest. For example, in the S+B context, it is difficult to conceive such an underlying partially latent regression that has policy relevance whose outcome is birth weight. The closest one can come in this context to specifying such a regression would be one in which the $\mathbf{Y}$ is birth weight as it would have manifested if all pregnancies had resulted in live births. Empirical analyses based on such a regression would, however, provide policy makers perhaps with little (no?) relevant inferential information regarding the causal effect of smoking during pregnancy on birth weight. In other words, in the context of the wage offer, analyzing a policy that affects a wage offer may be argued to matter even if a person chooses not to work after exposure to the underlying policy. This is perhaps because of the possible effect of the increase in the latent wage on future labor supply decisions or other contemporaneous outcomes. In the case of the S+B, however, for the subpopulation of pregnancies that end in non-live birth even after the relevant policy intervention, I do not see any reason that one would care about birth weight.[10] Even if one has an interest in this sub-population, because the observable (factual) data on the outcome does not exist, it is not possible to identify the EP non-parametrically. Therefore, in the S+B context, I assume

[10] Note that this is different from the effect of the smoking ban during pregnancy on the birth weight of those pregnancies that resolve in live birth only under the policy. This group whose pregnancy resolution switches from non-live birth to live birth is taken into account in the PQO modeling.

that interest lies in evaluating the effect of a counterfactually mandated smoking ban during pregnancy on the birth weight of the subpopulation of pregnancies that end in live birth.

3.4 The Rationale for Extending the CPOM in the Conventional GPOF

The salient feature of the GPOF discussed in chapter 2 is that it takes explicit account of counterfactuals in specifying the EP. In particular, two versions of the $\mathbf{X}$ and the $\mathbf{Y}$ are distinguished. In the PQO context, the definitions of the $\mathbf{X}$, X^* and X require no special consideration. For instance, in the context of the S+B example

$\mathbf{X} \equiv$ the number of cigarettes a woman smokes per day during pregnancy,

$X^* \equiv$ distribution of counterfactually imposed smoking levels (number of cigarettes per day) for the relevant population of pregnant women

and

$X \equiv$ the random variable representing the observable (factual) version of the distribution of smoking levels.

In the context of the S+B illustrative example that I consider (a counterfactual intervention that fully and effectively prevents (for non-smokers) and eliminates (for smokers) smoking during pregnancy),

$X^{pre} \equiv$ the counterfactually mandated pre-intervention distribution of the number of cigarettes smoked per day during pregnancy, and

$\Delta \equiv -X^{pre}$ the counterfactually imposed increment to pre-intervention smoking levels (representing fully effective prevention and cessation).

In the PQO context, however, setting up the EP as in (1) is tricky because defining the $\mathbf{Y}$, Y_{X^*} and Y in the GPOF-based PQO context (and, therefore the outcome in the S+B illustration) is not as straightforward.

The main issues in this regard clearly emerge when one attempts to define the $\mathbf{Y}$ in the S+B context. One cannot simply define it to be *birth weight for the live births* because that definition implicitly conditions on the occurrence of a live birth. This qualitative event is, however, itself affected by the $\mathbf{X}$. So, for instance, if we try to set up a counterfactual in which smoking levels are increased, we are confronted with the possibility that some pregnancies that would have resulted in live births (with manifested birth weight) in the pre-counterfactual scenario would have ended in non-live births in the post-counterfactual scenario (without a manifested birth weight) [assuming that smoking increases the likelihood of a non-live birth]. How can a meaningful EP be defined based on such an amorphous counterfactual? This is an example of the conceptual difficulty caused by so-called *bad control* – attempting to condition the analysis on a variate that is itself affected by the $\mathbf{X}$ (mentioned in 3.1 and 3.2).[11] As I will argue from a technical but practical perspective, bad control also precludes identification of the EP (regardless of how that EP is defined). As part of this research, definitions for the $\mathbf{Y}$, Y_{X^*} and Y are proposed in the GPOF-based PQO context that overcome this conceptual and practical impediment. I will discuss such definitions in detail later.

In the next section, I propose a version of the $\mathbf{Y}$ that is real-valued and is not subject to the bad control critique because it takes direct account of the effect of the $\mathbf{X}$ on the probability of occurrence for the qualitative event. In the S+B context, this $\mathbf{Y}$ accounts for the effect of smoking on the likelihood of a live birth. Although the proposed definition for

[11] In the context of S+B, the extant literature is subject to the bad control critique as it focuses on birth weight conditional on live birth (see Rosenzweig & Schultz, 1983; Evans & Ringel, 1999; Lumley et al., 2004; Lien & Evans, 2005; Abrevaya and Dahl, 2008). Therefore, results obtained in these studies are really not causally interpretable.

the $\mathbf{Y}$ does not have an observable version, as we will see, it yields a legitimate specification for the relevant EP and affords consistent estimation of that parameter using the observable data.

3.5 Extending the CPOM to the PQO Setting

As discussed in section 3.4 above, finding a clear, rigorous, and useful definition for the $\mathbf{Y}$ is difficult if it is partially qualitative. Without a definition for the $\mathbf{Y}$, specification of the CPOM in the GPOF as exemplified in (2) is not possible. In particular, if the $\mathbf{Y}$ is not real valued, specifying a pdf like (2) is not possible. Moreover, the only way to define the $\mathbf{Y}$ as real-valued in the PQO case is to set it as conditional on occurrence of the qualitative event. As we have discussed, however, this approach is plagued by the bad control problem. I seek a way around this apparent dilemma by extending the basic CPOM concept to allow for a PQO.

Recall that we are assuming that the subpopulation of interest, on which the empirical causal analysis is focused, comprises those for whom a quantitative outcome would actually (factually) be observed. For this subpopulation, I seek to specify, identify, estimate and conduct inference for a parameter characterizing the causal effect of a counterfactually imposed change in a presumed causal variable (X^*) on an outcome of interest (Y_{X^*}). For the sake of illustration, let us focus on the following version of the AIE as the effect parameter of interest

$$AIE(\Delta) = E[Y_{X^{pre}+\Delta}|Q = 0] - E[Y_{X^{pre}}|Q = 0] \tag{8}$$

where X^{pre} and Δ are defined above and $Y_{X^{pre}}$ and $Y_{X^{pre}+\Delta}$ denote the observed versions of the $\mathbf{Y}$ that correspond with X^{pre} and $X^{pre} + \Delta$, respectively. Q indicates the observability of the qualitative component (via the DGP) of the PQO. Conditioning on Q = 0 is tantamount to conditioning on the subpopulation for whom a quantitative outcome would actually manifest. In the S+B illustrative example, Q = 0 denotes the subpopulation of pregnancies that end in a live birth. Note that I have not yet given specificity to the definition of $(Y_{X^*}|Q = 0)$ [and, therefore, $(Y_{X^{pre}}|Q = 0)$] in the present PQO context. I now turn to this issue.

To fix ideas and to build upon the GPOF discussion in chapter 2, for the Q = 0 subpopulation, consider a random experiment corresponding to a counterfactually imposed version of X^* whose sample space comprises the union of a *qualitative* event $\mathcal{Q}_{X^*}$ (not a real value) and a specified subset of the real line ($\mathbb{R}$), say $\mathcal{R}_{X^*}$ whose typical element is r_{X^*} (note that it is possible that $\mathcal{R}_{X^*} = \mathbb{R}$). Correspondingly, $(Q_{X^*}|Q = 0)$ is defined to be the dichotomous random variable characterizing the stochastics of the qualitative component of the hypothetical experiment [$(Q_{X^*} \mid Q = 0) = 1$ if the qualitative potential outcome would occur and $(Q_{X^*} \mid Q = 0) = 0$ if the quantitative potential outcome would occur]. In the S+B context, $(Q_{X^*} \mid Q = 0) = 0$ if the pregnancy ends in a live birth at the counterfactually mandated level of smoking during pregnancy for those pregnancies that actually (factually) end in live birth. Finally, $(R_{X^*} \mid Q = 0, Q_{X^*} = 0)$ is defined to be the random variable characterizing the quantitative component of the counterfactual experiment conditional on non-occurrence of the qualitative potential outcome; the support of $(R_{X^*} \mid Q = 0, Q_{X^*} = 0)$ is $\mathcal{R}_{X^*}$. In the S+B example, $(R_{X^*}|Q = 0, Q_{X^*} = 0)$ is the potential birth weight of pregnancies that would end in live birth at the counterfactually mandated level of smoking during pregnancy for pregnancies that actually (factually) ended in live birth.[12]

In this context, one's first inclination for specifying $(Y_{X^*} \mid Q = 0)$ might be to set it equal to $(R_{X^*} \mid Q = 0, Q_{X^*} = 0)$ so that (8) would be rewritten as

$$\mathrm{AIE}(\Delta)^{\dagger} = E\left[R_{X^{pre} + \Delta} \middle| Q = 0, Q_{X^{pre} + \Delta} = 0\right] - E\left[R_{X^{pre}} \middle| Q = 0, Q_{X^{pre}} = 0\right] \qquad (9)$$

The problem with (9) is that it fixes $Q_{X^{pre}}$ and $Q_{X^{pre} + \Delta}$ at 0 and thus fails to account for the possible impact of the counterfactual change in X^{pre} on $Q_{X^{pre}}$ – an effect that will surely influence the potential outcome (whatever its definition). For this reason, (9) will be biased as an effect parameter intended to characterize the causal effect of the posited counterfactual change in X^{pre} on the quantitative potential outcome of interest (whatever its definition). It is this bias that prompts the use of the term *bad control* in describing the conditioning on $Q_{X^{pre}}$ and $Q_{X^{pre} + \Delta}$ that is required by (9). In general, bias due to bad control can be expected in any empirical causal analysis in which some elements of the vector of control variables can themselves be characterized as potential outcomes that would be impacted by the posited counterfactual change in the presumed causal variable. In the present context, I seek to avoid bad control by directly including the possible impact of counterfactual differences in the qualitative potential outcome in the specification of the

[12] To clarify this notation, consider the description for $(R_{X^*} | Q = 1, Q_{X^*} = 0)$. It is the potential birth weight of pregnancies that would have ended in live birth at the counterfactually mandated level of smoking during pregnancy but did not actually end in live birth. This implies that those pregnancies must have had levels of smoking during pregnancy that is different from X^*.

quantitative potential outcome of interest. To wit, the following specific definition for the generic term $(Y_{X^*} \mid Q = 0)$ in (8) is proposed.

$$(Y_{X^*} \mid Q = 0) \equiv \Pr(Q_{X^*} = 0 \mid Q = 0) \times (R_{X^*} \mid Q = 0, Q_{X^*} = 0) \tag{10}$$

where, $\Pr(Q_{X^*} = 0 \mid Q = 0)$ denotes the probability that, for a given version of X^*, the counterfactual quantitative outcome would manifest for the subgroup of the population for whom the quantitative outcome is actually (factually) observable. In the illustrative example of S+B, (10) is the potential birth weight of pregnancies that would end in live birth at X^* weighted by the probability of a live birth outcome for those pregnancies that actually ended in live birth. Clearly, (10) is designed to explicitly incorporate the impact of X^* on Q_{X^*}; thereby avoiding the bad control critique. Rewriting (8) accordingly we obtain

$$\begin{aligned} AIE(\Delta) = {} & E\left[\Pr(Q_{X^{pre}+\Delta} = 0 \mid Q = 0) \times (R_{X^{pre}+\Delta} \mid Q = 0, Q_{X^{pre}+\Delta} = 0)\right] \\ & - E\left[\Pr(Q_{X^{pre}} = 0 \mid Q = 0) \times (R_{X^{pre}} \mid Q = 0, Q_{X^{pre}} = 0)\right] \end{aligned} \tag{11}$$

Let us now turn to the estimation of (11) via regression methods.

As discussed in the chapter 2, in the GPOF, estimation of a CI parameter is predicated on an appropriately designed CPOM. The present PQO context is, however, a bit unorthodox. First, the relevant potential outcome as defined in (10) does not have a directly observable counterpart in the DGP [although, as we will see later, its components are identified (non-parametrically and parametrically)]. Secondly, and related to this, is the fact that the relevant CPOM will be defined in three parts. For the first component of the CPOM I assume that conditional on a vector of controls V

$$\Pr(Q_{X^*} = 0 \mid V, Q = 0) = \mathcal{P}(V, X^*; \tau_{\mathcal{Q}}) \tag{12}$$

where $\mathcal{P}(.\ ,\ .\ ;\ .)$ is a known function whose range is the unit interval and $\tau_{\mathcal{Q}}$ is a vector of unknown parameters.[13] This function is typically specified in terms of a cumulative distribution function (cdf). Secondly, the pdf for the quantitative component is specified as

$$\text{pdf}(R_{X^*} \mid V, Q = 0, Q_{X^*} = 0) = g(r_{X^*}, X^*, V; \tau_g) \tag{13}$$

where $g(r_{X^*}, X^*, V; \tau_g)$ is a known proper parametric pdf whose support is $\mathcal{R}_{X^*}$ ($r_{X^*} \in \mathcal{R}_{X^*}$) and τ_g is an unknown parameter vector.[14] Combining (11) through (13) and applying the law of iterated expectations we obtain

$$\text{AIE}(\Delta) = E\Big[E\big[\Pr(Q_{X^{\text{pre}}+\Delta} = 0 \mid V, Q = 0) \times (R_{X^{\text{pre}}+\Delta} \mid V, Q = 0, Q_{X^{\text{pre}}+\Delta} = 0)\big]$$

$$- E\big[\Pr(Q_{X^{\text{pre}}} = 0 \mid V, Q = 0) \times (R_{X^{\text{pre}}} \mid V, Q = 0, Q_{X^{\text{pre}}} = 0)\big]\Big]$$

$$= E\big[\mathcal{P}(V, X^{\text{pre}} + \Delta; \tau_{\mathcal{Q}}) \times E\big[R_{X^{\text{pre}}+\Delta} \mid V, Q = 0, Q_{X^{\text{pre}}+\Delta} = 0\big]$$

---

[13] To be more explicit, we might have written

$$\mathcal{P}(V, X^*; \tau_{\mathcal{Q}}) = \mathcal{P}_{(Q_{X^*} = 0 \mid V, Q = 0)}(V, X^*; \tau_{\mathcal{Q}})$$

[14] To be more explicit, we might have written

$$g(r_{X^*}, X^*, V; \tau_g) = g_{(R_{X^*} \mid V, Q = 0, Q_{X^*} = 0)}(r_{X^*}, X^*, V; \tau_g).$$

$$- \mathcal{P}(\mathrm{V}, \mathrm{X}^{\mathrm{pre}}; \tau_{\mathcal{Q}}) \times \mathrm{E}\left[\mathrm{R}_{\mathrm{X}^{\mathrm{pre}}} \mid \mathrm{V}, \mathrm{Q} = 0, \mathrm{Q}_{\mathrm{X}^{\mathrm{pre}}} = 0\right]\Big]$$

$$= \mathrm{E}\left[\mathcal{P}(\mathrm{V}, \mathrm{X}^{\mathrm{pre}} + \Delta; \tau_{\mathcal{Q}})\, m(\mathrm{V}, \mathrm{X}^{\mathrm{pre}} + \Delta; \tau_{g}) - \mathcal{P}(\mathrm{V}, \mathrm{X}^{\mathrm{pre}}; \tau_{\mathcal{Q}})\, m(\mathrm{V}, \mathrm{X}^{\mathrm{pre}}; \tau_{g})\right]$$

(14)

where

$$m(\mathrm{V}, \mathrm{X}^{*}; \tau_{g}) = \mathrm{E}\left[\mathrm{R}_{\mathrm{X}^{*}} \mid \mathrm{V}, \mathrm{Q} = 0, \mathrm{Q}_{\mathrm{X}^{*}} = 0\right] = \int_{\mathcal{R}_{\mathrm{X}^{*}}} r_{\mathrm{X}^{*}}\, g(r_{\mathrm{X}^{*}}, \mathrm{X}^{*}, \mathrm{V}; \tau_{g})\, \mathrm{d}r_{\mathrm{X}^{*}}$$

The third component of the CPOM is

$$\mathrm{pmf}(\mathrm{Q}_{\mathrm{X}^{*}} \mid \mathrm{V}) = h(\mathrm{Q}_{\mathrm{X}^{*}}, \mathrm{V}, \mathrm{X}^{*}; \tau_{\mathcal{Q}}) = \mathcal{P}(\mathrm{V}, \mathrm{X}^{*}; \tau_{\mathcal{Q}})^{1-\mathrm{Q}_{\mathrm{X}^{*}}} \left[1 - \mathcal{P}(\mathrm{V}, \mathrm{X}^{*}; \tau_{\mathcal{Q}})\right]^{\mathrm{Q}_{\mathrm{X}^{*}}}$$

(15)

which says that the parametric specification for the likelihood of the qualitative potential outcome in the first component of the CPOM for the relevant subpopulation (Q = 0) as given in (12), also holds true for the population at large (i.e., regardless of the value of Q).

Now if we had consistent estimates of the "deep" parameters of the model $\tau_{\mathcal{Q}}$ and τ_{g}, (say $\hat{\tau}_{\mathcal{Q}}$ and $\hat{\tau}_{g}$) we could consistently estimate the targeted effect parameter in (14) as

$$\mathrm{AIE}(\Delta) = \sum_{\mathrm{Q}=0} \frac{1}{\mathrm{n}_{\mathrm{Q}=0}} \left\{\mathcal{P}\left(\mathrm{V}_{\mathrm{i}}, \mathrm{X}_{\mathrm{i}}^{\mathrm{pre}} + \Delta_{\mathrm{i}}; \hat{\tau}_{\mathcal{Q}}\right) m\left(\mathrm{V}_{\mathrm{i}}, \mathrm{X}_{\mathrm{i}}^{\mathrm{pre}} + \Delta_{\mathrm{i}}; \hat{\tau}_{g}\right)\right.$$

$$\left. - \mathcal{P}\left(\mathrm{V}_{\mathrm{i}}, \mathrm{X}_{\mathrm{i}}^{\mathrm{pre}}; \hat{\tau}_{\mathcal{Q}}\right) m\left(\mathrm{V}_{\mathrm{i}}, \mathrm{X}_{\mathrm{i}}^{\mathrm{pre}}; \hat{\tau}_{g}\right)\right\} \qquad (16)$$

where $\Sigma_{\mathrm{Q}=0}$ indicates summation over the subsample of observations for whom $\mathrm{Q}_{\mathrm{i}} = 0$ and $\mathrm{n}_{\mathrm{Q}=0}$ denotes the size of that subsample.

Before embarking on the consistent estimation of the deep parameters of the model, identification must be established at the two levels as discussed in sections 2.3. First, aside from any particular parametric specification, the CPOM must be shown to be non-parametrically identified. In the context of the CPOM that I have posited here, according to Terza (2019a), non-parametric identification is established if it can be shown that the following DGP versions of (13) and (15) are valid

$$\mathrm{pdf}(R \mid V, Q = 0) = \mathcal{g}(r, X, V; \tau_g) \tag{17}$$

and

$$\mathrm{pmf}(Q \mid V) = \mathcal{h}(Q, V, X; \tau_Q) = \mathcal{P}(V, X; \tau_Q)^{1-Q}\left[1 - \mathcal{P}(V, X; \tau_Q)\right]^Q \tag{18}$$

In other words, the model (i.e. the CPOM) is non-parametrically identified if the relevant aspects of the DGP can be obtained by simply substituting the observable versions of the $\mathbf{X}$ and the relevant components of the $\mathbf{Y}$ (viz., X, Q and R) for their counterfactual counterparts (viz., X^* , Q_{X^*} and R_{X^*}) into the CPOM. As presented in section 2.3, the most important of these conditions, and the least likely to be true, is CIND between the potential outcomes and X given the vector of controls V. In the present context, CIND holds if Q_{X^*} and R_{X^*} are independent of X conditional on V. Among the list of sufficient conditions for non-parametric identification, CIND is the most troublesome because it is not only unlikely to be true but it is also untestable. For the present discussion, I will maintain that CIND and the other conditions for non-parametric identification [(as discussed in Terza (2019a) and reviewed in section 2.3] hold.

With non-parametric identification maintained, the second level of identification to be established is parametric identification.[15] As discussed in section 2.3.2, the parameters of the DGP model are identified if the chosen functional forms for the relevant aspects of the DGP (e.g. conditional mean, higher-order conditional moments, conditional pmf/pdf, etc.) are such that full knowledge of the values of those aspects of the DGP would imply knowledge of the values of the relevant parameters. Functional forms for $\mathcal{P}(\mathrm{V}, \mathrm{X}; \tau_{\mathcal{Q}})$ and $g(r, \mathrm{X}, \mathrm{V}; \tau_g)$ are chosen that afford parametric identification of $\tau_{\mathcal{Q}}$ and τ_g.

Given that the CPOM is non-parametrically and parametrically identified, based on (17) and (18), the parameter vectors $\tau_{\mathcal{Q}}$ and τ_g can be estimated using the following MLE

$$\hat{\tau}_{\mathcal{Q}} = \underset{\check{\tau}_{\mathcal{Q}}}{\arg\max} \sum_{i=1}^{n} q_{\mathcal{Q}}(\check{\tau}_{\mathcal{Q}}, Z_{\mathcal{Q}i}) \tag{19}$$

$$\hat{\tau}_g = \underset{\check{\tau}_g}{\arg\max} \sum_{i=1}^{n_{Q=0}} q_g(\check{\tau}_g, Z_{gi}) \tag{20}$$

where, n denotes the size of the full sample, $n_{Q=0}$ is the size of the subsample for whom $Q = 0$, $q_{\mathcal{Q}}(\check{\tau}_{\mathcal{Q}}, Z_{\mathcal{Q}i}) = \ln\left[h(Q_i, V_i, X_i; \check{\tau}_{\mathcal{Q}})\right]$, $q_g(\check{\tau}_g, Z_{gi}) = \ln\left[g(R_i, V_i, X_i; \check{\tau}_g)\right]$, $Z_{\mathcal{Q}i} = [Q_i \quad V_i \quad X_i]$ and $Z_{gi} = [R_i \quad V_i \quad X_i]$. The asymptotic standard errors of the estimators in (19) and (20) can be obtained by using the approach in Terza (2017).

3.6 Simulation Study

In this section, I demonstrate the implementation of the estimators in (19) and (20) for the PQO model and validate the consistency of the proposed AIE estimator using

[15] Note that if one cannot establish non-parametric identification as detailed in Terza (2019a) then subsequent discussion of parametric identification have no useful content from the perspective of causal inference.

simulated data. For the sampling design of the simulation study, consider the case in which (12) and (13) are specified such that

$$\mathcal{R}_{X^*} \equiv (0, \infty)$$

$$\mathcal{P}(V, X^*; \tau_{\mathcal{Q}}) \equiv \mathcal{M}\{(Q_{X^*} \mid Q = 0) = 0\} = \Phi(V\tau_{QV} + X^*\tau_{\mathcal{Q}X})$$

$$g(r_{X^*}, X^*, V; \tau_g) = gg(Y_{X^*}; V\tau_{gV} + X^*\tau_{gX}, \kappa, \sigma)$$

where

$$gg(A; b, c, d) = \frac{\nu^{\nu}}{dA\sqrt{\nu}\Gamma(\nu)} \exp(z\sqrt{\nu} - u) \qquad (21)$$

denotes the pdf of a generalized gamma variate A with parameters b, c and d; $\nu = |\kappa|^{-2}$, $z = \frac{\text{sign}(\kappa)[\log(A) - b]}{d}$, $u = \nu \times \exp(|\kappa|z)$, and $\Gamma(\ )$ is the gamma function. The three parameter Generalized Gamma (GG) distribution is chosen for simulating values for the quantitative component because it is very flexible distribution that subsumes several known distributions that are commonly used for non-negative random variables, such as the Weibull, Exponential and Log-normal among others.

3.6.1 The Simulated Data Generator

The data generator has two parts, i.e., (17) and (18). (18) generates a binary outcome indicator for the qualitative component of the model representing whether or not a pregnancy ends in a live birth. For this part, data will be generated for the full sample following the assumption in (15) that pregnancies that do and do not end in live birth have the same parametric specification for the likelihood of the qualitative potential outcome.

The second part, given in (17), generates strictly positive values for the quantitative component representing birth weight values of the pregnancies that end in live birth. Note that for this component, data is generated for the sub-sample of the pregnancies for which the simulated outcome for the first part of the generator is one that resulted in live births.

A uniformly distributed random variate data generator is used to simulate the data vectors V^o and X^o for the observable random variables V and X, respectively. To obtain these vectors, I specify the mean and variance of V^o and X^o. The change in the policy for the targeted AIE is given by $\Delta^o = -X^o$. Note that the post-intervention vector, $\Delta^o + X^o$, is a null vector. For the qualitative component, I specify an indicator function

$$(Q_{X^*} \mid Q = 0) = 0 \;\text{ if }\; I(V^o\tau^o_{\mathcal{Q}V} + X^o\tau^o_{\mathcal{Q}X} + \varepsilon > 0) = 1 \qquad (22)$$

where, $\varepsilon \sim N(0, 1)$, $\tau^o_Q = [\tau^o_{QX} \quad \tau^o_{QVo}{}']$ where $\tau^o_{QVo}{}' = [\tau^o_{QV} \quad \tau^o_{Qo}]$ denotes the vector of specific parameter values chosen for the sampling design of the simulation (detailed later).

Because birth weight is a strictly positive outcome, the quantitative component of the data generator simulates values from a GG random variable. Since the GG is a continuous random variable, I rely on the inverse transform theorem to generate values of the random variate (Ross, 1997). To implement the inverse transform method, I need to know the cdf of the GG variate and its inverse. According to Stacy and Mihram (1965), the conditional cdf of the GG variable is

$$GG_{(Y_{X^*} \mid V)}(Y^o, V^o, X^o, \tau^o_g, \kappa^o, \sigma^o) = \begin{cases} \dfrac{\gamma(\nu, (Y^o/a)^p)}{\Gamma(\nu)} & \text{if} \quad p > 0 \\[2ex] 1 - \dfrac{\gamma(\nu, (Y^o/a)^p)}{\Gamma(\nu)} & \text{if} \quad p < 0 \end{cases} \qquad (23)$$

$\gamma(b', c')$ denotes the incomplete gamma function defined as $\gamma(b', c') = \int_0^{c'} t^{b'-1} e^{-t} dt$,

$a = \frac{\exp(V^o \tau^o_{gV} + X^o \tau^o_{gX})}{\left(\frac{1}{|\kappa^o|^2}\right)^{\frac{\sigma^o}{\kappa^o}}}$, $\nu = \frac{1}{|\kappa^o|^2}$ and $p = \frac{\kappa^o}{\sigma^o}$. $\tau^o_g = [\tau^o_{gX} \quad \tau^o_{gVo}{}']$ where $\tau^o_{QVo}{}' = [\tau^o_{gV} \quad \tau^o_{go}]$ is the coefficient vector for the linear index component in a. Note that $\frac{\gamma(\nu, (Y^o/a)^p)}{\Gamma(\nu)} = SG(\nu, (Y^o/a)^p)$, where SG(h, j) is the standard gamma cdf evaluated at h with shape parameter j. Manning et al (2005) provides a crosswalk between this parameterization (a, ν and p) and the one introduced in the expression for the GG pdf in (21). From (23) it can be shown that when $p > 0$, we have

$$Y^o = GG^{-1}_{(Y_{X^*} \mid V)}(U[0, 1], V^o, X^o, \pi^o_g, \kappa^o, \sigma^o)$$

$$= a \times \gamma^{-1}(\nu, \Gamma(\nu) U[0, 1])^{\frac{1}{p}} \qquad (24)$$

where $GG^{-1}_{(Y_{X^*} \mid V)}(U[0, 1], V^o, X^o, \pi^o_g, \kappa^o, \sigma^o)$ denotes the inverse of the GG cdf as given in (23), $\gamma^{-1}(d', j)$ denotes the inverse incomplete gamma defined such that if $j = \gamma(d', k)$, then $k = \gamma^{-1}(d', j)$ and U[0, 1] is a unit uniform variate. It is not easy to get the inverse incomplete gamma function directly in Stata/Mata, but one can get it indirectly. Yang (2016) derived the following equivalent expression for (24) which can be simply implemented in Stata/Mata

$$Y^o = a \times SG^{-1}(\nu, U[0, 1])^{\frac{1}{p}} \qquad (25)$$

where $SG^{-1}(a, U[0, 1])$ denotes the inverse standard gamma variate with shape parameter a. Using (25) and by choosing the values $\tau^o_{\mathcal{G}}$, κ^o, σ^o, I generate the sample values of the quantitative component for the subsample such that $(Q_{X^*} \mid Q = 0) = 0$.

Once the two components of the data are generated, we obtain the true AIE in (14) using the specifications

$$\mathcal{P}(V^o, X^o; \tau^o_{\mathcal{Q}}) = \Phi(V^o\tau^o_{\mathcal{Q}V} + X^o\tau^o_{\mathcal{Q}X})$$

and

$$\mathcal{m}(V^o, X^o; \tau^o_{\mathcal{G}}) = \exp\Bigg[V^o\tau^o_{\mathcal{G}V} + X^o\tau^o_{\mathcal{G}X} + \left(\frac{\sigma^o}{\kappa^o}\right) \ln((\kappa^o)^2)$$
$$+ \ln\left(\Gamma\left\{\left(\frac{1}{(\kappa^o)^2}\right) + \left(\frac{\sigma^o}{\kappa^o}\right)\right\}\right) - \ln\left(\Gamma\left\{\left(\frac{1}{(\kappa^o)^2}\right)\right\}\right)\Bigg]$$

where $\Phi(\ )$ and $\mathcal{m}(\ )$ denote mean values of the binary and the GG variates, respectively.

3.6.2 The Sampling Design and the Simulation Results

The main motives for the simulation are to demonstrate the implementation of the PQO modeling using the estimators in (19) and (20), and to validate the consistency of the EP in (14). For the linear index component coefficient vector of the probit specification $\tau^o_{\mathcal{Q}} = [\tau^o_{\mathcal{Q}X} \quad \tau^o_{\mathcal{Q}V^o}{}']$ where $\tau^o_{\mathcal{Q}V^o}{}' = [\tau^o_{\mathcal{Q}V} \quad \tau^o_{\mathcal{Q}o}]$ for which the following parameter vector of values is specified.

$$\tau^o_{\mathcal{Q}} = [0.15 \qquad -0.5 \qquad 12]$$

Here, I assign $\tau^o_{\mathcal{Q}X} = 0.15$ for X^o in the probit model to mimic the evidence that smoking during pregnancy increases [decreases] the probability that a pregnancy ends in a non-live birth [live birth]. To obtain sample values of the GG variate representing birth weight for pregnancies that end in live birth, I specify the parameter values for the linear index component coefficient vector $\tau_g^o = [\tau_{gX}^o \quad \tau_{gV^o}^o{}']$ where $\tau_{QV^o}^o{}' = [\tau_{gV}^o \quad \tau_{go}^o]$ and for the two ancillary parameters κ^o and σ^o as follows:

$$\tau_g^o = [-0.004 \quad 0.002 \quad 8]$$

$$\kappa^o = 0.95 \text{ and } \sigma^o = 0.175$$

Like I do for the qualitative component, I impose a negative value ($\tau_{gX}^o = -0.004$) for X^o in the GG model to allow a reduction in birth weight due to smoking during pregnancy. To obtain V^o and X^o, I specify the mean and variances as $E[V^o] = 27$, $E[X^o] = 1.75$, $Var[V^o] = 49$ and $Var[X^o] = 1$, where E[ ] and Var[ ] denote the mean and variance functions. These values are chosen to get a distribution closer to birth weight for the outcome variable. As it can be seen on table A1 in appendix I, the average value of Y^o is 2939 which is somehow close to an average birth weight in grams.

After simulating the data vectors V^o, X^o, and Y^o for $(Q_{X^*} \mid Q = 0) = 0$, I obtain estimated values for the probit and the GG model deep parameters by applying the M-estimator in (19) and (20). Based on these values, I estimated the AIE and compared it to the true AIE obtained by plugging the specified parameter values into (16). A super sample of 2,000,000 observations is used to calculate the true AIE. To examine the consistency of the AIE estimator, I simulated samples of increasing size using the data generators detailed above. I then applied the MLE estimator for the targeted EP and calculated the absolute percentage bias (APB) of the estimated AIE for each of the simulated data of sample size n. The APB is calculated using the formula

$$\mathrm{APB} = \left|\frac{\mathrm{AIE}(\Delta)^{\mathrm{est}} - \mathrm{AIE}(\Delta)^{\mathrm{true}}}{\mathrm{AIE}(\Delta)^{\mathrm{true}}}\right| \times 100\% \qquad (26)$$

where $\mathrm{AIE}(\Delta)^{\mathrm{est}}$ and $\mathrm{AIE}(\Delta)^{\mathrm{true}}$ denote the estimated and the true AIE values, respectively.

Table 1 presents the results of the simulation. In general, the results provide evidence for the consistency of the AIE estimator. For instance, the APB of the estimated AIE is 11.7% when the model is simulated using a sample of 25,000 observations for the entire population of which 15181 observations are used to estimate the deep parameters of the GG model and the corresponding AIE. This APB further decreases to 5.4% [0.7%] when the entire sample size increases to 100,000 [250,000] of which the relevant subpopulation has 60,168 [150,067] observations. Therefore, the PQO modeling approach and the ML estimator I proposed provide consistent estimate for the specified AIE.

3.7 Application: Smoking and *Natility-Weighted Birth Weight*

This section presents an empirical application of the proposed PQO regression model to estimate the *natility-weighted birth weight* effect of a counterfactual intervention that fully and effectively bans smoking during pregnancy. I also estimate the AIE of the same hypothetical intervention on the conventional measure of birth weight that is subject to a bias due to bad control.

I use a probit specification to model the probability that a pregnancy ends in a live birth and a GG model to characterize birth weight for the sub-population of pregnancies that end in live birth. In particular, I estimate the AIE in (14) based on the M-estimators in (19) and (20) and using the following specific functional forms for the probability that a pregnancy ends in a live birth and for the conditional mean birth weight given that the pregnancy ends in a live birth

$$\mathcal{P}(V, X^*; \tau_Q) = \Phi(V\tau_{QV} + X^*\tau_{QX})$$

$$m(V, X^*; \tau_g) = \exp\Big[V\tau_{gV} + X^*\tau_{gX} + \left(\frac{\sigma}{\kappa}\right) \ln((\kappa)^2)$$

$$+ \ln\left(\Gamma\left\{\left(\frac{1}{(\kappa)^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\left(\frac{1}{(\kappa)^2}\right)\right\}\right)\Big]$$

Apart from the bad control problem, as discussed in section 3.5, estimating causally interpretable AIE requires CIND between the potential PQO (Q_{X^*} and R_{X^*}) and X conditional on V. In other words, the potential resolution of a pregnancy (whether or not it ends in a live birth) and potential birth weights are independent of the observed smoking level given the vector of observed variates. This is unlikely to hold in observational setting as one can imagine a number of unobserved factors such as propensity to engage in risky behavior, maternal health endowment and so on that correlate with both prenatal smoking and pregnancy as well as birth outcomes (Grossman & Joyce, 1990). This implies that smoking is endogenous.

Studies in the conventional birth weight literature find mixed evidence on the endogeneity of smoking during pregnancy. Using a difference-in-differences approach, Fertig (2010), for instance, finds that smoking was associated to a 261 grams reduction in birth weight in the year 2000 up from a reduction of 160 grams in 1958 sample. She argues that as information about the harms of cigarette smoking become widespread, women from higher socioeconomic status quit smoking at a larger rate implying that the observed association in the result from the year 2000 is likely confounded by unobserved factors that is correlated with smoking behavior of pregnant women from low socioeconomic status group. Evans & Ringel, (1999) use tax hike as instrumental variable and find that smoking during pregnancy reduces birth weight between 300 grams and 600 grams, a significantly larger reduction than the Ordinary Least Squares (OLS) estimates. This result suggests that there is probably advantageous selection – women with worse maternal health tend to quit smoking more than pregnant women whose health endowment during pregnancy is higher. On the other hand, Lien and Evans (2005) used a cigarette tax hike in four states and matched each state to similar control states. They find that smoking reduces birth weight by 185 grams. Their IV estimate was not, however, statistically different from the OLS estimate suggesting that smoking is perhaps exogenous. In addition to the fact that these studies are subject to bias due to bad control, the inconsistency in the results indicate that selection into smoking is also a concern.

3.7.1 Data Source and Descriptive Statistics

I use data from the National Survey of Family Growth (NSFG). NSFG is a nationally representative of women 15-44 years of age. From 2015 onwards the age range expanded to 15-49. Unlike many other data sources, NSFG asks each respondent woman detail information about her pregnancy history covering five years prior to the interview date. Importantly, it has information about conception and end of month information on each pregnancy, how each pregnancy ended and the smoking behavior of the mother during each pregnancy. I use data from survey periods 2002, 2006-2010, 2011-2013 and 2013-2015.

The outcome variables I consider are an indicator for whether the pregnancy ends in a live birth or not, and birth weight in grams for those pregnancies that end in live birth. The number of cigarettes smoked per day is the main policy variable in the empirical analysis. Given the possibility for endogeneity of observed level of smoking during pregnancy, as discussed in chapter 2, one would need to expand the DGP by exploiting an appropriate empirical identification strategy such as state cigarette tax rate as an instrument to deal with the endogeneity of smoking. Because I use the public access version of the data that do not have state identifier for pregnancies, I am not able to match state cigarette tax information to the pregnancies in the dataset. Therefore, the result can have a causal interpretation only under the strong assumption that smoking during pregnancy is exogenous conditional on the vector of observable control variables such as marital status during pregnancy, age at pregnancy, educational attainment, and race.

Table 2 presents the descriptive statistics of the data for the full sample and by live birth status of pregnancies. For the full sample, 21.6% of pregnancies do not end in live birth. As expected, relative to women who had a live birth, smoking rate is higher (by 5 percentage point) among women whose pregnancies did not end in live birth. Table 3 compares live birth outcome, birth weight and other characteristics of the sample by smoking status. Women who smoke during pregnancy have a 7.2 percentage point lower live birth rate than those who do not smoke. Among women who have live births, the average birth weight was 151.5 grams (0.33 pound) lower for infants born from mothers who were smoking during pregnancy.

3.7.2 Estimation Results

Table 4 presents the deep parameter estimates for the probit and the GG regression models. Columns 1 and 2 show the deep parameter estimates of the probit and the GG specifications, respectively. It can be seen that the coefficient on the number of smoking variable in both the qualitative and quantitative components of the PQO model are negative and statistically significant implying that smoking has a negative effect on *natility-*

weighted birth weight. The statistically significant coefficient in the probit model provides empirical evidence that bad control is indeed a concern in the context of S+B, suggesting that one cannot interpret an EP based only of the quantitative component of the model.

In column 1 of table 5, I present the AIE for the *natility-weighted birth weight* that has a direct causal interpretation (under exogeneity of smoking during pregnancy) and the AIE estimated based on the conventional measure of birth weight that is subject to the bad control problem. The latter is biased because it ignores the qualitative component that accounts for the effect of smoking during pregnancy on the probability that the pregnancy ends in a live birth. The estimated AIE shown in column 1 suggests that a fully effective smoking ban during pregnancy would increase *natility-weighted birth weight* by 32.7 grams (0.072 pounds). The result in column 2 shows that the policy increases the conventional measure of birth weight by 11.24 grams (0.025 pounds). This result, however, cannot have a causal interpretation even when smoking is truly exogenous.

It should be noted that the estimates in both cases are smaller than most of the estimates reported in the literature on the effect of smoking during pregnancy on birth weight. This is because the targeted effect parameter in our case applies to the entire population of pregnant women whose pregnancy ends in a live birth while the estimates in the conventional literature aims to target the subpopulation of pregnant women who actually smoked during pregnancy. In the language of the treatment effect literature, the latter is called treatment effect on the treated.

3.8 Summary, Discussion and Conclusion

A new regression-based approach is developed for specification, identification, and estimation of causally interpretable EP for an outcome of interest that manifests as either a value in a specified subset of the real line or a qualitative event -- a *partially qualitative outcome* (PQO). A PQO requires special attention because the only version of it to which the conventional form of the GPOF can be applied to is subject to bias due to the bad control problem. An outcome measure is proposed that maintains all of the essential features of a PQO but is entirely real-valued and is not subject to the bad control critique; the *P-weighted outcome* – the outcome weighted by the probability that it manifests as a quantitative (real) value. The practicality and usefulness of this new measure for specifying and identifying effect parameters that characterize the causal relationships between policy variables of interest and the PQO is discussed. Moreover, a regression-based estimation method is detailed for such effect parameters and, using simulated data, demonstrate its implementation and validate its consistency for the targeted effect parameter. The proposed approach is illustrated by conducting an empirical study to analyze a counterfactually mandated fully effective policy intervention that brings the smoking levels of all pregnant women down to zero on the *natility-weighted birth weight*. Using the NSFG public access data, I find that the smoking ban improves *natility weighted birth weight*, on average, by 32.7 grams over the entire sub-population of pregnancies that end in live birth. The corresponding AIE estimated for birth weight as in the conventional approach is an 11.24 grams increase in birth weight. While the latter result cannot be interpreted as causal effect, the causal interpretability of the AIE for *natility-weighted birth weight* relies on the assumption that smoking during pregnancy is exogenous conditional on the vector of observable variables included in the model. In practice, there may be several unobservable confounders such as risky behavior of pregnant women, maternal health endowment and so on that could invalidate this assumption, making causal interpretability of the estimates questionable. As an extension to this work, I plan to incorporate such endogeneity into the proposed regression framework in the future.

Chapter 4

Correcting Endogeneity Bias in Two-Part Models: Causal Inference from the Potential Outcomes Perspective

The two-part model (2PM) is one of the most widely applied empirical modeling and estimation framework in empirical health economics.[16] [17] It applies to cases in which the outcome variable is nonnegative with a non-trivial number of units having an observed value of zero. By design, the 2PM has two distinct components; a qualitative (binary) extensive margin (EM) characterizing an individual's participation (or not) in a specified activity, and an intensive margin (IM) representing the individual's level of activity (conditional on participation as determined at the EM). The 2PM allows the process that determines observed zero outcomes to systematically differ from that which determines non-zero observations.

In this chapter, a regression-based potential outcomes approach is developed to policy relevant causal inference in the context of 2PM in which the causal variable of interest is a continuous endogenous variable. In particular, the estimation objective is the AIE specified in (1). Endogeneity of the causal variable of interest is a common problem in 2PM like in any other econometric models that are specified to obtain CI parameter.

[16] The notion of explicitly accounting for a discrete mass at zero of a random variable was highlighted by Aitchison (1955). He derived the mean and variance of such random variables and demonstrated the enormous improvement in terms of obtaining a better fit by comparing the relative performance of fitting a truncated and a standard Poisson models to observed data on number of children in a household. The two-part model (2PM) was introduced in a regression framework by Cragg (1971).

[17] The seminal work by Duan et al (1983) was the first to apply the 2PM in health economics and health services research. Few of the numerous applications of the 2PM in health care and health service research include Biener et al 2020, Burney et al. 2016, Hyun et al. 2016, Li et al. 2016, Liu et al. 2010, Madden 2008, Buntin and Zaslavsky 2004, Ross and Chaloupka 2003, and Bradford et al. 2002.

Endogeneity is defined from the potential outcomes perspective detailed in chapter 2 as the partial observability of the unique conditioning vector, V, that induces CIND between the observed version of the $\mathbf{X}$ and the potential outcomes Y_{X^*}. With this definition, I extend the generic fully parametric 2PM (FP2PM) framework developed in Hao and Terza (2018) to encompass cases in which the $\mathbf{X}$ is endogenous. As pointed out in chapter 2, when endogeneity is present, an additional structure is needed to identify a CI parameter. The main objective of this chapter is specification of such a structure within the GPOF for the 2PM and a version of the estimator in (5) for the AIE in (1) that can be estimable using observable data. The chapter also highlights on the advantages of casting the causal inference problem in a FP2PM framework. These advantages include setting up of two important statistical tests. The performance of the proposed approach relative to other alternatives that are widely used to deal with endogeneity in the context of nonlinear models in general and the 2PM in particular is also demonstrated through a simulation study. In the last section, I implement the proposed approach to an empirical setting where the object of interest is estimation of the effect of a one unit increase in BMI on medical care spending among the youth in the US. I also estimate the medical care cost of a hypothetical change that moves all youth in the US from an average normal BMI to an average obese and severely obese BMI.

4.1 Specifying a Generic FP2PM with a Continuous Endogenous Variable within the GPOF

Here I extend the generic FP2PM specified within the GPOF in Hao and Terza (2018) to accommodate cases where the $\mathbf{X}$ is endogenous variable. This generic specification encompasses all the 2PM in the literature and also allows implementation of two important null hypotheses: "no 2PM is needed" and "the $\mathbf{X}$ is exogenous".

The salient feature of the 2PM is that the process that determines zero outcome values is allowed to systematically differ from that which determines strictly positive values for the outcome variable. Below I describe the specification of a generic FP2PM for the EM and IM components in which the $\mathbf{X}$ is a continuous endogenous variable. The model is specified using the GPOF notation to facilitate specification of a conditional mean function that is amenable for specifying CI parameter.

4.1.1 The Extensive Margin

$$Y_{X^*} = 0 \qquad \text{iff} \qquad \mathcal{U} < \mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X^*, X_o, X_u; \tau_{EM}) \tag{27}$$

where $\mathcal{U}$ is uniformly distributed on the unit interval and $\mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X^*, X_o, X_u; \tau_{EM})$ is the conditional cumulative density function (cdf) of $(\zeta^*_{EM} \mid X_o, X_u)$, written as a function of X_o, X_u and a vector of deep parameters for the EM, τ_{EM}, evaluated at ζ_{EM} (an unobserved parametric threshold). $\mathcal{G}^{EM}_{(A|C)}(A, B, C; \psi)$ denotes a conditional cdf of A conditional on C written as a function of A, B, C and the parameter vector ψ. To indicate that the $\mathbf{X}$ is endogenous, the unobservable scalar X_u is explicitly included in (27) and in subsequent specifications for the IM as well as for the conditional pdf detailed later.

4.1.2 The Intensive Margin

The generic FP2PM for the IM component is specified as follows:

$$\left(Y_{X^*} \mid Y_{X^*} > \zeta_{IM}\right) \text{ has a cdf } \mathcal{G}^{IM^*}\left(Y_{X^*}\right) \text{ iff}$$

$$\mathcal{U} \geq \mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X^*, X_o, X_u; \tau_{EM}) \qquad (28)$$

where $\mathcal{G}^{IM^*}\left(Y_{X^*}\right)$ is a shorthand for

$$\mathcal{G}^{IM^*}_{(Y_{X^*} \mid Y_{X^*} > 0, X_o, X_u)}\left(Y_{X^*}, X^*, X_o, X_u, \zeta_{IM}; \tau_{IM}\right)$$

$$= \frac{\mathcal{G}^{IM}_{\left(\zeta^*_{IM} \mid Y_{X^*} > 0, X_o, X_u\right)}(Y_{X^*}, X^*, X_o, X_u; \tau_{IM})}{1 - \mathcal{G}^{IM}_{(\zeta^*_{IM} \mid X_o, X_u)}(\zeta_{IM}, X^*, X_o, X_u; \tau_{IM})} \qquad (29)$$

$\mathcal{G}^{IM}_{\left(\zeta^*_{IM} \mid Y_{X^*} > 0, X_o, X_u\right)}(Y_{X^*}, X^*, X_o, X_u; \tau_{IM})$ is the specified conditional cdf of ζ^*_{IM} given $Y_{X^*} > 0$, X_o and X_u, written as a function of Y_{X^*}, X^*, X_o, X_u, an unknown parametric threshold ζ_{IM} and a vector of deep parameters of the distribution for the IM, τ_{IM}.

The systematic difference between the EM and IM components of the 2PM can arise from two sources. The first is when the structure (or the general function form) that generates the zero values differs from that which generates the strictly positive values. Second, even when there is no such structural difference (NSD) between the EM and the IM, the 2PM may still be needed if the deep parameters of the functional form for the two components differ. Unlike traditional 2PM specifications, the IM in (29) is written as a truncated cdf with a parametric and unknown truncation point. Hao and Terza (2018) show that such a truncated cdf specification for the IM is sufficient to set up a likelihood ratio test for the null that "there is no need for 2PM".

The conditional pdf of $\left(Y_{X^*} \mid X_o, X_u\right)$ can be written by combining (28) and (29) as

$$
\begin{aligned}
\mathrm{pdf}\left(Y_{X^*} \mid X_o, X_u\right) = {} & f_{\left(Y_{X^*} \mid X_o, X_u\right)}(Y_{X^*}, X^*, X_o, X_u; \pi) \\
& = \left[\, \mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X^*, X_o, X_u; \tau_{EM}) \,\right]^{I(Y_{X^*} = 0)} \\
& \times \Bigg[\left(1 - \mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X^*, X_o, X_u; \tau_{EM})\right) \\
& \times \frac{g^{IM}_{\left(\zeta^*_{IM} \mid Y_{X^*} > \zeta_{IM}, X_o, X_u\right)}(Y_{X^*}, X^*, X_o, X_u; \tau_{IM})}{1 - \mathcal{G}^{IM}_{(\zeta^*_{IM} \mid X_o, X_u)}(\zeta_{IM}, X^*, X_o, X_u; \tau_{IM})}\Bigg]^{1-I(Y_{X^*} = 0)}
\end{aligned}
\tag{30}
$$

$f_{(\,.\,)}(\,;\,)$ in the first equality denotes a known conditional pdf of Y_{X^*} given X_o and X_u, written as a function of Y_{X^*}, X_o, X_u and the vector of deep parameters $\pi = [\tau_{EM}{}' \quad \tau_{IM}{}']$. $g^{IM}_{\left(\zeta^*_{IM} \mid Y_{X^*} > \zeta_{IM}, X_o, X_u\right)}(Y_{X^*}, X^*, X_o, X_u; \tau_{IM})$ is the conditional pdf of Y_{X^*} given X_o and X_u for the subpopulation whose Y_{X^*} passes a parametric threshold ζ_{IM}. The indicator function I[A] equals 1 if the statement inside the bracket is true.

The conditional pdf in (30) implies the following conditional mean function for the outcome

$$\mathrm{E}[\mathrm{Y}_{\mathrm{X}^*} \mid \mathrm{X_o}, \mathrm{X_u}] = \left(1 - \mathcal{G}^{\mathrm{EM}}_{(\zeta^*_{\mathrm{EM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{EM}}, \mathrm{X}^*, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{EM}})\right) \times \frac{\int_{\zeta_{\mathrm{IM}}}^{\propto} \mathrm{Y}_{\mathrm{X}^*} g^{\mathrm{IM}}_{\left(\zeta^*_{\mathrm{IM}} \,\middle|\, \mathrm{Y}_{\mathrm{X}^*} > \zeta_{\mathrm{IM}}, \mathrm{X_o}, \mathrm{X_u}\right)}(\mathrm{Y}_{\mathrm{X}^*}, \mathrm{X}^*, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{IM}}) d\mathrm{Y}_{\mathrm{X}^*}}{1 - \mathcal{G}^{\mathrm{IM}}_{(\zeta^*_{\mathrm{IM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{IM}}, \mathrm{X}^*, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{IM}})} \tag{31}$$

Using the law of iterated expectation and (31), we can rewrite the EP in (1) as

$$\begin{aligned}\mathrm{AIE}(\Delta) = {} & \mathrm{E}\Bigg[\left(1 - \mathcal{G}^{\mathrm{EM}}_{(\zeta^*_{\mathrm{EM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{EM}}, \mathrm{X}^{\mathrm{pre}}+\Delta, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{EM}})\right) \\ & \times \frac{\int_{\zeta_{\mathrm{IM}}}^{\propto} \mathrm{Y}_{\mathrm{X}^*} g^{\mathrm{IM}}_{\left(\zeta^*_{\mathrm{IM}} \,\middle|\, \mathrm{Y}_{\mathrm{X}^*} > \zeta_{\mathrm{IM}}, \mathrm{X_o}, \mathrm{X_u}\right)}(\mathrm{Y}_{\mathrm{X}^*}, \mathrm{X}^{\mathrm{pre}}+\Delta, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{IM}}) d\mathrm{Y}_{\mathrm{X}^*}}{1 - \mathcal{G}^{\mathrm{IM}}_{(\zeta^*_{\mathrm{IM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{IM}}, \mathrm{X}^{\mathrm{pre}}+\Delta, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{IM}})}\Bigg] \\ & - \mathrm{E}\Bigg[\left(1 - \mathcal{G}^{\mathrm{EM}}_{(\zeta^*_{\mathrm{EM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{EM}}, \mathrm{X}^{\mathrm{pre}}, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{EM}})\right) \\ & \times \frac{\int_{\zeta_{\mathrm{IM}}}^{\propto} \mathrm{Y}_{\mathrm{X}^*} g^{\mathrm{IM}}_{\left(\zeta^*_{\mathrm{IM}} \,\middle|\, \mathrm{Y}_{\mathrm{X}^*} > \zeta_{\mathrm{IM}}, \mathrm{X_o}, \mathrm{X_u}\right)}(\mathrm{Y}_{\mathrm{X}^*}, \mathrm{X}^{\mathrm{pre}}, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{IM}}) d\mathrm{Y}_{\mathrm{X}^*}}{1 - \mathcal{G}^{\mathrm{IM}}_{(\zeta^*_{\mathrm{IM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{IM}}, \mathrm{X}^{\mathrm{pre}}, \mathrm{X_o}, \mathrm{X_u}; \tau_{\mathrm{IM}})}\Bigg]\end{aligned} \tag{32}$$

The AIE in (32) is not, however, identified because the underlying conditional pdf in (30) is written in terms of counterfactual entities that are not part of the DGP from which sample values are drawn. The AIE also contains a scalar control $\mathrm{X_u}$ that is part of the DGP yet is unobservable by the analyst. In the following section, I discuss identification of (31) which implies (32).

4.2 Identification of the Generic FP2PM with a Continuous Endogenous Variable

As discussed in chapter 2, identification of the AIE in (32) requires existence of a vector of control variables that induces CIND between X and Y_{X^*}. Suppose $V = [X_o \quad X_u]$ is such vector where X_o is represents the observable partition of V and X_u is a scalar comprising of the additional relevant unobservable controls needed to induce conditional independence. This implies that one can legitimately write the conditional pdf of Y given V and X by substituting X and Y for X^* and Y_{X^*} in (30) which yields

$$\begin{aligned}
\mathrm{pdf}(Y \mid X, X_o, X_u) = {} & \left[\mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X, X_o, X_u; \tau_{EM}) \right]^{I(Y=0)} \\
& \times \left[\left(1 - \mathcal{G}^{EM}_{(\zeta^*_{EM} \mid X_o, X_u)}(\zeta_{EM}, X, X_o, X_u; \tau_{EM})\right) \right. \\
& \left. \times \frac{g^{IM}_{\left(\zeta^*_{IM} \middle| Y_{X^*} > \zeta_{IM}, X_o, X_u\right)}(Y, X, X_o, X_u; \tau_{IM})}{1 - \mathcal{G}^{IM}_{(\zeta^*_{IM} \mid X_o, X_u)}(\zeta_{IM}, X, X_o, X_u; \tau_{IM})} \right]^{1-I(Y=0)} \qquad (33)
\end{aligned}$$

Although the conditional pdf in (33) is written virtually in terms of factual entities that are elements of the DGP, it contains a scalar X_u that is essential but unobservable. As a result, it is not possible to solely base identification and estimation of the AIE in (32) on the feature of the DGP shown in (33). Instead, one needs to impose additional structure that expands the DGP to include additional variables and relevant assumptions about the specifications of the conditional moments of the expanded DGP. Next, I present the proposed approach that adds structure to (33) to identification of its deep parameters and hence the AIE in (32). Note that the approach is specifically relevant to the 2PM for continuous nonnegative outcome with a continuous endogenous variable.

4.2.1 A Generalized Control Function Approach

A generalized control function (GCF) is specified based on a flexible distributional assumption for the endogenous variable. This approach is a specific case of an encompassing modeling and estimation framework proposed by Terza (2019c) for a very general class of nonlinear models involving an endogenous presumed causal variable. Under regular instrumental variables (IV) assumptions detailed later in this section, the GCF is a deep residual that necessarily induces CIND between X and Y_{X^*}. This property of the GCF is not satisfied by design in many other control function or predictor substitution approaches.[18]

As in Terza (2019c), the relationship between X and X_u implied by a fully specified parametric distribution for X can be written as

$$X = H^{-1}(X_u, W; \delta) \qquad (34)$$

where $H^{-1}(\,.\,)$ is the inverse of a known cdf of X given the vector $W = [X_o \quad W^+]$, W^+ is a vector of identifying instruments and δ is the parameter vector to be estimated.[19] The expression in (34) can be thought of as one resulting from the inverse transform theorem where X_u is a unit unform random variable (Ross, 1997). Thus, given that $H^{-1}(\,.\,)$ is continuous, we have

$$X_u = H(X, W; \delta) \qquad (35)$$

[18] Carlson (2020) demonstrates how the CIND assumption fails to hold in well-known control function approaches.

[19] $H^{-1}(\,.\,)$ could be written in long form as $H^{-1}_{(X|W)}(\,.\,)$.

where $X_u = U(0, 1)$, $U(a, b)$ denotes a random variable that is uniformly distributed on the interval (a, b) and H( ) is the conditional cdf of X given W. A vector of IV satisfying the regular IV assumptions completes identification of (33) and the relevant AIE in (32) which is a version of (1). These assumptions are reviewed below.

IV1. Exclusion Restriction

$$Y_{X^*} \perp W^+ | V \tag{36a}$$

where $\perp$ indicates statistical independence. IV1 states that Y_{X^*} and W^+ are independent conditional on V implying that W^+ is excluded from the conditional pdf of $Y_{X^*}|V$. In terms of the relevant DGP, (36a) amounts to[20]

$$f(Y \mid X, W, X_u; \pi) = f(Y \mid X, X_o, X_u; \pi)$$

where $f(. ; .)$ is a known conditional pdf and π is a parameter vector. For a minimally parametric specification of the 2PM, the relevant version of IV1 is

$$E[Y \mid X, W, X_u] = E[Y \mid X, X_o, X_u] = \mu(Y, X, X_o, X_u; \beta) \tag{36b}$$

where $\mu(. ; .)$ is a known conditional mean function and β is a sub-vector of π.

[20] The equality could be written in long form as $f_{(Y \mid X, W, X_u)}(. ; .) = f_{(Y \mid X, X_o, X_u)}(. ; .)$

IV2. Independence Between the Instrument and the Potential Value of the Endogenous Variable (Random assignment)

$$X_{W^+} \perp W^+ | X_o \tag{37a}$$

This assumption states that the identifying instrument W^+ and X_{W^+} are conditionally independent given X_o. X_{W^+} denotes a random variable representing the level of X that corresponds to an exogenously determined level of W^+. In the econometric literature on treatment effects, this IV assumption is called the random assignment to highlight the fact that units take on values of W^+ independent of unobservable characteristics that determine the observed version of $\mathbf{X}$ i.e., X. Alternatively, (37a) can be stated as $X_u \perp W^+|X_o$. IV1 and IV2 imply that W^+ affects Y only indirectly through X. For consistent estimation of the conditional mean, a weaker version of (37a) given below is sufficient.

$$E[X_u | W^+] = 0 \tag{37b}$$

IV3. Strength of the Instrument

Conditional on X_o, variation in W^+ should generate variation in X. In other words, $COV(X, W^+) \neq 0$ where $COV(a, b)$ denote the covariance between a and b. Given the specification for X in (34) and assumption IV2, it is clear that variation in W^+ generates an exogenous variation in X for a given value of X_o. Assumptions IV1- IV3 are sufficient to identify homogenous effect parameter i.e., where the effect of the X is constant for each unit in the population.

IV4. Monotonicity of the Instrument and the Unobservable in the Specification for X

Monotonicity assumption is especially important to identify an EP where the effect of the **X** is allowed to vary across units based on their X_u which can be thought to represent types of units. In such heterogenous effect models, except for the case where the causal variable of interest is binary and the instrument is discrete, monotonicity of X in the instrument W^+ and the unobservable scalar X_u are not equivalent (Imbens, 2007). Thus, I discuss each separately below.

IV4.1 Strict Monotonicity in the Instrument

If $H^{-1}(X_u = x_u, X_o, W^+ = w; \delta) > H^{-1}(X_u = x_u, X_o, W^+ = w'; \delta)$

for some ($X_u = x_u$, W^+=w, W^+=w'), then

$H^{-1}(X_u = x_u', X_o, W^+ = w; \delta) > H^{-1}(X_u = x_u', X_o, W^+ = w'; \delta)$ for all possible realizations of X_u. In other words, the instrument should move the value of X (also called the level of treatment) in the same direction for all types of units.

IV4.2 Strict Monotonicity in the Unobservable

If $H^{-1}(X_u = x_u, X_o, W^+ = w; \delta) > H^{-1}(X_u = x_u', X_o, W^+ = w; \delta)$

for some ($X_u = x_u$, $X_u = x_u'$, W^+=w), then

$H^{-1}(X_u = x_u, X_o, W^+ = w'; \delta) > H^{-1}(X_u = x_u', X_o, W^+ = w'; \delta)$ for all $W^+ = w'$. Alternatively stated, strict monotonicity in X_u indicates that as the value units take in (0, 1) increases, their corresponding level of X must move in the same direction for all possible realizations of W^+.

A sufficient condition for strict monotonicity of X in W^+ and X_u is that $H^{-1}(.)$ is either strictly increasing or strictly decreasing function in W^+ and X_u, respectively. By construction, the proposed GCF approach satisfies IV4.2 because once known fully parametric model is specified for X, by the inverse transform theorem, X_u, the cdf of X, is necessarily strictly monotonic. Therefore, if one can argue that assumption IV4.1 is satisfied, in addition to IV1-IV3, the GCF induces CIND between X and Y_{X^*} and hence identifies heterogenous effects.

THEOREM: Given the expression in (34) and under assumptions IV1-IV4 above, conditional on $V = [X_o \;\; X_u = H(X, W; \delta)]$, X and Y_{X^*} are independent.

Proof

By IV1-IV2 we have

$$Y_{X^*} \perp (W^+ | X_o, X_u)$$

Given (34) and IV4.2, (35) holds where $H(X, W; \delta)$ is one-to-one function of X_u.

$$\Rightarrow \quad Y_{X^*} \perp (W^+ | X_o, X_u = H(X, W; \delta))$$

IV2-IV3 (the variation in W^+ generates an exogenous variation in X that is independent of X_u) and (34) implies

$$Y_{X^*} \perp (H^{-1}(X_u, W; \delta) | X_o, X_u = H(X, W; \delta))$$

$$\Rightarrow \quad Y_{X^*} \perp (X | X_o, X_u = H(X, W; \delta)) \;\square$$

Theorem 1 establishes that the proposed approach satisfies the CIND needed to identify the conditional mean in (31) and the corresponding AIE in (32). Because X_u entails full information about the residual, I call the proposed approach as the generalized control function (GCF) approach.[21]

---

[21] The GCF identification argument here is closely related to those put forth for less parametric cases by Imbens and Newey (2002, 2009), Newey, Powell and Vella (1989)

Following the result from theorem 1, we can substitute H(X,W; δ) for X_u in (33) to obtain the relevant DGP as

$$\text{pdf}(Y|X, X_o, H(X,W;\delta)) = f(Y|X, W, H(X,W;\delta);\pi)$$

$$= \left[\mathcal{G}^{EM}_{(\zeta^*_{EM}|\, X_o,\, H(X,W;\delta))}(\zeta_{EM}, X, X_o, H(X,W;\delta); \tau_{EM}) \right]^{I(Y=0)}$$

$$\times \left[\left(1 - \mathcal{G}^{EM}_{(\zeta^*_{EM}|\, X_o,\, H(X,W;\delta))}(\zeta_{EM}, X, X_o, H(X,W;\delta); \tau_{EM})\right) \times \frac{g^{IM}_{\left(\zeta^*_{IM} \middle| Y_{X^*} > \zeta_{IM},\, X_o,\, H(X,W;\delta)\right)}(Y, X, X_o, H(X,W;\delta); \tau_{IM})}{1 - \mathcal{G}^{IM}_{(\zeta^*_{IM}|\, X_o,\, H(X,W;\delta))}(\zeta_{IM}, X, X_o, H(X,W;\delta); \tau_{IM})} \right]^{1-I(Y=0)}$$

(38)

The corresponding conditional mean is

$$E[Y \mid X,W, H(X,W;\delta)] = \left(1 - \mathcal{G}^{EM}_{(\zeta^*_{EM}|\, X_o,\, H(X,W;\delta))}(\zeta_{EM}, X, X_o, H(X,W;\delta); \tau_{EM})\right)$$

$$\times \frac{\int_{\zeta_{IM}}^{\infty} Y g^{IM}_{\left(\zeta^*_{IM} \middle| Y_{X^*} > \zeta_{IM},\, X_o,\, H(X,W;\delta)\right)}(Y, X, X_o, H(X,W;\delta); \tau_{IM})dY}{1 - \mathcal{G}^{IM}_{(\zeta^*_{IM}|\, X_o,\, H(X,W;\delta))}(\zeta_{IM}, X, X_o, H(X,W;\delta); \tau_{IM})}$$

(39)

Evaluating (39) at $X^{pre}+\Delta$ and X^{pre} and finding the difference and averaging over the entire X_o and H(X,W; δ) yields a version of the AIE in (32). The deep parameters of the model in (38) and the AIE based on (39) can be estimated by a two-stage GCF approach (more detail later). We first discuss in the next section a more efficient approach to estimating these deep parameters.

---

and Heckman and Robb (1986). It is also related to the fully parametric case in Terza (2009) where the **X** is assumed to be binary.

4.3 Full Information Maximum Likelihood Estimation of the Deep Parameters

A Full Information Maximum Likelihood (FIML) model for the 2PM is constructed based on the joint pdf of Y and X conditional on W and deep parameter vectors. The FIML model has the attractive that it implies a FIML estimation that provides asymptotically efficient estimates of the vector of deep parameters. We write a generic FIML model as

$$f_{(Y,\, X|X_u,\, W)}(Y, X, X_o, X_u; \tau, \delta) = \int_{-\infty}^{\infty} f_{(Y,\, X,\, X_u|W)}(Y, X, X_o, X_u; \tau) dX_u$$

since X_u is unit uniform, we have

$$= \int_0^1 f_{(Y,\, X,\, X_u|W)}(Y, X, W, X_u; \tau, \delta) dX_u$$

Further simplifying yields

$$= \int_0^1 f_{(Y|X,\, X_u,\, W)}(Y, X, W, X_u; \tau, \delta) \times g_{(X,\, X_u|W)}(X, W, X_u; \delta) dX_u$$

where $g_{(X,\, X_u|W)}(Y, X, W, X_u; \delta)$ is the joint pdf of X and X_u conditional on W and a vector of parameter δ. The GCF implies that X_u is unit uniform and the variation in X due to variation in W is independent of X_u. Thus,

$$g_{(X,\, X_u|W)}(X, W, X_u; \delta) = g_{(X|W)}(X, W, X_u; \delta) \times g_{(X_u)}(X_u)$$

where $g_{(X|\, W)}(X, W, X_u; \delta)$ is the pdf of X conditional on X_u, W and the parameter vector δ; $g_{(X_u)}(X_u)$ is the pdf of the unit-uniform random variable X_u which equals 1. Therefore, the joint pdf of Y and X conditional on W and deep parameter vectors is

$$f_{(Y,\, X|X_u,\, W)}(Y, X, X_o, X_u; \tau, \delta) = f_{(Y|X,\, X_u,\, W)}(Y, X, W, X_u; \tau, \delta) \times g_{(X|W)}(X, W, X_u; \delta)$$

Substituting $X_u = H(X,W;\delta)$ from (35), the joint pdf upon which the FIML model is based is given as

$$f_{(Y,\, X|X_u,\, W)}(Y, X, X_o, X_u; \tau, \delta) = f_{(Y|X,\, X_u,\, W)}(Y, X, W, H(X,W;\delta)\,;\tau)$$
$$\times g_{(X|W)}(X, W, H(X,W;\delta)) \tag{40}$$

The conditional pdf in (40) is a FIML model for the FP2PM with GCF that affords a FIML estimation. I call the MLE of the deep parameter vectors (τ, δ) based on the log-likelihood function of the pdf in (40) as the Generalized Control Function-Full Information Maximum Likelihood (GCF-FIML) estimator.

To obtain the specific GCF-FIML estimator of the deep parameter vectors, specific functional forms for the EM and IM need to be specified. In traditional 2PM framework, different functional forms are specified for the EM and the IM components. For instance, for modeling a continuous outcome in 2PM context, a logit EM and a gamma IM can be specified (Biener et al, 2020). However, as demonstrated in Hao and Terza (2018), such difference in structure is not needed. In a generic FP2PM, Hao and Terza (2018) analytically and via simulation show the robustness of maintaining NSD assumption for the EM and IM. In the context of the conditional pdf in (38) and a specific version of the joint conditional pdf in (40), this amounts to setting $\mathcal{G}^{EM}_{(\zeta^*_{EM}|\, X_o,\, H(X,W;\,\delta))}(\zeta_{EM}, X, X_o, H(X,W;\delta); \tau_{EM})$ to have a functional form that is the same as $\mathcal{G}^{IM}_{\left(\zeta^*_{IM}\middle|\, X_o,\, H(X,W;\,\delta)\right)}(Y,X, X_o, H(X,W;\delta); \tau_{IM})$. Specifying the FP2PM this way allows a formulation of a statistical test for the null hypothesis that "no 2PM is needed" in a particular empirical context. Once NSD is maintained the only source of the systematic difference between the process that generates zero values and that which determines the strictly positive values comes from a difference in the deep parameter vectors for the two components of the 2PM. This parametric difference refers to the possibility that the parameter vectors τ_{EM} and τ_{IM} in (38) are different. In the next subsection, a functional form is specified for the 2PM with NSD between the EM and the IM components. To appease any concern with misspecification of the fully parametric model, I assume a very flexible distribution for the EM and IM components of the 2PM as well as for the endogenous variable.

4.3.1 Specification of the Conditional Density with NSD

I specified a GG distribution for both the EM and IM components as well as for the endogenous variable. GG is parametrically very flexible distribution that subsumes several known distributions such as the Weibull, exponential, the standard gamma, and so on. The GG has been discussed and is being increasingly utilized in health economics and health service research methodology (Manning et al. 2005, Liu et al. 2010, Smith et al, 2015). The GG is particularly useful to fit continuous outcomes in empirical health economics (e.g health care cost) that are characterized by nonnegative values with long tail. Such a flexible distributional assumption appeases concern for misspecification of the functional form.[22]

[22] The generalized linear model (GLM) is extensively employed modeling framework to analyze the IM component in continuous 2PM (e.g Biener et al. 2020; Cawley et al. 2015; Chang and Mayerhoefer, 2016) Despite the fact that choosing an appropriate link and variance functions are the key for the performance of GLM estimators, no theoretically well-grounded procedure is available to guide these choices. Some tests such as the Park Test (Pregbon, 1980) and a modified Hosemer-Lemeshow test (Hosemer and Lemeshow, 1995) can be used to diagnose but not to fix misspecification of the link function. The

Accordingly, the specific version of the conditional joint pdf in (40) for the 2PM with a GG distributional assumption for both the outcome and the continuous endogenous variable is

$$
\begin{aligned}
& f_{(Y,\, X|W,\, H(X,W;\,\delta))}(Y, X, W, H(X,W;\delta);\pi) \\
&= \left[GG^{EM}_{(\zeta^{+}_{EM}\mid X, W, H(X,W;\,\delta))}(\zeta_{EM}, X\beta_{X1} + X_o\beta_{o1} + GG_{(X\mid W)}(X, W;\delta)\beta_{u1}, \sigma_{EM}, \kappa_{EM})\right]^{I(Y=0)} \\
&\times \Bigg[\left(1 - GG^{EM}_{(\zeta^{+}_{EM}\mid X, W, H(X,W;\,\delta))}(\zeta_{EM}, X\beta_{X1} + X_o\beta_{o1} + GG_{(X\mid W)}(X, W;\delta)\beta_{u1}, \sigma_{EM}, \kappa_{EM})\right) \\
&\quad \times \frac{gg^{IM}\left(\zeta^{+}_{IM} \,\middle|\, X, W, H(X,W;\delta)\right)^{(Y,\, X\beta_{X2} + X_o\beta_{o2} + GG_{(X\mid W)}(X, W;\,\delta)\beta_{u2};\, \sigma_{IM},\, \kappa_{IM})}}{1 - GG^{IM}_{(\zeta^{+}_{IM}\mid X, W, H(X,W;\,\delta))}(\zeta_{IM}, X\beta_{X2} + X_o\beta_{o2} + GG_{(X\mid W)}(X, W;\delta)\beta_{u2};\sigma_{IM}, \kappa_{IM})}\Bigg]^{1 - I(Y=0)} \\
&\times gg_{(X\mid W)}(X, W\delta;\sigma_\delta, \kappa_\delta)
\end{aligned}
\tag{41}
$$

where $\pi' = [\delta' \quad \beta_1' \quad \sigma_{EM} \quad \kappa_{EM} \quad \beta_2' \quad \sigma_{IM} \quad \kappa_{IM}]$ is the deep parameter vector such that $\delta' = [\delta_W \quad \sigma_\delta \quad \kappa_\delta]$; $\beta_1' = [\beta_{X1} \quad \beta_{o1}' \quad \beta_{u1}]$; $\beta_2' = [\beta_{X2} \quad \beta_{o2}' \quad \beta_{u2}]$; $\zeta^{+}_{\mathcal{X}M}$ is the component of the DGP that may be observed as a strictly positive outcome depending on whether or not it passes an unknown parametric threshold in the $\mathcal{X}$M component of the 2PM. $GG^{\mathcal{X}M}_{(\mathcal{A}|\mathcal{B})}(\mathcal{A}, \mathcal{b}, \mathcal{c}, \mathcal{d})$ and $gg^{\mathcal{X}M}_{(\mathcal{A}|\mathcal{B})}(\mathcal{A}, \mathcal{b}, \mathcal{c}, \mathcal{d})$ are the cdf and pdf, respectively, of a GG random variable $\mathcal{A}$ given $\mathcal{B}$ for the $\mathcal{X}$M component of the 2PM with location

modified Park Test suggested by Manning and Mullahy (2001) might be used to specify the variance function conditional on appropriately specified link function but this test relies on strong assumptions. On the other hand, the GG subsumes all the popular GLM specifications.

parameter b and shape parameters c and d. $GG_{(X \mid W)}(X, W; \delta)$ and $gg_{(X \mid W)}(X, W; \delta)$ are also the cdf and pdf, respectively, of a GG random variate X given W with parameter vector δ. The specific expression for the pdf and cdf of the GG are given in (21) and (23), respectively.

The shape parameters σ_{EM} and κ_{EM} are not, however, identified. Nevertheless, the following reduction is admissible (more detail later).[23]

$$GG^{EM}_{(\zeta^{+}_{EM}|X,\, W,\, G(X,W;\, \delta))}(\zeta_{EM}, X\beta_{X1} + X_o\beta_{o1} + GG(X,W;\delta)\beta_{u1}, \sigma_{EM}, \kappa_{EM})$$

$$= SG(\exp(X\beta^p_{X1} + X_o\beta^p_{o1} + GG(X,W;\delta)\beta^p_{u1}; \nu) \tag{42}$$

where SG (h, j) is the standard gamma variate evaluated at h with shape parameter j, $\beta^p_{X1} = -\,p_{EM}\beta_{X1}$, $\beta^p_{o1} = -\,p_{EM}\beta_{o1}$ and $\beta^p_{u1} = -\,p_{EM}\beta_{u1}$ with its constant term shifted by $\left[p_{EM}\ln(\zeta_{EM}) + \frac{1}{p_{EM}}\ln(\nu_{EM})\right]$. Combining (41) and (42), we obtain an approximation for the conditional joint pdf in (41) as

$$f_{(\tilde{Y},\, X|W,\, H(X,W;\, \delta))}(Y, X, W, GG(X,W;\delta); \pi)$$

$$= \left[SG(\exp(X\beta^p_{X1} + X_o\beta^p_{o1} + GG(X,W;\delta)\beta^p_{u1}; \nu)\right]^{I(Y=0)}$$

$$\times \left[\left(1 - SG(\exp(X\beta^p_{X1} + X_o\beta^p_{o1} + GG(X,W;\delta)\beta^p_{u1}; \nu\right)\right.$$

[23] See Hao and Terza (2018) for a simulation evidence on the consistency of an estimator for an AIE based on such an admissible reduction.

$$\times \frac{gg^{IM}_{\left(\zeta^{+}_{IM}\middle|X,\ W,\ H(X,W;\ \delta)\right)}(Y,\ X\beta_{X2}+X_o\beta_{o2}+GG(X,W;\ \delta)\beta_{u2};\ \sigma_{IM},\kappa_{IM})}{1-GG^{IM}_{(\zeta^{+}_{IM}|X,\ W,\ H(X,W;\ \delta))}(\zeta_{IM},\ X\beta_{X2}+X_o\beta_{o2}+GG(X,\ W;\ \delta)\beta_{u2};\ \sigma_{IM},\kappa_{IM})}\Bigg]^{1-I(Y=0)}$$

$$\times\ gg_{(X\ |\ W)}(X,\ W\delta\ ;\ \sigma_\delta,\ \kappa_\delta)$$

(43)

The vector of deep parameters π can be estimated by maximizing the following log-likelihood function based on (43).

$$L(\pi|Y,\ X,\ W)=\ \sum_{i=1}^{n}\ln\left(f_{\widetilde{(Y,\ X|W,\ H(X,W;\ \delta))}}(Y_i,\ X_i,\ W_i,\ GG(X_i,W_i;\ \delta);\ \pi)\right)$$

(44)

where, Y_i, X_i, and W_i are the values of Y, X, and W observed for the ith individual in the sample (i = 1, 2, …, n). ζ_{IM} is estimated by the minimum order statistics i.e., the smallest non-zero value in the sample. An estimator of the AIE in (32) can then be derived by plugging in the deep parameter estimates of the above log-likelihood function and the minimum order statistics ζ_{IM} for the corresponding deep parameter vectors and replacing the expectation operators by summation notation. Hence, the AIE estimator is

$$AIE(\Delta)\ =\sum_{i=1}^{n}\Big[\Big(1-SG(\exp((X_i^{pre}+\Delta_i)\hat{\beta}^{p}_{X1}+X_{oi}\hat{\beta}^{p}_{o1}+GG(X_i,W_i;\ \hat{\delta})\hat{\beta}^{p}_{u1};\ \nu\Big)$$

$$\times \frac{\int_{\hat{\zeta}_{IM}}^{Y_{max}} Y_i gg^{IM}(Y_i,\ (X_i^{pre}+\Delta_i)\hat{\beta}_{X2}+X_{oi}\hat{\beta}_{o2}+GG(X_i,W_i;\ \hat{\delta})\hat{\beta}_{u2};\ \hat{\sigma}_{IM},\ \hat{\kappa}_{IM})\ dY}{1-GG^{IM}(\hat{\zeta}_{IM},\ (X_i^{pre}+\Delta_i)\hat{\beta}_{X2}+X_{oi}\hat{\beta}_{o2}+GG(X_i,W_i;\ \hat{\delta})\hat{\beta}_{u2};\ \hat{\sigma}_{IM},\hat{\kappa}_{IM})}\Bigg]$$

$$-\sum_{i=1}^{n}\Big[\Big(1-SG(\exp(X_i^{pre}\hat{\beta}^{p}_{X1}+X_{oi}\hat{\beta}^{p}_{o1}+GG(X_i,W_i;\ \hat{\delta})\hat{\beta}^{p}_{u1};\ \nu\Big)$$

$$\times \frac{\int_{\hat{\zeta}_{IM}}^{Y_{max}} Ygg^{IM}(Y_i, X_i^{pre}\hat{\beta}_{X2}+X_{oi}\hat{\beta}_{o2}+GG(X_i,W_i;\hat{\delta})\hat{\beta}_{u2};\hat{\sigma}_{IM},\hat{\kappa}_{IM})\,dY}{1-GG^{IM}(\hat{\zeta}_{IM}, X_i^{pre}\hat{\beta}_{X2}+X_{oi}\hat{\beta}_{o2}+GG(X_i,W_i;\hat{\delta})\hat{\beta}_{u2};\hat{\sigma}_{IM},\hat{\kappa}_{IM})}\Bigg] \tag{45}$$

where $\hat{\pi}' = [\hat{\delta} \quad \hat{\beta}^{p}_{X1} \quad \hat{\beta}^{p'}_{o1} \quad \hat{\beta}^{p}_{u1} \quad \hat{\beta}_{X2} \quad \hat{\beta}^{p'}_{o2} \quad \hat{\beta}^{p}_{u2} \quad \hat{\sigma}_{IM} \quad \hat{\kappa}_{IM}]$ is a vector of the MLE estimates, $\hat{\zeta}_{IM}$ and Y_{max} are the minimum and maximum values of Y in the sample, respectively. For notational clarity, the subscripts in the gg and GG are suppressed. I call the AIE estimator in (45) the GCF-FIML based AIE estimator. The asymptotic standard errors of the MLE for the deep parameter in (45) can be obtained using the approach in Terza (2017).

4.3.2 Hypothesis Testing

The proposed approach affords easy-to-implement procedures for testing two crucial hypotheses in the context of the FP2PM framework where the $\mathbf{X}$ is a continuous endogenous variable. These are testing procedures for the null hypotheses that "no 2PM is needed" and "the causal variable of interest is exogenous". Below I discuss each of them.

4.3.2.1 Testing the "No 2PM is needed" Null Hypothesis

As mentioned earlier, the main features of the 2PM is that it allows the process governing the zero outcomes to systematically differ from that which determines strictly positive outcomes. Practically, the choice regarding the 2PM as the analytical framework in a given empirical context is guided by the presence of excess zeros and/or a theoretical ground. For example, in the health utilization research, analysts put forward two justifications for using the 2PM. First, a substantial proportion of individuals in a given sample have zero health care expenditure. Second, the decision to spend the first dollar is determined entirely by the patient while subsequent decisions are largely influenced by physicians (Pohlmeier and Ulrich, 1995). Testing statistically whether the 2PM is needed can, therefore, facilitate the choice of parsimonious model while providing insight into important theoretical predictions.

I developed a "no 2PM is needed" hypothesis testing procedure by utilizing the NSD assumption and by arguing that a certain reduction of unidentified ancillary parameters of the EM is admissible.[24] The NSD assumption implies that

$$\mathcal{G}^{EM}_{(\zeta^*_{EM} | X_o, X_u)}(\zeta_{EM}, X, X_o, H(X,W;\delta); \tau_{EM}) \text{ and } \mathcal{G}^{IM}_{\left(\zeta^*_{IM} \middle| X_o, X_u\right)}(Y, X, X_o, H(X,W;\delta); \tau_{IM})$$

have same functional forms. In our case, the underlying EM latent variable and the observed IM component for those units with strictly positive outcome are assumed to have a GG distribution. Next, I argue that the following reduction of the version of the model given in (41) is admissible.

$$f_{(Y, X|W, H(X,W;\delta))}(Y, X, W, H(X,W;\delta); \pi)$$

$$= \left[GG^{EM}_{(\zeta^+_{EM} | X, W, GG_{(X|W)}(X,W;\delta))}(\zeta, X\beta_{X1} + X_o\beta_{o1} + GG_{(X|W)}(X, W; \delta)\beta_{u1}; \sigma, \kappa)\right]^{I(Y=0)}$$

$$\times \left[\left(1 - GG^{EM}_{(\zeta^+_{EM} | X, W, GG_{(X|W)}(X,W;\delta))}(\zeta, X\beta_{X1} + X_o\beta_{o1} + GG_{(X|W)}(X, W; \delta)\beta_{u1}; \sigma, \kappa)\right)\right.$$

$$\left. \times \frac{gg^{IM}_{\left(\zeta^+_{IM} \middle| X, W, GG_{(X|W)}(X,W;\delta)\right)}(Y, X\beta_{X2} + X_o\beta_{o2} + GG_{(X|W)}(X, W; \delta)\beta_{u2}; \sigma, \kappa)}{1 - GG^{IM}_{(\zeta^+_{IM} | X, W, GG_{(X|W)}(X,W;\delta))}(\zeta, X\beta_{X2} + X_o\beta_{o2} + GG_{(X|W)}(X, W; \delta)\beta_{u2}; \sigma, \kappa)}\right]^{1 - I(Y=0)}$$

---

[24] See the discussion in Terza (1985) about admissible reduction.

$$\times\, gg_{(X \mid W)}(X, W\delta\,;\, \sigma_\delta, \kappa_\delta)$$

$$(46)$$

In other words, (46) implies that the following reduction of the model is admissible (i.e., it reduces the number of parameters to be estimated but does not distort the probability densities as assigned by the DGP).

$$\sigma_{EM} = \sigma_{IM} = \sigma$$

$$\kappa_{EM} = \kappa_{IM} = \kappa$$

$$\zeta^{+}_{EM} = \zeta^{+}_{IM} = \zeta^{+}$$

As the basis for this argument let

$$\alpha = \frac{\exp(X\beta_{X1} + X_o\beta_{o1} + GG_{(X \mid W)}(X, W;\, \delta)\beta_{u1})}{(\nu)^{\frac{1}{p}}}$$

where

$$\nu = \frac{1}{|\kappa_{EM}|^2}$$

and

$$p = \frac{\kappa_{EM}}{\sigma_{EM}}$$

Given that ζ^{+}_{EM} is a GG variate with parameters $X\beta_{X1} + X_o\beta_{o1} + GG_{(X \mid W)}(X, W;\, \delta)\beta_{u1}$, σ_{EM} and κ_{EM}, we have that

$$\left(\frac{\zeta_{EM}^{+}}{\alpha}\right)^{P}$$

is a standard gamma variate with shape parameter ν (Crooks, 2010). We seek a representation of the probability that

$$\zeta_{EM}^{+} \leq \zeta_{EM}$$

and, given that α is positive, this is equivalent to

$$\frac{\zeta_{EM}^{+}}{\alpha} \leq \frac{\zeta_{EM}}{\alpha}$$

Therefore, for p > 0

$$\left(\frac{\zeta_{EM}^{+}}{\alpha}\right)^{p} \leq \left(\frac{\zeta_{EM}}{\alpha}\right)^{p}$$

Now

$$\left(\frac{\zeta_{EM}}{\alpha}\right)^{p} = \exp\left(\ln\left[\left(\frac{\zeta_{EM}}{\alpha}\right)^{p}\right]\right)$$

$$= \exp\left(p \ln\left(\frac{\zeta_{EM}}{\alpha}\right)\right)$$

$$= \exp\left(p \ln\left(\zeta_{EM}\right) - p\left\{ \ln\left[\exp(X\beta_{X1} + X_o\beta_{o1} + GG(X, W; \delta)\beta_{u1})\right] - \frac{1}{p}\ln(\nu)\right\}\right)$$

$$= \exp\left(p \ln\left(\zeta_{EM}\right) - p\left(X\beta_{X1} + X_o\beta_{o1} + GG(X, W; \delta)\beta_{u1}\right) + \ln(\nu)\right)$$

$$= \exp\left(p \ln\left(\zeta_{EM}\right) + \ln(\nu) + X(-p\beta_{X1}) + X_o(-p\beta_{o1}) + GG(X, W; \delta)(-p\beta_{u1})\right)$$

$$= \exp\left(X\beta_{X1}^{o} + X_o\beta_{o1}^{o} + GG(X, W; \delta)\beta_{u1}^{o}\right)$$

where $\beta^o_{X1} = -p\beta_{X1}$, $\beta^o_{u1} = -p\beta_{u1}$ and β^o_{o1} is the same as $-p\beta_{o1}$ with its constant term shifted by $p\ \ln\left(\zeta_{EM}\right) + \ln(\nu)$. Therefore, for the probability that $\zeta^+_{EM} \leq \zeta_{EM}$ we have

$$\Pr(\zeta^+_{EM} \leq \zeta_{EM}) = GG(\zeta_{EM}; X\beta_{X1} + X_o\beta_{o1} + GG(X,W;\delta)\beta_{u1}, \sigma_{EM}, \kappa_{EM})$$

$$= SG\left(\exp(X\beta^o_{X1} + X_o\beta^o_{o1} + GG(X, W; \delta)\beta^o_{u1}\right); \nu). \tag{47}$$

This discussion clearly demonstrates the fact that the values of ζ_{EM} and σ_{EM} are absorbed by β^o_{X1}, β^o_{o1} and β^o_{u1} so that they are not identified. Their values are, in this sense, arbitrary. This, of course, means that

$$\zeta_{EM} = \zeta_{IM} = \zeta$$

$$\sigma_{EM} = \sigma_{IM} = \sigma$$

constitutes an admissible reduction. The argument for the admissibility of the reduction

$$\kappa_{EM} = \kappa_{IM} = \kappa$$

seems not as clear-cut. In essence, we seek to examine the extent to which κ_{EM} is absorbed by β^o_{X1}, β^o_{o1} and β^o_{u1}. In a simulation analysis, I demonstrated that fixing the value of κ_{EM}, and hence ν, at any arbitrary level does not affect the consistency of an AIE estimator, which is based on a likelihood function that is constructed from the conditional pdf in

(47).[25] Thus, the "no 2PM model is needed" null amounts to setting the β parameters of the EM and IM component equal. i.e.,

$$H_o: \beta_{i1} = \beta_{i2} = \beta$$

$$H_1: H_o \text{ is not true}$$

for i = 1,2,…K where K is the number of deep parameter coefficients. Thus, the relevant joint pdf conditional on X_o, and H(.) under the null becomes

$$f_{\widetilde{(Y,\, X|W,\, H(X,W;\,\delta))}}(Y, X, W, GG(X,W;\delta); \pi)$$

$$= \left[SG(\exp(p\ln(\zeta) + \ln(\nu) + X(-p\beta_X) + X_o(-p\beta_o) + GG(X, W; \delta)(-p\beta_u); \nu)\right]^{I(Y=0)}$$

$$\times \Bigg[\big(1 - SG(\exp(p\ln(\zeta) + \ln(\nu) + X(-p\beta_{X1}) + X_o(-p\beta_{o1}) + GG(X, W; \delta)(-p\beta_{u1}); \nu\big)$$

$$\times \frac{gg^{IM}_{\left(\zeta^{+}_{IM}\middle|X,\, W,\, H(X,W;\,\delta)\right)}(Y, X\beta_X + X_o\beta_o + GG(X,W;\delta)\beta_u; \sigma, \kappa)}{1 - GG^{IM}_{(\zeta^{+}_{IM}|X,\, W,\, H(X,W;\,\delta))}(\zeta, X\beta_X + X_o\beta_o + GG(X, W; \delta)\beta_u; \sigma, \kappa)}\Bigg]^{1 - I(Y=0)}$$

$$\times\, gg_{(X\,|\,W)}(X, W\delta\,; \sigma_\delta, \kappa_\delta)$$

(48)

This is a version of the joint pdf in (43) with the "no 2PM is needed" null imposed. The corresponding approximate log-likelihood function is

$$L_{H_o}(\lambda\,|\,Y, X, W) = \textstyle\sum_{i=1}^{n} \ln\,[f(Y_i, X_i, W_i, GG(X_i,W_i; \delta); \lambda)]$$

[25] See appendix II for a detail discussion of the simulation design and the results.

$L_{H_o}(.\,|\,.)$ denote the log-likelihood function to be maximized under the null. Given the unrestricted log-likelihood in (44), the likelihood ratio (LR) statistics

$$LR = -2\times[L(\pi|Y, X, W) - L_{H_o}(\lambda|Y, X, W)] \quad (49)$$

has a $\chi^2_{(K_\pi - K_\lambda)}$ where $\chi^2_{(K_a - K_b)}$ denotes the chi-square distribution with the degrees of freedom equal to the difference between the number of parameters in vectors a and b.

4.3.2.2 Testing for Exogeneity

In general, control function (CF) approaches address endogeneity by constructing a function that when conditioned on is supposedly induce CIND between the observed causal variable of interest and the potential outcomes. CF approaches naturally affords a simple procedure to test the null hypothesis that the causal variable of interest is exogenous. In particular, a t-test on the coefficient of the control function can be conducted to test if there is an evidence for endogeneity of the $\mathbf{X}$. The deep parameters in our case are, however, more than one. In particular, under the GG specification the null (H_o) and the alternative (H_1) hypotheses for testing exogeneity of the $\mathbf{X}$ are

H_o: $\beta_{1u}= 0$ *and* $\beta_{2u}=0$

H_1: H_o is not true

Similar to the test for "no 2PM is needed", the log-likelihood function under the null is a version of (48) with the restriction of the H_o. Because such log-likelihood is a nested version of the log-likelihood in (44), a likelihood ratio test statistic can be obtained by multiplying the difference in the log-likelihood values under H_1 and H_o by 2 which is distributed as $\chi^2_{(2)}$. Unlike two-step estimation protocols, because the proposed GCF-FIML is a one-step FIML estimator, it provides correct standard errors regardless of the decision about the above null hypothesis.

4.4 Simulation Study: Validating the Consistency of the GCF-FIML Based AIE Estimator and Comparing its Performance with Alternative Approaches

In this section, I demonstrate the implementation of the GCF approach, validate the consistency of the GCF-FIML based AIE estimator in (45) for a version of the AIE in (32) and compare its performance to four alternative estimators: namely, the two-stage residual inclusion (2SRI), the two-stage least square (2SLS), the two-stage predictor substitution (2SPS) and the two-stage Generalized Control Function (2SGCF) estimator. I first present the data generator for the true model in (38). Then the sampling design for generating a pseudo sample is presented. This will be followed by the analytical detail of the four alternative estimators. At the end of this section, the results of the simulation will be discussed.

4.4.1 The Data Generator

In order to conduct the simulation study, first, I develop Stata/Mata code for the true model in (41). The protocol for the simulator is as follows:

1) Choose the elements for the parameter vector

$\pi' = [\delta' \quad \sigma_\delta \quad \kappa_\delta \quad \beta_1' \quad \sigma_{EM} \quad \kappa_{EM} \quad \beta_2' \quad \sigma_{IM} \quad \kappa_{IM}]$ and the unobserved parametric thresholds ζ_{EM} and ζ_{IM}.

2) Generate a sample of simulated data for X_o and W^+; each assumed to be uniformly distributed with mean and variances chosen as part of the sampling design.

3) Generate a sample of simulated data for X from a GG distribution as

$$X = GG^{-1}_{(X \mid W)}(U[0, 1], X_o\delta'_o + W^+\delta'_{W^+}, \sigma_X, \kappa_X)$$

$$= a_X \times SG^{-1}(\nu_X, U[0, 1])^{\frac{1}{p_X}} \tag{50}$$

with a specified parameter vector $\delta' = [\delta'_W \ \ \sigma_X \ \ \kappa_X]$ where $\delta'_W = [\delta_{W^+} \ \ \delta_{X_o} \ \ \delta_{con}]$, the coefficient vector for the linear index δW that represents the location parameter, and σ_X and κ_X denote the shape parameters of the distribution. δ_{con} is an intercept term of the linear index δW. As specified in (24), $a_X = \frac{\exp(X_o\delta'_o + W^+\delta'_{W^+})}{\left(\frac{1}{|\kappa_X|^2}\right)^{\frac{1}{p_X}}}$, $\nu_X = \frac{1}{|\kappa_X|^2}$ and $p_X = \frac{\kappa_X}{\sigma_X}$.

4) Recover the values for X_u by calculating the cdf of X i.e.,

$$X_u = GG_{(X \mid W)}(X; X_o\delta'_o + W^+\delta'_{W^+}, \sigma_X, \kappa_X) \tag{51}$$

5) Generate a sample of outcomes at the EM (Y = 0 or not) i.e.,

EM = 1 iff

$$GG^{EM}_{(\zeta^+_{EM} \mid X, W, H(X,W;\delta))}(\zeta_{EM}, X\beta_{X1} + X_o\beta_{o1} + GG_{(X \mid W)}(X, W; \delta)\beta_{u1}, \sigma_{EM}, \kappa_{EM}) > \mathcal{U} \tag{52}$$

where $\mathcal{U}$ is a unit uniform random variable.

6) Complete the construction of the sample by simulating the IM values for the subpopulation of units for whom the $\text{EM} = 1$. These values need to be drawn from appropriately specified truncated distribution. Below I extend the derivation of the data generator for the truncated GG distribution discussed in Hao and Terza (2018) for the case where the $\mathbf{X}$ is endogenous.

A version of the cdf of a GG variate in (23) can be written in an important way as

$$\mathrm{GG}(\zeta, \mathcal{X}\beta, \sigma, \kappa) = \begin{cases} \mathrm{SG}(\nu, (\zeta/a)^{p}) & \text{if} \quad p > 0 \\ 1 - \mathrm{SG}(\nu, (\zeta/a)^{p}) & \text{if} \quad p < 0 \end{cases} \tag{53}$$

where $\mathcal{X}\beta$ is a linear index, $a = \frac{\exp(\mathcal{X}\beta)}{\left(\frac{1}{|\kappa|^2}\right)^{\frac{1}{p}}}$, $\nu_{IM} = \frac{1}{|\kappa|^2}$, $p = \frac{\kappa}{\sigma}$ and ζ is unobserved parametric threshold. (52) implies that for $p > 0$,

$$\mathrm{GG}^{-1}(\mathcal{P}, \mathcal{X}\beta, \sigma, \kappa) = a \times \left[\mathrm{SG}^{-1}(\mathcal{P}; \nu, 1)\right]^{\frac{1}{p}}$$

where $\mathcal{P}$ is a value in the unit interval and $\mathrm{GG}^{-1}(\mathcal{P}; c, d, e)$ is the inverse cdf of a GG variate with parameters c, d and e evaluated at $\mathcal{P}$. For a truncated GG random variable, a version of the cdf in (29) is given in a shorthand notation as

$$\mathrm{GG}^{*}(Y; X\beta_{X2} + X_o\beta_{o2} + \mathrm{GG}(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})$$

$$= \frac{\mathrm{GG}_{(Y \geq \zeta_{IM})}(Y; X\beta_{X2} + X_o\beta_{o2} + \mathrm{GG}(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})}{1 - \mathrm{GG}(\zeta_{IM}; X\beta_{X2} + X_o\beta_{o2} + \mathrm{GG}(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})} \tag{54}$$

where GG^{*} (Y; c, d, e, ζ) is the cdf of a truncated GG for Y with parameters c, d and e truncated at ζ. Now (54) implies that I can generate a truncated GG random variable y based on

$$\left[1 - GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM},\kappa_{IM})\right]\times U(0,1)$$

$$= GG_{(Y \geq \zeta_{IM})}(Y; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM}) \qquad (55)$$

We want to use GG^{-1} (Y; c, d, e) not $GG^{-1}_{(Y \geq \zeta_{IM})}$(Y; c, d, e) to generate the desired random variable. To achieve this, adding to both sides of (55) the following term $GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})$ gives

$$\left[1 - GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM},\kappa_{IM})\right]\times U(0,1) +$$
$$GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})$$
$$= GG(y; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM}) \qquad (56)$$

where the right side is the sum of $GG_{(Y \geq \zeta_{IM})}(Y; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})$ and $GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})$. From (56), it follows that

$$y = GG^{-1}(\mathcal{A}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM}) \qquad (57)$$

where

$$\mathcal{A} = \left[1 - GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})\right]\times U(0,1) +$$
$$GG(\zeta_{IM}; X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}, \sigma_{IM}, \kappa_{IM})$$

To complete the data generation,

5a) Generate U(0,1)

5b) Obtain the linear index $X\beta_{X2}+X_o\beta_{o2}+ GG(X,W;\delta)\beta_{u2}$ based on 1) to 4) above.

5c) Plug the values from 5a, 5b and the parameter values σ_{IM} and κ_{IM} into (56) to get the desired pseudo random variable y.

4.4.2 The Sampling Design

To validate the consistency of the GCF-FIML based AIE estimator for a version of the AIE in (32) and to compare its performance with the alternative estimators such as the 2SRI, 2SLS, 2SPS and 2SGCF, the following sample design is considered.

1) To generate pseudo values for X, I set

$$\delta_W' = [\delta_{W^+} \quad \delta_{X_o} \quad \delta_{con}] = [0.75 \quad -0.5 \quad -1]$$

for the linear index coefficients. The means and variances of X_o and W^+ are set to be $E[X_o] = 1$, $E[W^+] = 1$, $Var[X_o] = 0.45$ and $Var[W^+] = 1$. I also set values for the shape parameters as $\sigma_X = 0.51$ and $\kappa_X = 0.25$.

2) For generating the values at the EM, I set the values for the linear index coefficients, the shape parameters and the parametric threshold for (52) as follows

$$[\beta_{X1} \quad \beta_{u1} \quad \beta_{o1}'] = [0.4 \quad 0.8 \quad -0.5 \quad 0.25]$$

where $\beta_{o1}' = [\beta_{X_{o1}} \quad \beta_{cons1}]$ are the coefficients for X_o and the intercept, respectively.

$$\sigma_{EM} = 0.5$$

$\kappa_{EM} = 1$

$\zeta_{EM} = 0.75$

Note that σ_{EM}, κ_{EM} and ζ_{EM} are not identified in 2PM.

3) Similarly, to generate the pseudo values, the following parameter values for (57) are set

$$[\beta_{X2} \quad \beta_{u2} \quad \beta_{o2}'] = [0.5 \quad 0.25 \quad -0.5 \quad 0.5]$$

where $\beta_{o2}' = [\beta_{X_{o2}} \quad \beta_{cons2}]$ are the coefficients for X_o and the intercept, respectively.

$\sigma_{IM} = 1.5$

$\kappa_{IM} = 1.5$

$\zeta_{IM} = 2$

4) For testing the consistency of the AIE based on the proposed approach, samples of increasing size are generated based on the above sampling design. In particular, the samples are generated with the following sizes.

n = 1,000

n = 5,000

n = 15,000

n = 25,000

n = 50,000

n = 100,000

n = 250,000

n = 500,000

4.4.3 Alternative Approaches to Correcting Endogeneity in the 2PM

In sections 4.2 and 4.3 I presented detailed discussion on the proposed GCF identification approach and the implied GCF-FIML estimator for correcting endogeneity in a FP2PM framework from the potential outcomes perspective. In this sub-section, I layout four alternative approaches for correcting endogeneity mentioned at the beginning of section 4.4. These approaches except the 2SGCF are employed in empirical research in the context of the 2PM with continuous outcome and continuous endogenous variable. Like the GCF-FIML approach, all of these approaches use an IV to identify CI parameters. To facilitate comparison with the proposed approach, these alternative approaches are cast within the GPOF. Thus, I commence the discussion on this sub-section by assuming that the conditional pdf in (30) holds.

4.4.3.1 The Two-Stage Residual Inclusion Approach[26]

This approach requires only specification of the conditional mean function for the outcome and endogenous variables. To facilitate comparison with the proposed approach, I instead focus on a 2SRI approach based on a FP2PM. Such approach helps provide a correct specification for the conditional mean and also improves the efficiency of estimated parameters. Terza, Basu and Rathouz (2008) suggested the following auxiliary regression for the endogenous variable.

[26] The 2SRI estimation approach, popularized by Terza, Basu and Rathouz [TBR] (2008), is a control function approach for estimating causal effects in a general additive non-linear triangular model in which nonlinear models are specified both for the outcome and endogenous variables such that the former does not causally affect the later.

$$X = r(W; \delta) + X_u \tag{58}$$

where r(.) is a known conditional mean function. Extending the assumption of a GG distribution for the X to this context, we have

$$r(W; \delta) = \exp\left\{W\xi + (\sigma/\kappa) \times \ln(\kappa^2) + \ln[\Gamma(\,{}^{1}/_{\kappa^2}\,) + (\sigma/\kappa)] - \ln[\Gamma(\,{}^{1}/_{\kappa^2}\,)]\right\} \tag{59}$$

where the right-hand side is an expression for the conditional mean of a GG variate X. Although $\delta' = [\xi' \quad \sigma \quad \kappa]$ can be consistently estimated by using non-linear least square (NLS), by taking advantage of a known distributional assumption for X, a more efficient estimate for the parameter vector δ can be obtained via MLE. Invoking the standard IV conditions IV2-IV3 (in addition to the IV1 that ensures exclusion of W^+ from the specified model for the outcome), TBR argue that the following raw residual can serve as a control function that induces CIND between X and Y_{X^*}.

$$\widehat{X}_u = X - r\left(W; \hat{\delta}\right) \tag{60}$$

By plugging (60) for X_u into a version of (41) where the standard gamma admissible reduction replaces the GG in the EM, the relevant conditional pdf is given as

$$\mathrm{pdf}(Y \mid X, X_o, X_u) = \left[SG(\exp(X\beta_{X1}^{2SRI_p} + X_o\beta_{o1}^{2SRI_p} + \widehat{X}_u\beta_{u1}^{2SRI_p}; 1)\right]^{I(Y=0)}$$

$$\times \left[\left(1 - SG(\exp(X\beta_{X1}^{2SRI_p} + X_o\beta_{o1}^{2SRI_p} + \widehat{X}_u\beta_{u1}^{2SRI_p}; 1)\right)\right.$$

$$\times \frac{gg^{IM}_{\left(\zeta^*_{IM} \middle| Y_{X^*} > \zeta_{IM}, X_o, X_u\right)}(Y;\, X\beta^{2SRI}_{X2} + X_o\beta^{2SRI}_{o2} + \hat{X}_u\beta^{2SRI}_{u2},\, \sigma^{2SRI}_{IM},\, \kappa^{2SRI}_{IM})}{1 - GG^{IM}_{(\zeta^*_{IM} | X_o, X_u)}(\zeta_{IM};\, X\beta^{2SRI}_{X2} + X_o\beta^{2SRI}_{o2} + \hat{X}_u\beta^{2SRI}_{u2},\, \sigma^{2SRI}_{IM},\, \kappa^{2SRI}_{IM})}\Bigg]^{1-I(Y=0)} \tag{61}$$

The deep parameter vectors $\tau^{2SRI\prime}_{EM} = [\beta^{2SRI_p}_{X1} \quad \beta^{2SRI_p{}'}_{o1} \quad \beta^{2SRI_p}_{u1}]$ for the EM and $\tau^{2SRI\prime}_{IM} = [\beta^{2SRI}_{X2} \quad \beta^{2SRI'}_{o2} \quad \beta^{2SRI}_{u2} \quad \sigma^{2SRI}_{IM}]$ for the IM can then be estimated by maximizing the log-likelihood function based on (61).[27] The corresponding AIE estimator whose consistency for a version of the AIE in (32) will be evaluated is [28]

$$\mathrm{AIE}(\Delta) = \sum_{i=1}^{n} \Bigg[\left(1 - SG(\exp((X^{pre}_i + \Delta_i)\hat{\beta}^{2SRI_p}_{X1} + X_{oi}\hat{\beta}^{2SRI_p}_{o1} + \hat{X}_{ui}\hat{\beta}^{2SRI_p}_{u1};1\right)$$

$$\times \frac{\int_{\hat{\zeta}_{IM}}^{Y_{max}} Y_i gg^{IM}(Y_i, (X^{pre}_i + \Delta_i)\hat{\beta}^{2SRI}_{X2} + X_{oi}\hat{\beta}^{2SRI}_{o2} + \hat{X}_{ui}\hat{\beta}^{2SRI}_{u2};\, \hat{\sigma}^{2SRI}_{IM},\, \hat{\kappa}^{2SRI}_{IM})\, dY}{1 - GG^{IM}(\hat{\zeta}_{IM}, (X^{pre}_i + \Delta_i)\hat{\beta}^{2SRI}_{X2} + X_{oi}\hat{\beta}^{2SRI}_{o2} + \hat{X}_{ui}\hat{\beta}^{2SRI}_{u2};\, \hat{\sigma}^{2SRI}_{IM},\, \hat{\kappa}^{2SRI}_{IM})}\Bigg]$$

$$- \sum_{i=1}^{n} \Bigg[\left(1 - SG(\exp(X^{pre}_i\hat{\beta}^{2SRI_p}_{X1} + X_{oi}\hat{\beta}^{2SRI_p}_{o1} + \hat{X}_{ui}\hat{\beta}^{2SRI_p}_{u1};1\right)$$

$$\times \frac{\int_{\hat{\zeta}_{IM}}^{Y_{max}} Y gg^{IM}(Y_i, X^{pre}_i\hat{\beta}^{2SRI}_{X2} + X_{oi}\hat{\beta}^{2SRI}_{o2} + \hat{X}_{ui}\hat{\beta}^{2SRI}_{u2};\, \hat{\sigma}^{2SRI}_{IM},\, \hat{\kappa}^{2SRI}_{IM})\, dY}{1 - GG^{IM}(\hat{\zeta}_{IM}, X^{pre}_i\hat{\beta}^{2SRI}_{X2} + X_{oi}\hat{\beta}^{2SRI}_{o2} + \hat{X}_{ui}\hat{\beta}^{2SRI}_{u2};\, \hat{\sigma}^{2SRI}_{IM},\, \hat{\kappa}^{2SRI}_{IM})}\Bigg] \tag{62}$$

[27] The 2SRI model does not require a fully parametric specification for the outcome. In fact, by plugging (53) into a conditional mean function for the outcome, TBR show that the 2SRI is the best predictor in the sense of minimizing mean square prediction error.

[28] Corresponding to the result shown in the admissible reduction in (47), $\beta^{2SRI_p}_{X1} = -p^{2SRI}_{EM}\beta^{2SRI}_{X1}$, $\beta^{2SRI_p}_{o1} = -p^{2SRI}_{EM}\beta^{2SRI}_{o1}$ and $\beta^{2SRI_p}_{u1} = -p^{2SRI}_{EM}\beta^{2SRI}_{u1}$ with its constant term shifted by $\left[p^{2SRI}_{EM}\ln(\zeta^{2SRI}_{EM}) + \frac{1}{p_{EM}}\ln(\nu^{2SRI}_{EM})\right]$.

The consistency of the above 2SRI estimators for the AIE depends on whether or not the specified raw residual based on the auxiliary regression function for X_u in (60) is correct.

4.4.3.2 The Two-Stage Least Squares Approach

The 2SLS approach is one of the most widely used approach to estimate causal effects when the causal variable of interest is endogenous. Its applicability for 2PM is, however, limited because it ignores the nonlinearity in the EM and IM components of the 2PM that is often inherent in many empirical settings. Some applied research, however, still use it mainly for its simplicity in estimation and interpretation. Under the 2SLS, the auxiliary regression model for X is specified as

$$X = \delta W + X_u \tag{63}$$

which is typically estimated using the OLS approach. By IV3 and the minimally parametric version of IV2, the above model generates a variation in X that is independent of Y_{X^*}, making X_u redundant in the conditional mean function for the outcome. This exogenous variation in X is obtained from the predicted values based on the OLS estimates of δ in (63). These predicted value are given as

$$\widehat{X} = \widehat{\delta} W \tag{64}$$

On the other hand, the minimally parametric version of IV1 ensures that W^+ is excluded from the linear conditional mean function specified under 2SLS. This implies that the conditional mean is given by

$$E[Y \mid X, X_o, X_u] = E[Y \mid X, W]$$

$$= [\alpha_{X1}\widehat{X} + \alpha_{o1}X_o] \times [\alpha_{X2}\widehat{X} + \alpha_{o2}X_o] \qquad (65)$$

where $P(Y > 0|X, W) = [\alpha_{X1}(\hat{\delta}W) + \alpha_{o1}X_o]$ and $E[Y \mid Y > 0, X, W] = [\alpha_{X2}(\hat{\delta}W) + \alpha_{o2}X_o]$ are the conditional mean functions for the EM and IM components of the 2PM, respectively. The 2SLS AIE estimator for a version of (32) is, thus,

$$AIE(\Delta) = \sum_{i=1}^{n}\left[[(\widehat{X}_i^{pre} + \Delta_i)\hat{\alpha}_{X1} + \hat{\alpha}_{o1}X_{oi}] \times [(\widehat{X}_i^{pre} + \Delta_i)\hat{\alpha}_{X2} + \hat{\alpha}_{o2}X_o]\right]$$

$$-\sum_{i=1}^{n}\left[[\widehat{X}_i^{pre}\hat{\alpha}_{X1} + \hat{\alpha}_{o1}X_{oi}] \times [\widehat{X}_i^{pre}\hat{\alpha}_{X2} + \hat{\alpha}_{o2}X_o]\right] \qquad (66)$$

When the object of interest is causal estimation, Angrist (2001) argues that linear probability model for the EM component can often provide a good approximation to the conditional probability whether the $\mathbf{X}$ is binary, count or continuous. Black et al (2018), for instance, specified a linear IV model for the EM component of the 2PM to estimate health care cost of childhood obesity and find identical estimate as logit estimates. They also argue in favor of using 2SLS for the IM component of the 2PM because log health cost is approximately normally distributed. Linear approximation to the IM component is, however, problematic due to skewness and heavy tail that typically characterize many of the semi-continuous outcomes cast in a 2PM context. Although transformation of the continuous component of the IM mitigates these data problems, the issues involved in the retransformation of outcomes to the original scale makes 2SLS unattractive.

4.4.3.3 The Two-Stage Predictor Substitution Approach

The 2SPS is a rote extension of the 2SLS approach. The argument for the first stage regression is the same as the 2SLS approach where the auxiliary regression for X is specified as in (63) and the corresponding predicted values are obtained as in (64). These predicted values are then substituted in the second stage in a specified nonlinear conditional mean function. Like the 2SRI approach, I specify a FP2PM to evaluate the consistency of an AIE estimator for a version of (32) based on the 2SPS approach. The protocol for estimating an AIE under 2SPS approach is as follows:

First stage

Obtain the predicted values from a first stage linear regression of X on the vector of controls and instruments W like the one in (64).

Second Stage

Note that like the 2SLS, the 2SPS approach assumes that under the minimally parametric version of IV1-IV2 and IV3, the X_u is redundant. Therefore, the relevant conditional pdf can be obtained by substituting (64) for X in a version of (33) where X_u is excluded. Assuming a GG for both the EM and IM components of the 2PM and using the standard gamma admissible reduction for the EM component, we have the relevant conditional pdf upon which the second stage estimation is based as

$$\mathrm{pdf}(Y \mid X, X_o, X_u) = \left[\mathrm{SG}(\exp(\widehat{X}\beta_{X1}^{2SPS_p} + X_o\beta_{o1}^{2SPS_p}; 1)\right]^{I(Y=0)}$$

$$\times \left[\left(1 - \mathrm{SG}(\exp(\widehat{X}\beta_{X1}^{2SPS_p} + X_o\beta_{o1}^{2SPS_p}; 1)\right)\right.$$

$$\times \frac{gg^{IM}_{\left(\zeta^*_{IM} \middle| Y_{X^*} > \zeta_{IM}, X_o, X_u\right)}(Y; \hat{X}\beta^{2SPS}_{X2} + X_o\beta^{2SPS}_{o2} + \hat{X}_u\beta^{2SPS}_{u2}, \sigma^{2SPS}_{IM}, \kappa^{2SPS}_{IM})}{1 - GG^{IM}_{(\zeta^*_{IM}| X_o, X_u)}(\zeta_{IM}; \hat{X}\beta^{2SPS}_{X2} + X_o\beta^{2SPS}_{o2} + \hat{X}_u\beta^{2SPS}_{u2}, \sigma^{2SPS}_{IM}, \kappa^{2SPS}_{IM})}\Bigg]^{1-I(Y=0)}$$

(67)

The deep parameters of the above model can be estimated by maximizing a log-likelihood function based on (67) above. The corresponding 2SPS AIE estimator for a version of (32) is [29]

$$\mathrm{AIE}(\Delta) = \sum_{i=1}^{n} \Bigg[\Big(1 - SG(\exp((\hat{X}^{pre}_i + \Delta_i)\hat{\beta}^{2SPS_p}_{X1} + X_{oi}\hat{\beta}^{2SPS_p}_{o1}; 1\Big)$$

$$\times \frac{\int_{\hat{\zeta}_{IM}}^{Y_{max}} Y_i gg^{IM}(Y_i, (\hat{X}^{pre}_i + \Delta_i)\hat{\beta}^{2SPS}_{X2} + X_{oi}\hat{\beta}^{2SPS}_{o2}; \hat{\sigma}^{2SPS}_{IM}, \hat{\kappa}^{2SPS}_{IM})\, dY}{1 - GG^{IM}(\hat{\zeta}_{IM}, (\hat{X}^{pre}_i + \Delta_i)\hat{\beta}^{2SPS}_{X2} + X_{oi}\hat{\beta}^{2SPS}_{o2}; \hat{\sigma}^{2SPS}_{IM}, \hat{\kappa}^{2SPS}_{IM})}\Bigg]$$

$$-\sum_{i=1}^{n} \Bigg[\Big(1 - SG(\exp(\hat{X}^{pre}_i\hat{\beta}^{2SPS_p}_{X1} + X_{oi}\hat{\beta}^{2SPS_p}_{o1}; 1\Big)$$

$$\times \frac{\int_{\hat{\zeta}_{IM}}^{Y_{max}} Y gg^{IM}(Y_i, \hat{X}^{pre}_i\hat{\beta}^{2SPS}_{X2} + X_{oi}\hat{\beta}^{2SPS}_{o2}; \hat{\sigma}^{2SPS}_{IM}, \hat{\kappa}^{2SPS}_{IM})\, dY}{1 - GG^{IM}(\hat{\zeta}_{IM}, \hat{X}^{pre}_i\hat{\beta}^{2SPS}_{X2} + X_{oi}\hat{\beta}^{2SPS}_{o2}; \hat{\sigma}^{2SPS}_{IM}, \hat{\kappa}^{2SPS}_{IM})}\Bigg] \quad (68)$$

[29] Corresponding to the result shown in the admissible reduction in (47), $\beta^{2SPS_p}_{X1} = -p^{2SPS}_{EM}\beta^{2SPS}_{X1}$, $\beta^{2SPS_p}_{o1} = -p^{2SPS}_{EM}\beta^{2SPS}_{o1}$ and $\beta^{2SPS_p}_{u1} = -p^{2SPS}_{EM}\beta^{2SPS}_{u1}$ with its constant term shifted by $\left[p^{2SPS}_{EM}\ln\left(\zeta^{2SPS}_{EM}\right) + \frac{1}{p_{EM}}\ln(\nu^{2SPS}_{EM})\right]$.

4.4.3.4 The Two-Stage Generalized Control Function Estimator

The 2GCF is a two-stage estimator of the conditional pdf given in (38). It is an estimator of the deep parameter vector in (38) that is computationally less demanding yet less efficient than the GCF-FIML estimator, a one-step MLE estimator based on the joint pdf in (43). The procedure for estimating the deep parameter vector in (38) under the 2SGCF approach is as follows:

First stage

Estimate the deep parameters of the fully parametric model for the endogenous variable and obtain its cdf based on the estimated parameters. In the specific case we consider here, the first stage involves estimating δ, the GG parameter vector for X, and obtain the cdf of X as $\widehat{X}_{ui} = GG(X_i, W_i; \hat{\delta})$.

Second Stage

Substitute as $\widehat{X}_{ui} = GG(X_i, W_i; \hat{\delta})$ into the conditional pdf in (38) and estimate its deep parameters by maximizing the implied log-likelihood function. Under the conditions outlined in theorem 1, the MLE obtained in this way is necessarily consistent. Although the 2SGCF is computationally less burdensome, it leads to unnecessary efficiency loss as it ignores relevant information entailed in the joint pdf of Y and X. The 2SGCF approach is also amenable to conduct the two hypotheses discussed in sections 4.3.2.1 and 4.3.2.2.

4.4.4 Simulation Results

As discussed above the true model outlined in section 4.4.1 is based on (38). I generated a super sample of n =1,500,000 to compute the true AIE. The proposed GCF-FIML, 2SGCF and the three alternative AIE estimators are applied on each of the samples generated. The absolute percentage bias (APB) for each AIE estimate is then computed using the formula in (26).

Table 6 presents the result of the simulation. The true AIE based on the sampling design is 2.1449. The AIE estimates from six different estimators are presented in successive columns. These estimators are the GCF-FIML, 2SGCF, 2SRI, 2SPS, 2SLS and the Full Information Maximum Likelihood Estimator with exogenous causal variable (FIML-EXOG). The last estimator is based on a FIML model that ignores the endogeneity of the causal variable. The APB of the GCF-FIML and 2SGCF based AIE estimates are almost the same for most of the sample sizes. The estimated AIEs from these two estimators have small APB that gets close to 0% as the sample size of the simulated data increases. On the other hand, the 2SLS based AIE estimates have an APB of around 30% even for very large sample sizes. The 2SPS AIE estimates are also biased with an APB of around 25% for very large samples.[30] This APB is comparable to the APB of the FIML estimator that ignores the endogeneity of the $\mathbf{X}$. While the 2SRI based AIE estimator performs better than the 2SPS, 2SLS and FIML-EXOG based AIE estimators, the APB for its estimates hoovers around 10% even for very large sample sizes.

To examine how sensitive the estimates are to the amount of endogeneity and nonlinearity, I design a different sample with lower parameter values for β_{u1}, β_{u2} and β_{X2}. [31] The resulting model based on this sample has much lower endogeneity and nonlinearity. As shown in table 7, the new true AIE computed based on a super sample of n =1,500,000

[30] Convergence was not achieved for sample sizes of $n = 25{,}000$, $n = 50{,}000$ and $n = 100{,}000$.

[31] See the entire sampling design in appendix IV.

is 2.009. The table also presents the AIE estimates based on the six estimators outlined above using new samples of increasing size. Again, the GCF-FIML and 2SGCF based AIE estimators perform very well in terms of consistency as the APB of the estimates gets close to zero for large sample sizes. The 2SRI and 2SLS based AIE estimators perform well. The 2SPS based AIE estimator, however, still has a substantial APB even for very large samples. In fact, in this particular sampling design, the 2SPS AIE estimator is even worse than the FIML-EXOG based AIE estimator that ignores endogeneity. The above results imply that the proposed GCF-FIML and the 2SGCF based AIE estimators are consistent for a version of the AIE in (32) and outperforms the alternative three estimators that are commonly used in the context of the 2PM.

4.5 Application: The Medical Care Cost of Obesity in Youth in the US

Obesity rate among children in the US has more than quadrupled from 5% in 1971-1974 to 20.5% in 2015-2016. (Anderson and Butcher, 2006; Ogden et al., 2012; Ogden et al., 2016; Hales et al., 2018). Obesity is linked to several chronic diseases such as diabetes, high blood pressure, asthma, depression, musculoskeletal diseases and cardiovascular diseases. A growing number of policies and programs are implemented to curb this alarming trend. To determine the optimal level of spending towards addressing obesity in children, it is imperative to have an accurate estimate of the effect of obesity on the healthcare system. Medical care expenditure is suitable for two-part modeling because a substantial proportion of youth has zero medical spending in a given year. For instance, 31.9% of the youth in our sample have observed zero amount of medical expenditure. Pohlmeier and Ulrich (1995) also discussed theoretical grounds that justify two-part modeling of medical utilization.

In this section, I illustrate the econometric models and methods discussed in this chapter to estimating *the average incremental effect* of a one-unit increase in BMI on the total medical care spending among the youth in the US. I also estimate the total medical care spending AIE of a hypothetical event that moves every youths' BMI from an average normal to an average obese and severely obese BMI. The potential outcomes specification for the AIE is

$$\begin{aligned}
\mathrm{AIE}(\Delta) = \mathrm{E}\Bigg[&\Big(1 - \mathrm{SG}^{\mathrm{EM}}_{(\zeta^*_{\mathrm{EM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{EM}}, \exp((\mathrm{X}^* + \Delta)\beta^{\mathrm{p}}_{\mathrm{X1}} + \mathrm{X_o}\beta^{\mathrm{p}}_{\mathrm{o1}} + \mathrm{X_u}\beta^{\mathrm{p}}_{\mathrm{u1}}; 1)\Big) \\
&\times \frac{\int_{\zeta_{\mathrm{IM}}}^{\propto} \mathrm{Y}_{\mathrm{X}^*} gg^{\mathrm{IM}}_{\left(\zeta^*_{\mathrm{IM}} \mid \mathrm{Y}_{\mathrm{X}^*} > \zeta_{\mathrm{IM}}, \mathrm{X_o}, \mathrm{X_u}\right)}(\mathrm{Y}_{\mathrm{X}^*}, (\mathrm{X}^* + \Delta)\beta^{\mathrm{p}}_{\mathrm{X1}} + \mathrm{X_o}\beta^{\mathrm{p}}_{\mathrm{o1}} + \mathrm{X_u}\beta^{\mathrm{p}}_{\mathrm{u1}}; \sigma_{\mathrm{IM}}, \kappa_{\mathrm{IM}}) d\mathrm{Y}_{\mathrm{X}^*}}{1 - \mathcal{G}^{\mathrm{IM}}_{(\zeta^*_{\mathrm{IM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{IM}}, (\mathrm{X}^* + \Delta)\beta^{\mathrm{p}}_{\mathrm{X1}} + \mathrm{X_o}\beta^{\mathrm{p}}_{\mathrm{o1}} + \mathrm{X_u}\beta^{\mathrm{p}}_{\mathrm{u1}}; \sigma_{\mathrm{IM}}, \kappa_{\mathrm{IM}})}\Bigg] \\
- \mathrm{E}\Bigg[&\Big(1 - \mathrm{SG}^{\mathrm{EM}}_{(\zeta^*_{\mathrm{EM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{EM}}, \exp(\mathrm{X}^* \beta^{\mathrm{p}}_{\mathrm{X1}} + \mathrm{X_o}\beta^{\mathrm{p}}_{\mathrm{o1}} + \mathrm{X_u}\beta^{\mathrm{p}}_{\mathrm{u1}}; 1))\Big) \\
&\times \frac{\int_{\zeta_{\mathrm{IM}}}^{\propto} \mathrm{Y}_{\mathrm{X}^*} g^{\mathrm{IM}}_{\left(\zeta^*_{\mathrm{IM}} \mid \mathrm{Y}_{\mathrm{X}^*} > \zeta_{\mathrm{IM}}, \mathrm{X_o}, \mathrm{X_u}\right)}(\mathrm{Y}_{\mathrm{X}^*}, \mathrm{X}^* \beta^{\mathrm{p}}_{\mathrm{X1}} + \mathrm{X_o}\beta^{\mathrm{p}}_{\mathrm{o1}} + \mathrm{X_u}\beta^{\mathrm{p}}_{\mathrm{u1}}; \sigma_{\mathrm{IM}}, \kappa_{\mathrm{IM}}) d\mathrm{Y}_{\mathrm{X}^*}}{1 - \mathcal{G}^{\mathrm{IM}}_{(\zeta^*_{\mathrm{IM}} \mid \mathrm{X_o}, \mathrm{X_u})}(\zeta_{\mathrm{IM}}, \mathrm{X}^* \beta^{\mathrm{p}}_{\mathrm{X1}} + \mathrm{X_o}\beta^{\mathrm{p}}_{\mathrm{o1}} + \mathrm{X_u}\beta^{\mathrm{p}}_{\mathrm{u1}}; \sigma_{\mathrm{IM}}, \kappa_{\mathrm{IM}})}\Bigg]
\end{aligned} \tag{69}$$

4.5.1 Identification of the AIE

Our discussion in section 4.2 made it clear that identification of (69) is predicated on the requirement that X_o and X_u induce CIND between X and Y_{X^*}. Because X_u is an essential unobservable confounder, the $\mathbf{X}$ is endogenous and completing the identification requires one to add structure to the conditional pdf that imply the AIE in (69). In the context of the empirical setting that I consider, BMI/obesity can be endogenous because of unobserved health behavior that determine both the observed BMI and the potential medical care spending. For instance, those youth with higher BMI may also have unobserved health behavior that leads to higher medical care spending implying that one would overestimate the AIE without addressing endogeneity. On the other hand, youth from lower socioeconomic status households tend to have higher BMI and lower access to medical care. In this case, ignoring the endogeneity of BMI causes underestimation of the AIE. Thus, addressing the endogeneity of BMI is crucial to obtain causally interpretable AIE.

I follow Biener et al (2020) and use mothers' BMI as an instrumental variable for child BMI. The validity of this instrument is discussed extensively in the literature.[32] Given the IV, we have the GCF-FIML and 2SGCF estimators of the deep parameters of (44) and (38), respectively, based on which the AIE in (69) is estimated. For comparison, I also estimate (69) using the 2SRI, 2SLS and 2SPS estimators given in (62), (66) and (68) respectively. In addition, the FIML-EXOG based AIE estimator is applied to the data. The log-likelihood from the FIML-EXOG MLE is used to test the null hypothesis that X is exogenous. I also estimate a one-part version of the GCF-FIML model to test the null that "no 2PM is needed".

4.5.2 Data and Descriptive Statistics

The data source for the empirical application is the public use version of the Medical Expenditure Panel Survey (MEPS). The household component of the MEPS is a comprehensive, nationally representative data for US civilian population. Information for each member of a participant household is obtained through 5 rounds of survey over two years. The unique feature of MEPS is that information on medical expenditure is

[32] For discussion of Instrumental Variables approach in estimating the medical care cost of obesity, see Biener et al (2020); Cawley and Meyerhoefer (2012); Cawley et al (2015) and Chang and Meyerhoefer (2016).

supplemented by expenditure data directly collected from participants' medical providers and pharmacies through the medical provider component. Like Biener et al (2020), I limit the sample to individuals aged 11-17 as data is missing for several variables for those younger than this age group. The constructed sample also excludes underweight children because the focus is on obesity. The data I use covers the period 2009-2015. Following Biener et al (2020), the total medical care spending is top coded at $50,000.

Beiner et al (2020) exclude children who live with stepmother as doing so minimizes the concern for weak instrument. The biological linkage information is, however, part of the restricted use MEPS data. Thus, I do not distinguish stepmothers' BMI from biological mothers' BMI. To the extent that considerable proportion of children in the sample live with stepmothers, the strength of mothers' BMI as instrument of child's BMI reduces. This is not a concern for our case because the main goal of the empirical application is to demonstrate the implementation of the proposed approach and compare the estimates for the targeted parameters to those obtained using alternative estimators. In fact, the fact that the 2SPS AIE closely matches that in Biener et al (2020) implies that the biological linkage information is unlikely to have substantial effect on estimated AIEs.

Table 8 presents the descriptive statistics of the data for the full sample and by mothers' obesity status (34% of the sample have obese and severely obese mothers). The outcome variable at the EM is whether a child has any medical spending and, in the sample, only 68.1% have positive total medical care spending, suggesting that the 2PM may provide the right framework for this empirical setting. For those children with positive annual expenditure, the average total expenditure is $993.46 with no statistically significant difference by mothers' obesity status.

The average BMI for the full sample is 21.23 while for children with obese and non-obese mothers the corresponding averages are 22.8 and 20.4, respectively. Children's weight is classified into severely obese, obese, overweight and normal weight using the gender-age-specific CDC growth rate. Among children with obese mothers, 33% of children are obese and severely obese and the corresponding rate among those with non-obese mothers is 17.1%, suggesting a potential for high strength of the instrument. The sample statistics for many of the control variables indicate significant differences by mothers obesity status especially for the race, insurance and mother's health status variables. Biener et al (2020) analyzed whether differences in many other controls exist within race/insurance category and found that within most of the groups these differences disappear. Table 9 presents the descriptive statistics for the full sample and by children's obesity status. A striking point in this table is that spending a positive amount and the average expenditure for those who spend is higher among children with overweight and normal weight BMI relative to obese and severely obese children. This could be due to unobserved socioeconomic status that is positively correlated with obesity and negatively correlated with access to medical care.

4.5.3 Empirical Results

I intended to estimate the AIE using the five estimators discussed earlier. The GCF-FIML estimator is burdensome and had convergence problems. I used the 2SGCF estimates as initial values for the one-step GCF-FIML estimator.

4.5.3.1 Estimated AIE Across Different Approaches

Table 11 presents the estimated AIE of the five estimators discussed in sections 4.3.1 and 4.4.3 (i.e., GCF-FIML, 2SGCF, 2SPS, 2SLS and 2SRI) and the FIML-EXOG, GCF-FIML with “no 2PM is needed” null and OLS based AIE estimators.[33] Panel A and B in table 11 report, respectively, the estimated AIE of a one-unit increase in BMI and of a hypothetical change that moves every youth BMI from an average normal to an average obese and severely obese.

The first and second rows in panel A of table 11 show that the GCF-FIML and 2SGCF AIE estimates are almost identical. In both cases, a one unit rise in BMI across the entire youth population in the US leads to a $14.8 AIE on the total medical expenditure. The corresponding estimates based on the 2SPS, 2SRI and 2SLS approaches are $69.6, $117.41 and $45.88, respectively. The result I obtain for the 2SPS is close to that reported in Biener et al (2020) who find a $76 increase in total medical care spending.[34] The GCF-FIML AIE estimator that ignores the two-part structure of medical expenditure estimates a $23.17 increase in total medical expenditure to a one-unit rise in BMI. The FIML-EXOG AIE estimator which ignores endogeneity estimates a $1.64 AIE of a one-unit rise in BMI. Formal statistical tests, however, reject the “no 2PM is needed” and “$\mathbf{X}$ is exogenous” null hypotheses. Finally, I also estimate the AIE based on a linear model that ignores endogeneity of the $\mathbf{X}$. The estimated OLS based AIE is a $3.08 increase in total medical expenditure.

---

[33] Each of these AIE are estimated based on deep parameter estimates from an underlying model. Table 10, for instance, shows the deep parameter estimates of the GCF-FIML model, which are used to compute the GCF-FIML based AIE.

[34] Note that Beiner et al (2020) specified a Generalized Linear Model with gamma variance structure and log link while the specification for the 2SPS approach in this chapter is a GG that subsumes the gamma distribution. Biener et al (2020) also used the biological linkage restricted use MEPS data to exclude children who live with stepmothers.

Like the results from panel A discussed above, the different approaches lead to substantially different estimates for the AIE of a hypothetical change in BMI that moves every youth BMI from average normal to an average obese and severely obese.[35] The GCF-FIML and 2SGCF based AIE estimates reported in the first two rows on panel B show that such change in BMI would lead to an AIE of $143. The corresponding estimates based on the 2SPS, 2SRI and 2SLS approaches are $705.73, $1401.5, and $442.56, respectively. On the other hand, the FIML-EXOG, GCF-FIML with one-part null and OLS estimates indicate that the total medical expenditure, on average, increases by $15.49, $224.35, and $29.01, respectively.

4.5.3.2 Likelihood Ratio Test Results

I applied the likelihood ratio test discussed in section 4.3.2 to test whether parametric distinction between the EM and the IM is needed. The log-likelihoods based on the joint pdf under the "no 2PM is needed" null in (47) and the 2PM with NSD are $\widehat{\mathrm{L}}_{\text{one-part}} = -150{,}638.61$ and $\widehat{\mathrm{L}}_{\text{GCF-FIML}} = -150{,}470.45$, respectively. Thus, the likelihood ratio test statistics is $\mathrm{LR} = -2 \times [-150{,}638.61 + 150{,}470.45] = 336.32$, and $\mathrm{P}(\chi^2_{(31)} > 336.32) = 0$ implying that the 2PM is relevant to this empirical context. I also tested whether BMI is exogenous using a likelihood ratio test. The loglikelihood of the FIML model with no endogeneity and the GCF-FIML model where X is endogenous are $\widehat{\mathrm{L}}_{\text{no endogeneity}} = -150{,}475.37$ and $\widehat{\mathrm{L}}_{\text{GCF-FIML}} = -150{,}470.45$, respectively. The corresponding

[35] In the sample, the average BMI among obese and severely obese youth is 27.77 while the average BMI among normal weight youth is 18.35.

likelihood ratio test statistics is LR= –2 × [-150,475.37 + 150,40.45] = 9.84 and $P(\chi^2_{(2)} > 9.84) = 0.00006$. Therefore, BMI is endogenous.

4.6 Summary and Conclusion

In this chapter a regression-based approach is developed for causally interpretable AIE in the context of a generic FP2PM from the potential outcomes perspective. I consider the case where the IM component of the 2PM is continuous and the causal variable of interest is continuous and endogenous. By casting the AIE within the GPOF, I give unambiguous definition of endogeneity. I propose a new approach – a generalized control function (GCF) – to specify, identify and estimate causally interpretable AIE. Under a distributional assumption for the endogenous variable and regular IV conditions, the approach is shown to satisfy the CIND assumption that is difficult to hold in alternative approaches. Given a FP2PM for the outcome and a fully specified model for the endogenous variable, a FIML model and estimation method is developed for obtaining consistent estimates of the targeted effect parameter. A GG distribution is specified for the EM and IM components of the 2PM as well as for the endogenous variable. The proposed approach is suitable to conduct two important statistical tests: a test for a one-part null and a test for exogeneity of the causal variable. A simulation analysis is conducted to demonstrate the implementation of the GCF-FIML based AIE estimator and validate its consistency. A comparison of this estimator with conventional estimators shows that the proposed estimator performs better. Using data from the MEPS, I apply the approach to estimate the medical care spending effect of an increase in BMI by one-unit, and of moving every youth aged 11-17 from an average normal BMI level to an average obese and severely obese BMI. Following Biener et al (2020), I use mothers' BMI as an instrumental variable for children's BMI. The GCF-FIML and the 2SGCF based AIE estimates are substantially smaller than those obtained using conventional estimators, which are demonstrated to be inconsistent in the simulation study.

Chapter 5

Summary, Discussion and Conclusions

In this dissertation, two regression-based approaches are developed for specification, identification, and estimation of causally interpretable (CI) average incremental effects (AIE). Both of the approaches are cast within the General Potential Outcomes Framework (GPOF) in which effect parameters (EPs) are specified based on relevant counterfactuals. The GPOF also makes clear the conditions under which such EPs can be identified and estimated using observed version of the data. The first approach is developed for specifying, identifying, and estimating causally CI AIE for a Partially Qualitative Outcome (PQO) – an outcome that manifests either as a value in the real line or a qualitative event. Casting a regression model for a PQO within the conventional GPOF is difficult because the only version of a PQO that would be amenable to conventional potential outcomes framework is the one that is conditioned on non-occurrence of the qualitative event. Such conditioning, however, would lead to bias due to bad control. By extending the GPOF, a new measure is proposed that maintains all the essential features of a PQO that is real-valued and is not subject to the bad control critique: a P-weighted conditional outcome. The second approach provides a Fully Parametric Two-Part Model (FP2PM) potential outcomes framework that allows a continuous causal variable to be endogenous. The two-part model (2PM) applies to cases in which the outcome of interest is nonnegative with large fraction of zeros. To accommodate endogeneity within the FP2PM, a generalized control function (GCF) model is specified in which a full information residual is recovered from a fully specified model for the endogenous variable, which in turn serve as a control function for the unobservable in the FP2PM. Given a correct specification for the endogenous variable and under regular instrumental variables assumptions, the GCF approach is shown to satisfy the conditional independence assumption – a key assumption for obtaining CI parameter – that is difficult to hold in alternative approaches. The FP2PM with a distributional assumption for the endogenous variable gives a Full Information Maximum Likelihood (FIML) model which can be estimated via a FIML method. Using flexible distributional assumption, the consistency of the GCF-FIML based estimator for the specified targeted parameter is validated. Although a very flexible distributional assumption is used, to appease for any concern for misspecifications, in the future I plan to develop a distribution free version of the GCF-FIML approach.

Tables

Table 1: Simulation Results of the Partially Qualitative Regression Model

Parameter	**TRUE**	**ML Estimates**						
		500*	**1000**	**5000**	**25000**	**50000**	**100000**	**250000**
$\boldsymbol{\tau}^{\mathbf{o}}_{\mathcal{Q}\mathbf{X}}$	0.15	0.249	0.363	0.088	0.169	0.162	0.144	0.152
$\boldsymbol{\tau}^{\mathbf{o}}_{\mathcal{Q}\mathbf{V}}$	-0.5	-0.5	-0.581	-0.486	-0.49	-0.502	-0.499	-0.5
$\boldsymbol{\tau}^{\mathbf{o}}_{\mathcal{Q}\mathbf{o}}$	12	11.88	13.81	11.76	11.7	12.03	12	12.02
$\boldsymbol{\tau}^{\mathbf{o}}_{g\mathbf{X}}$	-0.004	0.0053	-0.0103	-0.0031	-0.0045	-0.0047	-0.0038	-0.0038
$\boldsymbol{\tau}^{\mathbf{o}}_{g\mathbf{V}}$	0.002	0.0001	0.0018	0.0006	0.0027	0.0021	0.0019	0.002
$\boldsymbol{\tau}^{\mathbf{o}}_{g\mathbf{o}}$	8	8.032	8.01	8.03	7.98	8	8	8
$\boldsymbol{\sigma}^{\mathbf{o}}$	0.175	0.183	0.176	0.177	0.177	0.177	0.175	0.175
$\boldsymbol{\kappa}^{\mathbf{o}}$	0.95	0.839	0.78	0.908	0.963	0.921	0.961	0.949
AIE	66.26	48.49	131.18	43.71	74.02	71.53	62.67	65.77
APB		26.8%	97.9%	34%	11.7%	8%	5.4%	0.7%
Subsample Size		307	596	3011	15181	30217	60168	150067

*Sample sizes are indicated in the second row starting from the third column.

Table 2: Descriptive Statistics of the NSFG Data (Full Sample and By Live Birth Status)

Variable Name	**Full sample**		**Live Birth**		**Non-live Birth**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Outcome Variable						
Live birth	0.784	0.411	1	0		
Birth weight (grams)*	3283.57	613.48	3283.57	613.48		
Policy variable						
If smoker	0.136	0.343	0.125	0.331	0.175	0.38
# cigarettes smoked per day	1.12	3.979	0.908	3.42	1.89	5.477
Control variables						
age	27.04	6.05	26.892	5.838	27.639	6.746
Proportion that is						
Married	0.484	0.5	0.502	0.5	0.42	0.494
Hispanic	0.264	0.441	0.278	0.448	0.214	0.41
Non-Hispanic white	0.455	0.498	0.444	0.497	0.493	0.5
Black	0.227	0.419	0.222	0.415	0.244	0.43
Other race	0.053	0.224	0.055	0.228	0.049	0.215
Less than High school	0.259	0.438	0.262	0.44	0.249	0.432
High school complete	0.269	0.443	0.271	0.445	0.259	0.438
Some college	0.265	0.441	0.259	0.438	0.287	0.453
College	0.207	0.405	0.208	0.406	0.205	0.404
N	15,658		12,274		3,384	

*number of observations for calculating the mean and standard deviation of birth weight in columns 2 and 3, respectively, is based on 12,274 live births.

Table 3: Descriptive Statistics of the NSFG Data (Full Sample and By Smoking Status)

Variable Name	Smoking Women		Nonsmoking Women	
	Mean	**SD**	**Mean**	**SD**
Outcome Variable				
Live birth	0.722	0.448	0.794	0.405
Birth weight (grams)*	3151.15	622.406	3302.48	609.88
Policy variable				
If smoker	1	0	0	
# cigarettes smoked per day	8.249	7.604		
Control variables				
age	25.812	5.84	27.236	6.064
Proportion that is				
Married	0.267	0.443	0.519	0.5
Hispanic	0.1	0.3	0.29	0.454
Non-Hispanic white	0.675	0.469	0.421	0.494
Black	0.192	0.394	0.233	0.423
Other race	0.033	0.178	0.057	0.231
Less than High school	0.417	0.493	0.234	0.424
High school complete	0.31	0.463	0.262	0.44
Some college	0.231	0.422	0.269	0.444
College	0.041	0.198	0.233	0.423
N	2,126		13532	

*number of observations for Birth weight is calculated for a subsample that equals the proportion of live birth by N in the respective column

Table 4: Deep Parameter Estimates of the PQO Model

	Probit estimates		**GG estimates**	
	Column (2)		**Column (4)**	
Variable name	$\hat{\tau}_{\mathcal{Q}}$	**s.e**	$\hat{\tau}_{g}$	**s.e**
# of cigarettes smoked per day	-0.0259***	0.0027	-0.00387***	0.0004
Age at pregnancy	-0.0204***	0.0021	0.00147***	0.0003
Marital status at pregnancy	0.301***	0.0271	0.0135***	0.0035
Hispanic	0.188***	0.0306	-0.00531	0.0039
Black	0.0534*	0.0306	-0.0491***	0.0041
Other races	0.111**	0.053	-0.0210***	0.0067
Less than High school	0.00278	0.0322	-0.00284	0.0041
Some college	-0.0926***	0.0372	0.00428	0.0041
College	-0.0419	0.0371	-0.00199	0.0048
Constant	1.201***	0.0583	2.013***	0.0081
Ancillary Parameters				
Sigma			0.163***	0.001
Kappa			0.934***	0.0153
N	15658		12274	

Note: ***P<0.01, **P<0.05, *P<0.1

Table 5: Estimated AIE of Smoking Ban During Pregnancy

Natility-weighted Birth weight		**Bad Control Birth weight**	
AIE	**se**	**AIE**	**se**
32.68***	3.178	11.35***	0.036

Note: ***P<0.01, **P<0.05, *P<0.1

Table 6: Simulation Results for GCF-FIML Based and Alternative AIE Estimators

True AIE	**2.1449**											
	GCF-FIML		**2SGCF**		**2SRI**		**2SPS**		**2SLS**		**FIML-EXOG**	
n	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**
1K	2.262	5.5%	2.250	4.9%	2.544	18.9%	2.712	26.4%	3.315	54.6%	2.860	33.3%
5K	2.255	5.1%	2.258	5.3%	2.573	20.0%	2.711	26.4%	2.926	36.4%	2.763	28.8%
15K	2.184	1.8%	2.185	1.9%	2.446	14.0%	2.968	38.4%	2.899	35.1%	2.662	24.1%
25K	2.059	4.0%	2.057	4.1%	2.282	6.4%	**	**	2.674	24.7%	2.624	22.3%
50K	2.159	0.6%	2.160	0.7%	2.425	13.1%	**	**	3.051	42.2%	2.704	26.1%
100K	2.109	1.7%	2.109	1.7%	2.360	10.0%	**	**	2.800	30.5%	2.620	22.2%
250K	2.141	0.2%	2.141	0.2%	2.366	10.3%	2.708	26.2%	2.859	33.3%	2.700	25.9%
500K	2.133	0.6%	2.133	0.6%	2.361	10.1%	2.687	25.3%	2.786	29.9%	2.643	23.2%

** Convergence was not achieved.

Table 7: Simulation Results for GCF-FIML Based and Alternative AIE Estimators with Lower Endogeneity and Nonlinearity

True AIE	**2.009**											
	GCF-FIML		**2SGCF**		**2SRI**		**2SPS**		**2SLS**		**FIML-EXOG**	
n	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**	**Est AIE**	**APB**
1K	2.109	5.0%	2.107	4.9%	2.236	11.3%	1.595	20.6%	2.073	3.2%	2.349	16.9%
5K	2.042	1.7%	2.043	1.7%	2.149	7.0%	1.693	15.7%	2.049	2.0%	2.172	8.1%
15K	2.106	4.8%	2.106	4.8%	2.161	5.6%	1.765	12.1%	2.082	3.6%	2.267	12.8%
25K	1.968	2.0%	1.968	2.0%	2.031	1.0%	1.545	23.1%	1.872	6.8%	2.120	5.5%
50K	2.033	1.2%	2.033	1.2%	2.130	6.0%	1.681	16.3%	2.006	0.2%	2.184	8.7%
100K	1.966	2.1%	1.966	2.1%	2.048	1.9%	1.649	17.9%	1.934	3.7%	2.097	4.4%
250K	1.992	2.3%	1.992	2.3%	2.066	2.8%	1.656	17.6%	1.953	2.8%	2.139	6.5%
500K	2.020	0.5%	2.020	0.5%	2.092	4.1%	1.654	17.9%	1.974	1.7%	2.150	7.0%

Table 8: Descriptive Statistics for the MEPS Data (Full Sample and By Mother's Obesity Status)

Variable Name	Full Sample		Mother BMI > 30		Mother BMI < 30	
	Mean	Std. D	Mean	Std. D	Mean	Std. D
Outcomes						
If Any Expenditure	0.681	0.466	0.693	0.461	0.674	0.469
Total Medical Expenditure	993.556	3493.458	1013.307	3350.074	983.362	3565.313
Policy Variable						
BMI	21.227	5.216	22.805	5.950	20.412	4.586
Severely Obese	0.081	0.273	0.134	0.341	0.054	0.225
Obese	0.144	0.351	0.195	0.396	0.118	0.322
Overweight	0.183	0.387	0.209	0.407	0.169	0.375
Normal	0.592	0.491	0.461	0.499	0.660	0.474
Instrument						
Mom BMI	28.319	6.683	35.666	5.577	24.527	3.067
Control Variables						
Age in Month	140.983	37.457	141.641	37.149	140.643	37.611
Female	0.487	0.500	0.492	0.500	0.485	0.500
Hispanic	0.381	0.486	0.439	0.496	0.351	0.477
Black	0.250	0.433	0.257	0.437	0.247	0.431
Medicaid Insurance	0.375	0.484	0.457	0.498	0.332	0.471
Private Insurance	0.534	0.499	0.451	0.498	0.577	0.494
Uninsured	0.065	0.246	0.064	0.244	0.065	0.247
West	0.314	0.464	0.281	0.450	0.331	0.471
Midwest	0.193	0.394	0.201	0.401	0.188	0.391
south	0.358	0.479	0.408	0.491	0.332	0.471
Mother is married	0.915	0.27854	0.897	0.304	0.925	0.264
Mothers' Age group						
35-44 years	0.506	0.500	0.490	0.500	0.515	0.500
45-54 years	0.223	0.416	0.203	0.402	0.233	0.423
55-64 years	0.014	0.117	0.012	0.109	0.015	0.120
Mothers' Education						
High School	0.198	0.398	0.226	0.418	0.183	0.387
Some College	0.236	0.424	0.262	0.440	0.222	0.416
BA Degree	0.145	0.352	0.095	0.293	0.171	0.376
More than BA	0.115	0.319	0.072	0.259	0.138	0.344
Mothers' Health						
Poor or fair (overall health)	0.108	0.310	0.157	0.364	0.083	0.275

Continued

Poor or fair (mental health)	0.053	0.224	0.072	0.258	0.044	0.204
Activity Limitation	0.117	0.322	0.161	0.368	0.095	0.293
Father is married	0.918	0.274	0.900	0.300	0.928	0.259
Fathers' Age group						
35-44 years	0.473	0.499	0.490	0.500	0.465	0.499
45-54 years	0.309	0.462	0.284	0.451	0.321	0.467
55-64 years	0.052	0.223	0.044	0.204	0.057	0.232
Fathers' Education						
High School	0.230	0.421	0.266	0.442	0.211	0.408
Some College	0.192	0.394	0.194	0.395	0.191	0.393
BA Degree	0.122	0.327	0.079	0.269	0.144	0.351
More than BA	0.116	0.320	0.064	0.245	0.143	0.350
N	17,307		5,892		11,415	

Table 9: Descriptive Statistics for the MEPS Data (Full Sample and By Child's Obesity Status)

Variable Name	Full Sample		Obese & above		Below Obese	
	Mean	Std. D	Mean	Std. D	Mean	Std. D
Outcomes						
If Any Expenditure	0.681	0.466	0.663	0.473	0.686	0.464
Total Medical Expenditure	993.55	3493.45	869.83	3317.04	1029.47	3542.34
Policy Variable						
BMI	21.227	5.216	27.771	5.674	19.327	3.118
Severely Obese	0.081	0.273	0.360	0.480	0	0
Obese	0.144	0.351	0.640	0.480	0	0
Overweight	0.183	0.387	0	0	0.236	0.425
Normal	0.592	0.491	0	0	0.764	0.425
Instrument						
Mom BMI	28.319	6.683	30.847	7.421	27.586	6.265
Control Variables						
Age in Month	140.98	37.46	129.59	37.10	144.29	36.91
Female	0.487	0.500	0.413	0.493	0.509	0.500
Hispanic	0.381	0.486	0.499	0.500	0.346	0.476
Black	0.250	0.433	0.199	0.400	0.265	0.441
Medicaid Insurance	0.375	0.484	0.493	0.500	0.340	0.474
Private Insurance	0.534	0.499	0.410	0.492	0.571	0.495
Uninsured	0.065	0.246	0.065	0.247	0.064	0.246
West	0.314	0.464	0.297	0.457	0.319	0.466
Midwest	0.193	0.394	0.178	0.383	0.197	0.398
south	0.358	0.479	0.405	0.491	0.344	0.475
Mother is married	0.915	0.279	0.873	0.333	0.927	0.259
Mothers' Age group						
35-44 years	0.506	0.500	0.491	0.500	0.511	0.500
45-54 years	0.223	0.416	0.167	0.373	0.239	0.426
55-64 years	0.014	0.117	0.009	0.093	0.015	0.123
Mothers' Education						
High School	0.198	0.398	0.228	0.419	0.189	0.392
Some College	0.236	0.424	0.227	0.419	0.238	0.426
BA Degree	0.145	0.352	0.096	0.294	0.159	0.366
More than BA	0.115	0.319	0.074	0.262	0.127	0.333
Mothers' Health						
Poor or fair (overall health)	0.108	0.310	0.146	0.353	0.097	0.296
Poor or fair	0.053	0.224	0.063	0.243	0.050	0.218

(mental health)

Continued

Activity Limitation	0.117	0.322	0.124	0.330	0.115	0.320
Father is married	0.918	0.274	0.878	0.327	0.930	0.255
Fathers' Age group						
35-44 years	0.473	0.499	0.504	0.500	0.464	0.499
45-54 years	0.309	0.462	0.249	0.433	0.326	0.469
55-64 years	0.052	0.223	0.035	0.184	0.057	0.233
Fathers' Education						
High School	0.230	0.421	0.267	0.442	0.219	0.413
Some College	0.192	0.394	0.171	0.377	0.198	0.398
BA Degree	0.122	0.327	0.070	0.254	0.137	0.344
More than BA	0.116	0.320	0.064	0.245	0.131	0.338
N	17,307		3,894		13,413	

Table 10: Deep Parameter Estimates of the GCF-FIML Model

	GG for X			EM			IM		
	Column (1)			Column (1)			Column (1)		
	$\hat{\delta}$	Std. Err.		$\hat{\tau}_{EM}$	Std. Err.		$\hat{\tau}_{IM}$	Std. Err.	
BMI				-0.005	0.0059		0.017	0.0057	***
Xu				0.124	0.0966		-0.271	0.0940	***
Mom BMI	0.004	0.0002	***						
Medicaid	0.008	0.0046	*	-0.310	0.0532	***	-0.034	0.0534	
Private	-0.002	0.0046		-0.091	0.0531	*	0.308	0.0537	***
Uninsured	0.002	0.0065		0.698	0.0643	***	-0.375	0.0877	***
Female	-0.011	0.0024	***	0.022	0.0273		-0.064	0.027	***
Age (months)	0.002	0.0000	***	0.002	0.0005	***	0.001	0.0005	***
Hispanic	0.030	0.0030	***	-0.047	0.0346		-0.142	0.0355	***
Black	-0.003	0.0031		-0.090	0.0372	**	0.009	0.0354	
Midwest	0.000	0.0042		0.100	0.0531	*	0.016	0.0468	
South	0.009	0.0038	**	0.235	0.0477	***	-0.068	0.0428	
West	-0.003	0.0039		0.348	0.0482	***	-0.200	0.0443	***
Mom High School	0.001	0.0039		0.014	0.0424		-0.041	0.0451	
Mom Some College	-0.004	0.0038		-0.156	0.0436	***	0.162	0.0443	***
Mom BA	-0.008	0.0046	*	-0.267	0.0546	***	0.208	0.0524	***
Mom BA plus	-0.005	0.0048		-0.312	0.0570	***	0.296	0.0540	***
Mom Age 35-44	-0.002	0.0035		-0.043	0.0402		0.085	0.0406	**
Mom Age 45-54	0.002	0.0048		-0.058	0.0546		0.100	0.0554	*
Mom Age 55- 64	-0.011	0.0116		-0.168	0.1343		0.598	0.1331	***
Dad High School	0.000	0.0036		0.028	0.0410		0.040	0.0425	
Dad Some College	-0.010	0.0039	*	-0.071	0.0456		-0.055	0.0454	
Dad BA	-0.015	0.0048	***	-0.111	0.0566	*	0.042	0.0541	
Dad BA plus	-0.015	0.0047	***	-0.107	0.0554	*	0.055	0.0536	
Dad Married	-0.018	0.0045	***	-0.138	0.0496	***	0.003	0.0530	
Dad Age 35-44	0.011	0.0040	***	-0.083	0.0451	*	-0.073	0.0459	*
Dad Age 45-54	0.003	0.0048		0.047	0.0544		-0.063	0.0557	
Dad Age 55-64	0.001	0.0073		0.028	0.0826		-0.001	0.0840	
Mom Overall Health	0.004	0.0044		-0.164	0.0517	***	0.104	0.0482	**
Mom Mental Health	-0.006	0.0059		-0.160	0.0727	**	0.410	0.0642	***
Mom Activity Limit	-0.001	0.0039		-0.160	0.0472	***	0.247	0.0431	***
Constant	2.491	0.0100	***	-0.887	0.1118	***	5.338	0.1121	***
Ancillary Parameters									
Sigma-X	-0.156	0.0087	***						

Kappa-X	-1.051	0.0252	***						
Sigma-Y							1.474	0.007	***
Kappa-Y							-0.095	0.0202	***
N		17,307			17,307			11,782	

Note: ***P<0.01, **P<0.05, *P<0.1

Table 11: Estimated AIEs of BMI and Obesity on Total Medical Care Cost

	Panel A Δ = a 1unit increase in BMI			
Estimator	**AIE**	**Asy-SE**	**Asy-t-stat**	**P-value**
GCF-FIML	$14.87	1.77	8.42	0.000
2SGCF	$14.79	1.82	8.11	0.000
2SPS	$69.6	4.41	15.79	0.000
2SRI	$117.41	7.64	15.36	0.000
2SLS	$45.88	27.46	1.67	0.095
FIML-EXOG	$1.64	0.86	1.91	0.056
GCF-FIML (ONE PART)	$23.17	7.52	3.08	0.002
OLS	$3.08	6.85	0.45	0.653

	Panel B Δ = average normal BMI – average of obese and severely obese BMI			
Estimator	**AIE**	**Asy-SE**	**Asy-t-stat**	**P-value**
GCF-FIML	$143.52	17.23	8.32	0.000
2SGCF	$142.74	17.78	8.02	0.000
2SPS	$705.73	45.43	15.53	0.000
2SRI	$1401.5	82.33	17.02	0.000
2SLS	$442.56	267.53	1.65	0.098
FIML-EXOG	$15.49	8.19	1.89	0.058
GCF-FIML (ONE PART)	$224.35	75	2.99	0.002
OLS	$29.01	64.48	0.45	0.653

Appendices

Appendix I

Table A1: Summary Statistics for the Simulated Data in Section 3.6

Variables**	**Observations**	**Mean**	**Std. Dev**	**Min**	**Max**
$\mathbf{X^o}$	250,000	1.75	1	0.018	3.482
$\mathbf{V^o}$	250,000	26.97	7.01	14.88	39.12
Q	250,000	0.4	0.5	0	1
$\mathbf{Y^o}$	150,067	2932.6	585.4	409.9	4908.5
$\mathbf{X^o}$	150,067	1.73	0.99	0.018	3.482
$\mathbf{V^o}$	150,067	31.68	4.44	17.76	39.12

**The first three rows are based on the entire population and the last three rows are calculated for the subpopulation for which the qualitative event does not occur.

Appendix II

Simulation Evidence on the Arbitrariness of κ_{EM}

Here the objective is to establish the validity of an AIE estimator based on the density of the standard gamma distribution, where the value of κ_{EM} is fixed arbitrarily to any value. For simplicity, consider the following version of (47), where X is not endogenous.

$$\Pr(\zeta^{+}_{EM} \leq \zeta_{EM}) = SG\left(\exp(X\beta^{o}_{X1} + X_o\beta^{o}_{o1}); \nu\right). \qquad (47')$$

where $\beta^{o}_{X1} = -p\beta_{X1}$ and β^{o}_{o1} are defined analogous to (47). I conducted the simulation following the steps outlined below.

Step 1 – I picked the values for the parameters that are conjectured to constitute an admissible reduction. In particular, $\zeta_{EM} = 2.5$, $\sigma_{EM} = 1.5$ and $\kappa_{EM} = 2$ and set $\nu = |\kappa_{EM}|^{-2} = 0.25$.

Step 2 – I generated X (not endogenous) and X_o (a two-dimensional vector including a constant term say $X_o = [X^{+}_{o} \quad 1]$. X and X^{+}_{o} are uniform random variables with means and variances E[X] = 1.5, $E[X^{+}_{o}] = 1$, Var[X] = 1 and $Var[X^{+}_{o}] = 0.25$.

Step 3 - For generating the values at the EM, I set the values for the linear index coefficients, as

$$[\beta_{X1} \quad \beta_{o1}'] = [0.15 \quad 1 \quad -0.75]$$

Step 4 – Generate (non-endogenous X) binary EM data $\{I(Y = 0)\}$ based on (47'). The data is generated with a sample size n = 1,500,000.

Step 5 – Set $\nu_{fixed} = |\kappa_{EM}|^{-2}$ where κ_{EM} is the value used to generate the data. Apply the maximum likelihood binary outcome estimator to the simulated data based on the following conditional pdf

$$f_{(Y|X, X_o)}(Y, X, X_o; \beta^o_{X1}, \beta^o_{o1})$$

$$= SG\left(\exp(X\beta^o_{X1} + X_o\beta^o_{o1}); \nu_{fixed}\right)^{I(Y = 0)}$$

$$\times [1 - SG\left(\exp(X\beta^o_{X1} + X_o\beta^o_{o1}), \nu_{fixed}\right)]^{(1 - I(Y = 0))} \qquad (47\text{'a})$$

Note that the log-likelihood function is optimized with respect to β^o_{X1} and β^o_{o1}; not with respect to $\nu_{fixed} = |\kappa_{EM}|^{-2}$ which is held fixed.

Step 6 – Estimate the AIE(1) for X as

$$AIE(1) = [1 - SG\left(\exp((X_i{+}1)\beta^o_{X1} + X_{io}\beta^o_{o1}),\nu_{fixed}\right)]$$

$$-[1 - SG\left(\exp(X_i\beta^o_{X1} + X_{io}\beta^o_{o1}),\nu_{fixed}\right)]$$

Step 7 – I set $\nu_{fixed} = |\kappa_{EM}|^{-2}$ at a different value and repeat Steps 5 and 6. Note that the same large, simulated sample is used to estimate the parameters of (47'a) and the corresponding AIE. I repeated steps 5-7, the results of which is presented in the following table.

Table A2: Simulation Result for Arbitrariness of κ_{EM}

$\boldsymbol{\kappa_{EM}}$	$\boldsymbol{\nu_{fixed} = \lvert\kappa_{EM}\rvert^{-2}}$	**Estimated AIE**
0.1	100	0.0267732
0.25	16	0.026765
0.5	4	0.0267342
1	1	0.0266065
2	0.250	0.0261212
3	0.111	0.0257261
4	0.0625	0.0256069
5	0.04	0.0255895
7.5	0.01778	0.0255879
10	0.01	0.0255879

The simulation results implies that the estimated value of κ_{EM} does not affect the consistency of an AIE estimator. Hence, the reduction of the model with respect to κ_{EM} is admissible.

Appendix III

Sampling Design for a FP2PM with GG EM, GG IM and GG Endogenous Variable: Lower Endogeneity and Nonlinearity

1) To generate pseudo values for X, I set

$$\delta_W' = [\delta_{W^+} \quad \delta_{X_o} \quad \delta_{con}] = [0.75 \quad -0.5 \quad -1]$$

for the linear index coefficients. The means and variances of X_o and W^+ are set to be $E[X_o] = 1$, $E[W^+] = 1$, $Var[X_o] = 0.45$ and $Var[W^+] = 1.5$. I also set values for the shape parameters as $\sigma_X = 0.51$ and $\kappa_X = 0.25$.

2) For generating the values at the EM, I set the values for the linear index coefficients, the shape parameters and the parametric threshold for (51) as follows

$$[\beta_{X1} \quad \beta_{u1} \quad \beta_{o1}'] = [0.5 \quad 0.2 \quad -0.5 \quad 0.25]$$

where $\beta_{o1}' = [\beta_{X_{o1}} \quad \beta_{cons1}]$ are the coefficients for X_o and the intercept, respectively.

$\sigma_{EM} = 0.5$

$\kappa_{EM} = 1$

$\zeta_{EM} = 0.5$

Note that σ_{EM}, κ_{EM} and ζ_{EM} are not identified.

3) Similarly, to generate the pseudo values, the following parameter values for (56) are set

$$[\beta_{X2} \quad \beta_{u2} \quad \beta_{o2}'] = [0.2 \quad 0.1 \quad -0.5 \quad 0.5]$$

where $\beta_{o2}' = [\beta_{X_{o2}} \quad \beta_{cons2}]$ are the coefficients for X_o and the intercept, respectively.

$\sigma_{IM} = 1.5$

$\kappa_{IM} = 1.5$

$\zeta_{IM} = 10$

4) For testing the consistency of the AIE based on the proposed approach, samples of increasing size are generated based on the above sampling design. In particular, the samples are generated with the following sizes.

n = 1,000

n = 5,000

n = 15,000

n = 25,000

n = 50,000

n = 100,000

n = 250,000

n = 500,000

References

Abrevaya, J., and Dahl, C.M. (2008): "The Effects of Birth Inputs on Birth Weight: Evidence from Quantile Estimation on Panel Data," *Journal of Business & Economic Statistics*, 26(4), 379–397.

Aitchison, J. (1955): "On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin," *Journal of American Statistical Association*, 50, 901–908.

Anderson, P.M. and Butcher, K.F. (2006): "Childhood Obesity: Trends and Potential Causes," *The Future of Children*, 16, 19-45.

Angrist, J. D. (2001): "Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business & Economic Statistics*, 19, 2–28.

Angrist, J. D., and Pischke, J.S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Biener, A., Cawley, J., and Meyerhoefer, C., (2020): "The Medical Care Cost of Obesity and Severe Obesity in Youth: An Instrumental Variables Approach," *Health Economics*, 29, 624–639.

Black, N., Hughes, R., Jones, A.M., and Kassenboehmer, S. (2018): "The Health Care Costs of Childhood Obesity in Australia," *Economics of Human Biology*, 31, 1-13.

Bradford, W.D., Kleit, A.N., Krousel-Wood, M.A., and Re R.N. (2002): "A Two-Part Model of Treatment for Benign Prostatic Hyperplasia and the Impact of Innovation," *Applied Economics*, 34, 1291-1299.

Buntin, M.B., and Zaslavsky, A.M. (2004): "Too Much Ado about Two-Part Models and Transformation? Comparing Methods of Modeling Medicare Expenditures," *Journal of Health Economics*, 23, 525-542.

Burney, N.A., Alenezi, M., Al-Musallam, N., and Al-Khayat, A. (2016): "The Demand for Medical Care Services: Evidence from Kuwait Based on Households' Out-of-Pocket Expenses," *Applied Economics*, 48(28), 2636-2650.

Carlson, A. (2020): "Relaxing the Conditional Independence in an Endogenous Binary Response Model," Unpublished Manuscript, Department of Economics, University of Missouri.

Cawley, J., and Meyerhoefer, C. (2012): "The Medical Care Costs of Obesity: An Instrumental Variables Approach," *Journal of Health Economics*, 31(1), 219–230. https://doi.org/10.1016/j.jhealeco.2011.10.003

Cawley, J., Meyerhoefer, C., Biener, A., Hammer, M., and Wintfeld, N. (2015): "Savings in Medical Expenditures Associated with Reductions in Body Mass Index Among US Adults with Obesity, by Diabetes Status," *PharmacoEconomics*, 33(7), 707–722.

Chang, H., and Meyerhoefer, C.D. (2016): "The Causal Effect of Education on Farm-Related Disability: Evidence from a Compulsory Schooling Reform in Taiwan," *American Journal of Agricultural Economics*, 98(5), 1545-1557.

Cragg, J.G. (1971): "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica*, 39(5), 824-844.

Crooks, G.E. (2010): "The Amoroso Distribution," arXiv preprint arXiv:1005.3274.

Duan, N., Manning, W.G., Morris, C.N., and Newhouse, J.P. (1983): "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business & Economic Statistics*, 1(2), 115-126.

Evans, W.N., and Ringel, J.S. (1999): "Can Higher Cigarette Taxes Improve Birth Outcomes?" *Journal of Public Economics*, 72(1), 135–154.

Fertig, A.R. (2010): "Selection and the Effect of Prenatal Smoking," *Health Economics*, 19(2), 209–226.

Grossman, M., and Joyce, T.J. (1990): "Unobservables, Pregnancy Resolutions, and Birth Weight Production Functions in New York City," *Journal of Political Economy*, XCVIII, 983–1007.

Hales, C.M., Fryar, C. D., Carroll, M. D., Freedman, D. S., and Ogden, C.L. (2018): "Trends in Obesity and Severe Obesity Prevalence in US Youth and Adults by Sex and Age, 2007-2008 to 2015-2016," *JAMA*, 319(16), 1723-1725.

Hao, Z., and Terza, J.V. (2018): "Causal Analysis Using Two-Part Models: A General Framework for Specification, Estimation and Inference" Unpublished Manuscript, Department of Economics, Indiana University School of Liberal Arts at IUPUI.

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Heckman, J.J., and Robb, R. (1986): "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," In: Wainer,

H. (Ed) *Drawing Inference from Self-Selected Samples*, New York. Springer, pp. 63–107.

Heckman, J.J., and Navarro-Lozano, S. (2004): "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics*, 86(1), 30–57.

Hernández-Díaz1, S., Schisterman, E.F., and Hernán M.A. (2006): "The Birth Weight "Paradox" Uncovered," American Journal of Epidemiology, 164(11), 115-1120.

Hosmer, DW., and Lemeshow, S. (1995): *Applied Logistic Regression* 2, New York: John Wiley & Sons

Hyland, A., Piazza, K., Hovey, K.M., Tindle, H.A., Manson, J.E., Messina, C., Wactawski-Wende, J. (2016): "Associations between Lifetime Tobacco Exposure with Infertility and Age at Natural Menopause: The Women's Health Initiative Observational Study," *Tobacco Control*, *25*(6), 706–714. https://doi.org/10.1136/tobaccocontrol-2015-052510

Hyun, Kyung-Rae, Sungwook Kang and Sunmi Lee (2016): "Population Aging and Healthcare Expenditure in Korea," *Health Economics*, 25, 1239-1251.

Imbens, G., and Newey, W.K (2002): "Identification and Estimation of Triangular Simultaneous Equations Model without Additivity," *NBER Technical Working Paper* 285.

Imbens, G. (2007): "Non-additive Models with Endogenous Regressors," In: Blundell, R., Newey, W.K., and Persson, T. (Eds), *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, 3rd ed, Cambridge and New York: Cambridge University Press.

Imbens, G., and Newey W.K. (2009): "Identification and Estimation of Triangular Simultaneous Equations Model without Additivity," *Econometrica*, 46, 33-50.

Li, B., Cairns, J., Fotheringham, J., and Ravanan, R. (2016): "Predicting Hospital Costs for Patients Receiving Renal Replacement Therapy to Inform an Economic Evaluation," *European Journal of Health Economics*, 17, 659-668.

Lien, D. S., and Evans, W.N. (2005): "Estimating the Impact of Large Cigarette Tax Hikes: The Case of Maternal Smoking and Infant Birth Weight," *Journal of Human Resources*, *XL*(2), 373–392. https://doi.org/10.3368/jhr.XL.2.373

Liew, Z., Olsen, J., Cui, X., Ritz, B.R., and Arah, O.A. (2015): "Bias from Conditioning on Live Birth in Pregnancy Cohorts: An illustration Based on Neurodevelopment in Children After Prenatal Exposure to Organic Pollutants," *International Journal of Epidemiology*, 44(1), 345–354. https://doi.org/10.1093/ije/dyu249

Lisonkova, S., and Joseph, K.S. (2015): "Left Truncation Bias as a Potential Explanation for the Protective Effect of Smoking on Preeclampsia," *Epidemiology* (Cambridge, Mass.), 26(3), 436–440. https://doi.org/10.1097/EDE.0000000000000268

Liu, Lei, Robert L., Strawderman, Mark E. Cowen and Ya-Chen T. Shih (2010): "A Flexible Two-Part Random Effects Model for Correlated Medical Costs," *Journal of Health Economics*, 29, 110-123.

Lumley, J., Oliver, S., Chamberlain, C., and Oakley, L. (2004): "Interventions for Promoting Smoking Cessation During Pregnancy." In: *The Cochrane Collaboration (Ed.)*, Cochrane Database of Systematic Reviews (p. CD001055.pub2). https://doi.org/10.1002/14651858.CD001055.pub2

Madden, D. (2008): "Sample Selection Versus Two-Part Models Revisited: The Case of Female Smoking and Drinking," *Journal of Health Economics*, 27, 300-307.

Manning, W.G., Basu, A., and Mullahy, J. (2005): "Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data," *Journal of Health Economics*, 20, 465–488.

Manning, W.G. and Mullahy, J. (2001): "Estimating Log Models: To Transform or Not to Transform?" *Journal of Health Economics*, 20, 461–494.

Mishra, G.D., Dobson, A.J., and Schofield, M.J. (2000): "Cigarette Smoking, Menstrual Symptoms and Miscarriage among Young Women," *Australian and New Zealand Journal of Public Health*, *24*(4), 413–420.

Ness, R.B., Grisso, J.A., Hirschinger, N., Markovic, N., Shaw, L.M., Day, N.L., and Kline, J. (1999): "Cocaine and Tobacco Use and the Risk of Spontaneous Abortion," *New England Journal of Medicine*, *340*(5), 333–339.

Newey, W. K., Powell J. L., and Vella, F. (1989): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Ogden, C.L., Carroll, M.D., Kit, B.K., and Flegal, K.M. (2012): "Prevalence of Obesity and Trends in Body Mass Index Among US Children and Adolescents, 1999-2010," *Journal of the American Medical Association*, 307, 483-490.

Ogden, C.L., Carroll, M.D., Lawmand, H.G., Fryar, C.D., Kruszon-Moran, D., Kit, B.K., Flegal, K.M. (2016): "Trends in Obesity Prevalence Among Children and Adolescents in the United States, 1988-1994 through 2013-2014," *Journal of the American Medical Association*, 315(21), 2292-2299.

Pineles, B.L., Park, E., and Samet, J.M. (2014): "Systematic Review and Meta-analysis of Miscarriage and Maternal Exposure to Tobacco Smoke During Pregnancy," *American Journal of Epidemiology*, *179*(7), 807–823.

Pohlmeier, W., and Ulrich, V. (1995): "An Econometric Model of the Two-Part Decision-Making Process in the Demand for Health Care," *Journal of Human Resources*, 30, 339–361.

Pregibon D. (1980): "Goodness of Link Tests for Generalized Linear Models," *Applied Statistics*, 29, 15–24.

Ross, H., and Chaloupka, F.J. (2003): "The Effect of Cigarette Prices on Youth Smoking," *Health Economics*, 12, 217-230.

Rosenzweig, M.R., and Schultz, T.P. (1983): "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and their Effects on Birth Weight," *Journal of Political Economy*, 91(5), 723–746.

Ross, S. (1997): *Simulation*, 2nd Ed., San Diego: Academic Press.

Rubin, Donald B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66 (5):688-701. doi: 10.1037/h0037350.

Smith, V.A, Neelon B, Preisser J.S, and Maciejewski M.L. (2015): "A Marginalized Two-part Model for Longitudinal Semicontinuous Data," *Statistical Methods for Medical Research*, 26(4), 1949-1968.

Stacy, E.W., and Mihram, G.A. (1965): "Parameter Estimation for a Generalized Gamma Distribution," *Technometrics*, 7, 349‑358.

Suarez, E.A., Landi, S.N., Conover, M.M., and Funk, J.M. (2018): "Bias from Restricting to Live Births When Estimating Effects of Prescription Drug Use on Pregnancy Complications: A Simulation," *Pharmacoepidemiology and Drug Safety*, 27(3), 307–314. https://doi.org/10.1002/pds.4387

Terza, J. (1985): "Reduced-Form Trinomial Probit: A Quantal Response Model without A Priori Restrictions," *The Journal of Business and Economics Statistics*, 3, 54-59.

Terza, J.V., Basu, A., and Rathouz, P.J. (2008): "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling," *Journal of Health Economics*, 27(3), 531-43.

Terza, J.V. (2009): "Parametric Nonlinear Regression with Endogenous Switching." *Econometric Reviews*, 28, 555-580.

Terza, J.V. (2016a): "Simpler Standard Errors for Two-Stage Optimization Estimators," *the Stata Journal*, 16, 368-385.

________(2016b): "Inference Using Sample Means of Parametric Nonlinear Data Transformations," *Health Services Research*, 51, 1109-1113.

________(2016c): "Supplementary Appendix to 'Inference Using Sample Means of Parametric Nonlinear Data Transformations,'" *Health Services Research*, DOI: 10.1111/1475-6773.12494.

________ (2017): "Causal Effect Estimation and Inference Using Stata," the *Stata Journal*, 939-961.

________ (2019a): "Regression-Based Causal Analysis from the Potential Outcomes Perspective," *Journal of Econometric Methods*, 9:1, DOI: https://doi.org/10.1515/jem-2018-0030.

________ (2019b): "Enodogeneity and Regression-Based Causal Analysis from the Potential Outcomes Perspective," Unpublished Manuscript, Department of Economics, Indiana University, School of Liberal Arts at IUPUI.

________ (2019c): "Estimating Nonlinear Models with Endogenous Regressors: Explicit and Implicit Control Function Methods," Unpublished Manuscript, Department of Economics, Indiana University, School of Liberal Arts at IUPUI.

U.S. Department of Health and Human Services. (2009): "Reducing Alcohol-exposed Pregnancies," A report of the National Task Force on Fetal Alcohol Syndrome and Fetal Alcohol Effect.

Walsh, R.A. (1994): "Effects of Maternal Smoking on Adverse Pregnancy Outcomes: Examination of the Criteria of Causation," *Human Biology*, 66(6), 1059–1092.

Wilcox, A.J. (1993): "Birth Weight and Perinatal Mortality: the Effect of Maternal Smoking," *American Journal of Epidemiology*, 137, 1098–1104.

Wilcox, A. (2001): "On the Importance and the Unimportance of Birthweight," *International Journal Epidemiology*, 30, 1233–1241.

Wooldridge, J. (2005): "Violating Ignorability of Treatment by Controlling for Too Many Factors." *Econometric Theory*, 21, 1026–1028.

Curriculum Vitae

Daniel Abebe Asfaw

Position

- Postdoctoral Associate, School of Public Health, Department of Health Law, Management and Policy, Boston University, March 2021-

Education

- Ph.D. in Economics, Indiana University, Indianapolis, 2021.
- Advanced Master's Degree in Development and Globalization, University of Antwerp, Belgium, 2016.
- MSc in Economics, Addis Ababa University, Ethiopia, 2013.
- BA in Economics, Haramaya University, Ethiopia, 2007

Research Fields

- Applied Econometrics, Health Economics, Development Economics.

Work Experience

- Teaching Fellow at AEA Summer Program at Michigan State University, Summer, 2019.
- Research Assistant, HKP Properties inc. Indianapolis, IN Sept 2016- Sept 2017.
- Instructor, IUPUI, 2017-2020
- Research Assistant, Institute for African Economic Studies, 2011.
- Instructor, Haramaya University, 2007-2014

Awards and Scholarships

- IUPUI-Carlin award for Outstanding Graduate Student Paper in Empirical Economics, March 2020.
- Outstanding Teaching Fellow Award – AEA Summer Program, Summer 2019
- Loretta-Lunsford Scholarship ($5000) – IUPUI Spring 2019
- Annual Health Econometrics Workshop Student Scholarship – September 2019
- Graduate and Professional Education Grant, Fall 2019, Fall 2018
- Graduate Student Travel Fellowship Award
- IUPUI, Department of Economics Travel Grant
- African Economic Research Consortium (AERC) Scholarship – Summer 2011

Conference Presentations

- Mini Health Economics Conference, IU Bloomington, December 12-13, 2019.
- Southern Economics Association, 88th Annual Meeting, Washington DC, November 18-20, 2018.
- The 26th Midwest Econometrics Group, University of Wisconsin Madison, October 26-27, 2018.
- The 89th Indiana Academy of Social Sciences (IASS), Indiana University Southeast, Indiana, October 12, 2018.
- The 88th Midwest Economics Association Annual Meeting, Evanston, Illinois, March 18-20, 2018.
- New Generation Dataset on Child Development, Human Capital and Economic Opportunity (HCEO) Group, November 2-3, 2018, Brooklyn, New York. (Attended by Invitation)