ROBUST PARAMETRIC ESTIMATORS FOR HEALTH ECONOMETRIC MODELS WITH SKEWED OUTCOMES AND ENDOGENOUS REGRESSORS

By

MUJDE ZEYNEP ERTEN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2010

To my family

ACKNOWLEDGMENTS

I am grateful to Dr. Joseph V. Terza, my advisor, without his guidance this dissertation would not be possible. He was not only an advisor but a mentor to me during this experience. His deep knowledge and great enthusiasm for the subject combined with his tolerant and pleasant personality helped me through this process.

I would like to thank my committee members Dr. David E. M. Sappington, Dr. Chunrong Ai and Dr. Bruce Vogel for their help and guidance through constructive comments.

I owe my deepest gratitude to Dr. Steven Slutsky and Dr. Jonathan Hamilton for their support throughout my studies. I would also like to thank all my professors in the Department of Economics at the University of Florida for generating the most supportive and creative environment to pursue my graduate studies. I have had invaluable experience during my teaching and research assistantships.

I am grateful to the Institute for Child Health Policy at the University of Florida, the Department of Economics at the University of Florida, and the Agency for Healthcare Research and Quality under a grant to Dr. Joseph V. Terza (#R01 HS017434-01) for providing me financial support throughout my graduate studies.

I would like to extend my gratitude to Dr. Sema Aydede and Dr. Elizabeth Shenkman for introducing me to the health economics area. The work I did with them ignited my interest to health economics.

I would also like to thank Dr. Bruce Stuart and Dr. John Mullahy for providing the data sets used in this dissertation.

My family has been there for me through all the good times and the bad times. I am grateful for their support, my parents Nail Erten and Muhterem Erten, and my sisters

Selda Erten, Ferah C. Erten, Hande K. Erten, and Hale Erten. Special thank goes to my sister Hale Erten for her invaluable support, and incessant help.

Last but not the least I would like to thank Ritwik Kumar for his endless support and help – especially for all the brainstorming and programming.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ROBUST PARAMETRIC ESTIMATORS FOR HEALTH ECONOMETRIC MODELS WITH SKEWED OUTCOMES AND ENDOGENOUS REGRESSORS

By

Mujde Z. Erten

August 2010

Chair: Joseph V. Terza
Major: Economics

Empirical models in health economics and health services are commonly characterized by skewness in the regression outcome and endogeneity among the regressors. Though some methods in the literature have addressed skewness and endogeneity separately, none comprehensively accounts for both of them together. A new estimator is introduced that combines the generalized gamma approach, which flexibly accommodates skewness in the dependent variable; and the two-stage residual inclusion method, designed to account for endogeneity. This extended generalized gamma model is the first application of two-stage residual inclusion method to healthcare expenditure analysis in the full information maximum likelihood context. In order to evaluate the generalized gamma with endogeneity estimator, extensive simulation analyses were conducted with data generated using various sampling designs. The simulations results show that this method matches or outperforms other estimation methods that are widely used in the literature. Using the generalized gamma with endogeneity method, the offsetting effect of appropriate prescription drug utilization on hospital costs was estimated in a two-part modeling context. The generalized gamma model, though effective in addressing skewness, imposes a fixed functional

form for the conditional mean regression function. This may lead to misspecification bias. Using a variant of the inverse Box-Cox model, in this dissertation, we developed a more robust version of the generalized gamma method. Using simulated data, this new estimator is tested using a flexible conditional mean specification and compared with alternative estimators. The simulation analyses highlight the advantages of our model in terms of bias and precision. This estimator is applied to estimate the effect of cigarette smoking during pregnancy in a model of birthweight. Finally, this dissertation extends the generalized gamma model in both directions by incorporating the flexible inverse Box-Cox functional form and accommodation for endogenous regressors via the two-stage residual inclusion method. Using simulated data, the bias and efficiency properties of this new estimator are evaluated in comparison with the alternative parametric estimators. The suggested model is applied to birthweight data to estimate the effect of potentially endogenous cigarette smoking variable.

CHAPTER 1
INTRODUCTION

Health outcomes data has unique properties that need to be carefully addressed in empirical modeling. The researcher is mainly interested in outcome variables such as health costs and utilization that are non-negative by nature. Healthcare expenditures and utilization in the typical population is distributed asymmetrically with a small proportion of the population with chronic illnesses have high levels of utilization and spending and a very high proportion of the population that is healthy spend very little. This high concentration of health care expenditures and utilization by the chronically ill generates a skewed distribution of the health outcomes data. In the health econometrics area, researchers have proposed different estimators to deal with the skewness issue. These range from the nonlinear least squares (NLS) method that is based on only a conditional mean assumption to the full information maximum likelihood (FIML) estimation techniques that allow for all of the conditional moments. All of these methods have their shortcomings. Although the NLS model is unbiased since it does not make any assumptions related to skewness it is less efficient. FIML methods are prone to misspecification bias. They do, however, take into account higher order conditional moments including skewness. In the presence of this trade-off, the search is for models that are both efficient and flexible.

Endogeneity in the regressors is another common problem in health outcomes modeling. This problem is caused by the presence of unobservable confounders – latent variables that are correlated with both the dependent variable and one or more of the regressors. This typically causes bias in estimation because the conventional regression methods that do not account for the problem will spuriously attribute the

effects of the unobservable confounders to the observable regressors. There is a sizable literature on the development and application of methods to resolve this issue. One of the common methods used in the health outcomes literature is the two-stage least squares (2SLS) estimator. This method generates consistent estimates for linear models, however, for nonlinear models the results are not consistent. Health outcomes models are generally nonlinear so alternative techniques for endogeneity correction are needed. One method that has recently come to the fore in this context is the two stage residual inclusion (2SRI) method (Terza et al., 2008a). Recent applications of this method can be found in Terza et al. (2008b), Carpio et al. (2008), Zhang (2008), Vandegrift and Yavas (2009), Bradford et al. (2010), Etile and Jones (2009), Richardson et al. (2010), etc. Terza et al. (2008a) show that rote analogs to 2SLS in the nonlinear context are not consistent while proving that 2SRI is a consistent method in the nonlinear setting.

In regression analysis the conditional mean is typically defined as a linear combination of the regressors. This linearity assumption might lead to bias if the true conditional mean regression model is actually nonlinear. Wooldridge (1992) employed the inverse Box-Cox (IBC) transformation for the conditional mean function for the first time in a regression analysis. Applications of the IBC transformation include Kenkel and Terza (2001), Terza et al. (2008a), Terza et al. (2008b) and Basu and Rathouz (2005). The first three of these implement NLS with IBC conditional mean regression specification. This approach is unbiased yet inefficient since it is based on the NLS. Basu and Rathouz (2005) included the IBC transformation in the generalized linear model (GLM) and seek to improve efficiency by explicitly accounting for

heteroscedasticity in the estimation. GLM modeling is, however, prone to misspecification bias and is not as flexible as an NLS model.

In the literature there are some papers pointing out these challenges in modeling health outcomes, however, each of these studies focuses on only one of these problems at a time. Manning et al. (2005) proposes the use of a FIML model that not only takes into account the higher order moments such as skewness, kurtosis, etc. but is also flexible by nature. The fully parametric formulation of this model specifies the density of the dependent variable conditional on the regressors as a generalized gamma (GG). The GG model subsumes widely used distributions in itself such as the log-normal, the Weibull, the exponential and the standard gamma. Manning et al. (2005) concentrates on the efficiency properties of the GG model for skewed, thick tailed and heteroscedastic outcomes. The use of GG model is also suggested in other papers such as Basu and Manning (2006) and Hill and Miller (2009).

The endogeneity issue is widely examined in all areas of empirical econometrics, since it is a common problem in empirical modeling. Most of the instrumental variable models proposed to handle this are very restrictive in their assumptions. In case of non-linear models there are some options but they have their own shortcomings. In this regard the 2SRI model comes forward since it is a consistent model. The second chapter in this dissertation proposes a model that would not only take care of skewness in the outcome variable but will also take care of endogeneity in the non-linear context by taking advantage of the two aforementioned methods – FIML-GG (Manning et al., 2005) and 2SRI (Terza et al., 2008a). This new model incorporates the 2SRI approach into the GG framework. We name this new model generalized gamma with endogeneity

(GGE). It is flexible since it subsumes all of the aforementioned FIML models that are members of the GG family of distributions. Because it is cast in the framework of a FIML estimator, GGE is likely to approach full efficiency in parametric estimation. The 2SRI component of the GGE method takes care of biases due to endogeneity which ensures the consistency of the method.

In the third chapter the issue of endogeneity is, for the moment, set aside. The focus in this chapter is on misspecification bias due to the functional form of the conditional mean regression model. Here we incorporate the IBC conditional mean regression specification into the GG model. This new hybrid model (generalized gamma with inverse Box-Cox transformation (GG-IBC)) yields efficient parameter estimates because it is a FIML estimator. The inclusion of the IBC transformation leads to a more flexible regression estimator. The NLS model with an IBC conditional mean regression specification (NLS-IBC) (Wooldridge, 1992; Kenkel and Terza, 2001; Terza et al., 2008a; Terza et al., 2008b; Basu and Rathouz, 2005) is a reasonable alternative to the GG-IBC model, because both models will yield consistent estimates but the proposed GG-IBC method will likely be more efficient. We compare the two models in our simulation and real data analysis.

In the fourth chapter, we introduce an estimator that brings together all the important elements of the second and the third chapters – conditional density flexibility (skewness accommodation) of the GG model, endogeneity correction as provided by the 2SRI approach, and flexibility in the conditional mean regression as provided by the IBC model. We call the proposed estimator the generalized gamma with endogeneity and IBC transformation (GGE-IBC). Comparisons of this model with the NLS with

endogeneity and IBC transformation (NLS-IBC-2SRI) and the GGE are also provided as part of our simulation and real data analysis.

This dissertation is organized in the following manner. The Chapter 2 introduces the GGE model including a description of the GG and 2SRI models. A simulation analysis comparing the GGE estimator to widely used alternative estimators in the literature is conducted. The bases for comparison are the bias and the efficiency of the estimators. The estimator is then also applied to real world data in an examination of the effect of prescription drug utilization on the hospital cost offsets. The Chapter 3 develops and details the proposed GG-IBC model and follows a similar simulation analysis and a real data application agenda as that of Chapter 2. Here the GG-IBC estimator is compared to the GG and the NLS-IBC estimators. As our real data experiment we re-examine Mullahy's (1997) model of the effect of mothers' cigarette smoking during pregnancy on newborn birthweight (ignoring the potential endogeneity of smoking in this model). The Chapter 4 describes the GGE-IBC model and details simulation and real data comparisons with the GGE method and the 2SRI variant of the NLS-IBC estimator (NLS-IBC-2SRI). The analysis of the impact of smoking during pregnancy on birthweight introduced in Chapter 3 is extended in Chapter 4 to account for potential endogeneity. Finally in Chapter 5 we summarize our findings, discuss topics for future research and present concluding remarks.

CHAPTER 2
SKEWED OUTCOMES AND ENDOGENOUS REGRESSORS: PRESCRIPTION DRUG UTILIZATION AND HOSPITAL COST OFFSETS

2.1 Introduction and Background

Skewness in the regression outcome (y) is prevalent in empirical models in health economics. The most common examples are health care expenditures and utilization data (e.g., hospital expenditure, the number of physician visits, the number of prescriptions filled, etc.).[1] There are a variety of parametric estimators that can be implemented in such cases. These range from the nonlinear least squares (NLS) methods that require only a conditional mean regression assumption to the full information maximum likelihood (FIML) approaches that require knowledge of the conditional probability density of the dependent variable given the regressors (x).

The NLS methods are relatively robust to misspecification bias because only the conditional mean regression need be specified. They do little, however, to account for skewness [or any of the other higher order moments of the conditional distribution of y given the regressors (y | x) for that matter] and are, therefore, relatively inefficient. The FIML estimators are more susceptible to bias, but are more efficient because they impose maximal parametric structure. Manning et al. (2005) argue the importance of taking into account skewness in health econometric models. They suggest a FIML estimator that accounts for skewness but avoids misspecification bias because it is based on a flexible distributional form – the generalized gamma (GG). The flexibility of the GG is evidenced by the fact that it subsumes, as special cases, some of the

[1] Health care utilization and expenditures data are generally skewed to the right due to the high frequency of healthy people with zero health care utilization or expenditures, and the low frequency of people with chronic diseases.

commonly used distributions found in the literature such as the standard gamma, the exponential, and the Weibull. According to the simulation based findings of Manning et al. (2005), the GG estimator offers substantial potential efficiency gains due to the fact that it accounts for skewness [and all other higher-order moments of (y | x)].

Another important issue commonly encountered in empirical models in health economics and health services research is endogeneity among the regressors – i.e., the presence of unobservable confounding influences – resulting in biased estimates of the model parameters and related causal effects.[2] Endogeneity is commonly caused by omitted regressors, simultaneity between a regressor and the outcome variable y, or errors in regressors. For example, Shea et al. (2007) analyzed the effect of prescription drug coverage on the number of drug prescriptions filled – an outcome variable that is highly skewed. Here, prescription drug coverage is likely to be endogenous because the unobservable variables with which it is correlated are also likely to influence drug utilization (e.g., unobserved health status – a confounding influence).

In nonlinear settings, endogeneity is often handled by linearizing the model and applying the conventional two-stage least squares (2SLS) method, or by implementing a two-stage predictor substitution (2SPS) approach – a direct analog to 2SLS for nonlinear models. Terza et al. (2008b) demonstrate that the former is likely to be substantially biased and Terza et al. (2008a) prove that the latter is generally inconsistent. Terza et al. (2008a) also established the consistency of an alternative method for endogeneity correction in nonlinear setting called two-stage residual inclusion (2SRI). As noted above, though some methods in the literature have handled

[2] By definition, a confounding influence (variable) is one that affects y but is correlated with one or more of the elements of x.

skewness and endogeneity separately (e.g., Manning et al., 2005; Terza et al., 2008a, respectively), none comprehensively accounts for both of them together. In this paper, to address this void, we extend the GG estimator as presented by Manning et al. (2005) by incorporating the 2SRI method discussed by Terza et al. (2008a).[3] We call this new estimator the GG model with endogeneity (GGE). This model simultaneously maintains the substantial potential efficiency gains provided by the GG specification and corrects for endogeneity via the 2SRI method. It is noteworthy that this is the first application of 2SRI to expenditure analysis in the full information maximum likelihood (FIML) context.

The rest of the chapter is organized as follows: In Section 2.2, we introduce the GGE estimator and describe how 2SRI can be implemented in conjunction with the GG model to deal with endogeneity. In Section 2.3, the details of a simulation study are presented. Therein we: (i) summarize the different data generation techniques used in the simulation analysis; (ii) briefly describe the estimation techniques to be included in the simulation comparison; (iii) detail the statistical criteria for comparison; and (iv) discuss the simulation analysis results. Then, for the purpose of illustration, in Section 2.4, the new estimator is applied to the same data analyzed by Stuart et al. (2009) in their examination of the impact of prescription drug use on hospital costs. Finally in Section 2.5, we summarize and conclude.

2.2 Accounting for Endogeneity in the Generalized Gamma Estimation Framework

A common problem in the health outcomes data is asymmetry in the distribution of the dependent variable (y), namely skewness. Usually the existence of an upper or

[3] Note that, for the special case in which the outcome regression and the auxiliary regression are both linear, 2SRI is equivalent to the conventional linear IV (two-stage least squares) method (Terza et al., 2008a).

lower bound on y leads to this type of asymmetry – as is the case for many health outcome variables such as prescription drug utilization or outpatient expenditures. Estimation methods that require only a conditional mean regression assumption, while relatively robust to conventional misspecification error, do not account for skewness in the data and are, therefore, relatively inefficient. The FIML estimators that account for additional information on the other moments of the distribution of the outcome conditional on the regressors are more prone to bias but are relatively efficient. The GG estimator discussed by Manning et al. (2005) provides a reasonable and practical compromise for the robustness vs. efficiency tradeoff in that it accounts for skewness in a parametrically flexible way.

Endogeneity among the regressors is another important problem widely observed in health data models. Endogeneity can result in biased estimates of the model parameters and related causal effects. Sampling is subject to endogeneity when one or more of the regressors are correlated with the unobservable determinants of the outcome variable. Specific forms of endogeneity include, omitted regressors, measurement error and simultaneity (reverse causality) between regressors and the outcome. Our aim is to introduce a new version of the GG estimator that corrects for endogeneity. The main source of difficulty encounter by applied researchers seeking to use instrumental variables (IV) in this context is the typical nonlinearity of the relevant regression model (and often the other conditional moments). The GG model is no exception in this regard. We resolve this issue by incorporating the 2SRI approach into the GG model.

2.2.1 The Generalized Gamma Model and Popular Special Cases

The GG estimator is based on a full parametric assumption, i.e., the conditional probability density function is fully specified. The GG probability density function has three parameters – κ, μ and σ. It embodies many different distributions as special cases, such as the standard gamma, the log-normal, the exponential, and the Weibull. Because these models are nested within the GG framework, standard likelihood-ratio and/or Wald tests can be used as a means of detecting and selecting specific cases. The following gives a brief summary of these nested cases for particular values of the parameters κ and σ, and the specification of the conditional mean in each case.

The conditional probability density, f(y | x), is assumed to have the following form

$$f(y;\kappa,\mu,\sigma) = \frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\Gamma(\gamma)}\exp\left[z\sqrt{\gamma} - u\right] \qquad y \geq 0 \tag{2-1}$$

where $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)\{\ln(y) - \mu\}/\sigma$, and $u = \gamma\exp(|\kappa|z)$; with κ, μ and σ as the basic parameters of the distribution ($1/|\kappa|^2 > 0, \kappa/\sigma > 0,$ and $\exp(\mu)/\left(1/|\kappa|^2\right)^{\sigma/\kappa} > 0$).

Moreover, it is assumed that

$$\mu = x\beta^* \tag{2-2}$$

where x is the $1 \times K^*$ row vector of regressors, $\beta^* = \left[\beta_1^* \;\; \beta_2^* \;\; \ldots \;\; \beta_K^*\right]'$ is a column vector of regression parameters conformable with x. Assuming that the first element of x is equal to one, β_1^* is the regression constant term. Under the assumed distribution in (2-1), it can be shown that

$$E[y \mid x] = \exp\left(x\beta^* + C(\sigma,\kappa)\right) \tag{2-3}$$

where

$$C(\sigma,\kappa) = \left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right). \tag{2-4}$$

We can express equation (2-3) as

$$E[y \mid x] = \exp(x\beta) \tag{2-5}$$

where $\beta = \left[\beta_1 \;\; \beta_2^* \; ... \; \beta_K^*\right]'$ and β is the same as β* except for its first element, the constant term, which becomes $\beta_1 = \beta_1^* + C(\sigma,\kappa)$.

The GG estimator encompasses a number of familiar special cases corresponding to various combinations of values for the parameters κ and σ.

Standard gamma: When the shape parameters are equal, i.e., $\kappa = \sigma$, and are strictly positive, the GG distribution reduces to the standard gamma distribution, i.e.,

$$f(y \mid x) = \frac{\gamma^\gamma}{y\Gamma(\gamma)}\exp\left[z\sqrt{\gamma} - \gamma\exp(\sigma z)\right] \qquad y \geq 0. \tag{2-6}$$

In this case, since $\kappa = \sigma$, the constant term shift (2-4) in the reparametrization of the conditional mean (2-5) is $C(\sigma, \sigma) = 0$.[4]

Weibull: If the value of the shape parameter κ is equal to 1, the GG distribution reduces to the Weibull distribution, i.e.,

$$f(y \mid x) = \frac{1}{\sigma y}\exp\left[z - \exp(z)\right] \qquad y \geq 0. \tag{2-7}$$

Here the shift in the constant term is $C(\sigma, 1) = \ln\left(\Gamma\left(1+\sigma\right)\right)$.[5]

[4] Appendix A includes the formal derivation of the reparametrization of the conditional mean of the standard gamma.

[5] Appendix A includes the formal derivation of the reparametrization of the conditional mean of the Weibull.

Exponential: When both shape parameters are set to unity, i.e., $\kappa = \sigma = 1$, the GG distribution reduces to the exponential distribution, i.e.,

$$f(y \mid x) = \exp\left[\frac{y}{\exp(\mu)} - \mu\right]. \tag{2-8}$$

For the exponential distribution $C(1,1) = 0$.[6]

Log-normal: If κ approaches zero in the limit, the GG distribution approaches the log-normal distribution, i.e.,

$$f(y \mid x) = \frac{1}{\sigma y\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \qquad y \geq 0. \tag{2-9}$$

For the log-normal $C(\sigma, \kappa \to 0) = \frac{\sigma^2}{2}$.[7]

2.2.2 Accounting for Endogenous Regressors

Now suppose that some of the elements of x are endogenous. To allow for this possibility, we combine the GG formulation defined in (2-1) with the 2SRI method discussed by Terza et al. (2008a). There are many applications of 2SRI in the literature including Shea et al. (2007), Stuart et al. (2009), DeSimone (2002), Baser et al. (2003), Norton and Van Houtven (2006), Gibson et al. (2006), Shin and Moon (2007), Lindrooth and Weisbrod (2007), Terza et al. (2008b), and Gavin et al. (2007). Our extended model is the first application of 2SRI to expenditure analysis in the FIML context. All other applications of 2SRI as discussed by Terza et al. (2008a) – i.e., those in Terza et al.

[6] Appendix A includes the formal derivation of the reparametrization of the conditional mean of the exponential.

[7] Computed using Maple®.

(2008a) and those that reference Terza et al. (2008a) – were cast and estimated in the NLS framework.

Following the 2SRI approach, we partition x in the following way

$$x = [x_o \quad x_e \quad x_u], \tag{2-10}$$

where x_o is the (1 $\times$ K) vector of observable exogenous regressors whose first element is constant, x_e is the (1 $\times$ S) vector of endogenous variables, and x_u is the (1 $\times$ S) vector of unobservables that are correlated with both x_e and y (unobservable confounding influences). We can now rewrite (2-5) as

$$E[y \mid x_o, x_e, x_u] = \exp(x_o\beta_o + x_e\beta_e + x_u\beta_u). \tag{2-11}$$

The unobservable confounders are the source of the endogeneity problem. To control for x_u, potentially nonlinear auxiliary equations are specified with the following form

$$x_{es} = r_s(w\alpha_s) + x_{us} \qquad \text{for } s = 1,\ldots,S, \tag{2-12}$$

where w = [x_o w^+] and w^+ denotes a 1 $\times$ S* vector of identifying instrumental variables and α is a (K+S*)$\times$1 vector of coefficient parameters.[8] In the first-stage of the 2SRI method, the appropriate estimator (e.g., NLS) is applied to (2-12) and the residuals from that regression are computed as

$$\hat{x}_{us} = x_{es} - r_s(w\hat{\alpha}_s) \qquad \text{for } s = 1,\ldots,S, \tag{2-13}$$

where $\hat{\alpha}_s$ denotes the estimated value of α_s. In the second stage, the predicted residuals (2-13) from the auxiliary equations are substituted for x_u in the regression model (2-11) and FIML (i.e., GG, as described above) is applied to obtain estimates of

[8] If the model is to be identified it must be true that S*≥S.

the elements of β. Terza et al. (2008a) show that this estimator is consistent. We analytically derive the details of this GGE estimator and program it along with its correct asymptotic inferential statistics in Stata/Mata 10®.

2.3 Simulation Analysis

Our objectives in this simulation analysis are threefold: First, we seek to verify the theoretical properties of the GGE estimator, such as statistical efficiency (variance) and unbiasedness. For the former, we compare the GGE estimator with 2SRI versions of two popular alternative methods – ordinary least squares (OLS) applied to a log-linear model and a generalized linear model with a log link and a gamma family (GLM-Gamma) – using the mean squared error of the marginal effect as the criterion. For the latter, we seek to study the marginal effect of endogenous policy variables using above mentioned estimators with average percentage absolute bias as our measurement. Secondly, we want to compare the performances of the above mentioned estimators, both in terms of efficiency and bias, with GG, OLS and GLM models, respectively, when they do not account for endogeneity. Finally, we test the nested model selection capability of the GGE model. For all the cases outlined above, we used simulated data from a variety of sampling designs nested within the GG framework.

The requisite data is generated using Monte Carlo simulations. The focus here is on strictly positive outcome data skewed to the right. There are five different data generation methods that are used in our study. Each of these data generation methods satisfies the exponential conditional mean property as specified in (2-5). These sampling designs include lognormal distributed data, gamma distributed data, Weibull distributed data, exponential distributed data and GG distributed data. The first four data generation techniques are chosen since they are specific versions of GG. We use

average percentage absolute bias as our metric for comparing the marginal effects in the various sampling designs.

2.3.1 Sampling Designs

2.3.1.1 The observable and unobservable confounders, the instrumental variables and the endogenous variable

As mentioned in Section 2.2.2, x_e is the (S x 1) vector of endogenous variables, x_o is the vector of observable exogenous regressors, and x_u is the (S x 1) vector of unobservable confounders (i.e., unobservable variables that are correlated with both x_e and y). For our simulations we generate a single observable regressor x_o, and a single unobservable confounder x_u, uniformly distributed over the [0, 2] and [0, 0.5] intervals, respectively. The instrumental variable w^+ is uniformly distributed over the [0, 2] interval. The endogenous variable x_e is defined as a linear function of x_o and w^+ with constant term equal to zero. Specifically, $r(\cdot)$ in (2-12) is the identity function with $w = [1 \quad x_o \quad w^+]$ and $\alpha = [0 \quad .3 \quad .3]$. The simulations are repeated 500 times for each of the seven sample sizes – 500, 1,000, 2,500, 5,000, 10,000, 100,000 and 500,000. Here we generate samples of increasing size in order to explore the small sample properties and the asymptotic properties of the estimators.

2.3.1.2 The outcome variable

The outcome variable y is generated using five different distributions. The scale parameter in GG is specified as $\mu = c + x_o\beta_o^* + x_e\beta_e^* + x_u\beta_u^*$ across all sampling designs for the outcome, where the constant c is chosen as 1.0. The coefficients of the endogenous and observable regressors are both equal to 0.5, whereas the coefficient of the unobservable regressor is equal to 1.0. In the following we describe the outcome data generator and define the values of the parameters of κ and σ, and the value of the

term C(σ, κ) that shifts the constant term in the conditional expectation of the outcome variable for each of the particular sampling designs.

Log-normal: The data generator we used for the log-normal outcome variable y is $y = \exp(c + x_o\beta_o^* + x_e\beta_e^* + x_u\beta_u^* + \varepsilon)$, where ε is standard normally distributed. Recall that the log-normal is the special case of the GG in which the parameter κ approaches zero and the parameter σ is permitted to vary freely. In generating the log-normal data we assume that σ = 1. It follows from the discussion in Section 2.2.1 that $C(1, \kappa \to 0) = \frac{1}{2}$.

Standard gamma: For the next case we generate standard gamma distributed data. The Stata® data generator command rgamma(a, b), that generates the gamma random variates, is used to generate the outcome variable y, where "a" and "b" are the gamma shape and scale parameters, respectively. The standard gamma distribution has a shape parameter $(1/|\kappa|^2)$ and a scale parameter $(\exp(\mu)|\kappa|^2)$. We chose the shape parameter $1/|\kappa|^2 = 2.0$ $(\kappa = \sigma \cong 0.71)$ in order to create skewed outcome data. The independent variables are generated according to the above descriptions with the same coefficients. The conditional mean of the outcome variable is defined according to equations (2-3) and (2-4), where $\kappa = \sigma$, and $C(\sigma,\sigma) = 0$ following the discussion in Section 2.2.1.

Weibull: In the third case we generate a Weibull distributed outcome variable. We set the distribution parameters so as to obtain right skewed outcome data for our simulation. We have used the Inverse Transform Method (Rubinstein, 1981) to create the Weibull distributed data. The random number generator for Weibull is $y = (-\ln(U))^{\sigma} \exp(\mu)$, where U is uniformly distributed over the interval [0, 1]. Here the

scale parameter is equal to $\exp(\mu)$, where μ is given by equation (2-2). We used two different values of the shape parameter ($\sigma = 0.5$, and $\sigma = 3.0$) in order to generate data with varying degrees of skewness. Accordingly, the conditional mean of y is $E(y \mid x) = \exp\left[x'\beta^* + C(\sigma, 1)\right]$, and the constant term is defined as $C(\sigma, 1) = \ln\left(\Gamma(1+\sigma)\right)$. For the value of the shape parameter $\sigma = 0.5$, $C(0.5, 1)$=-0.12; and for $\sigma = 3.0$, $C(3.0, 1)$=1.79. The independent variables are generated as described above.

Exponential: The fourth data generating process produces exponential distributed data. The rate parameter is defined as $(1/\exp(\mu))$. The exponential is a special case of the standard gamma in which both parameters are equal to unity. The Stata® data generator command rgamma(a, b) is used to generate the exponential outcome variable where the shape parameter $1/|\kappa|^2$ is equal to unity. The conditional mean of y is given by $E(y \mid x) = \exp[x'\beta^*]$, since the constant term – $C(1,1)$ – is zero when $\kappa = \sigma = 1$. We have the same endogenous, observable, unobservable and instrumental variables as defined above.

Generalized gamma: The last data generation method is the GG. In choosing the values for GG parameters we avoid the specific combinations of values corresponding to the distributions used in the other four sampling designs. We generate GG distributed random variables from a standard gamma using the following transformations, as demonstrated by Tadikamalla (1979):

$$y = v^{1/c} * b, \qquad (2\text{-}14)$$

where v is standard gamma distributed and y is the generalized gamma distributed random variable (i.e., the outcome variable in this case). The parameters "b" and "c" are defined as $\exp(\mu)/(1/|\kappa|^2)^{\sigma/\kappa}$ and κ/σ, respectively.

Given the considerable flexibility of the GG distribution we tried three different parameter settings in an attempt to explore the distributional landscape. We assign $1/|\kappa|^2 = 0.5$ ($\kappa \cong 1.41$), and use three different values for $\kappa/\sigma = 0.5, 2.0,$ and 5.0 ($\sigma \cong 2.83, 0.71,$ and 0.28, respectively) in our simulations. The conditional mean of y is given in (2-3), where $C(0.71, 1.41) = -0.23$, $C(2.83, 1.41) = 1.10$ and $C(0.28, 1.41) = -0.17$, respectively.

2.3.2 Estimators to be Evaluated and Compared

The GG estimator is a maximum likelihood estimator. It is assumed that the parameter μ is defined as $\mu = x\beta^*$ and the other two parameters κ and σ are estimated from the data. For the estimation of GG we use the streg[9] command in Stata®. The alternative estimators that are included in the comparisons are OLS applied to a log-linear model and GLM with a log link and a gamma family. These are described below.

Log-normal model: In this setting, the log-normal model is based on the assumption that $\ln(y) = x\beta^* + \varepsilon$. Next, the following two cases should be considered: In the first case, the distribution of (ε | x) can be assumed to be unspecified, $E[\varepsilon \mid x] = 0$, and $E[\exp(\varepsilon) \mid x] = \psi$, where ψ is an unspecified constant that does not depend on x. On

[9]streg command is used to estimate parametric survival models by maximum likelihood estimation. It can be used to estimate various distributions, such as Gompertz, logistic, Weibull, generalized gamma etc. In our implementation we used streg by assigning the number of failures equal to zero with the generalized gamma model.

the other hand in the second case, the distribution of (ε | x) can be assumed to be normally distributed $N(0,\sigma^2)$.

In the log-normal model, the logarithm of the dependent variable y is regressed on the independent variables x using OLS. The conditional mean is specified by

$$E[y \mid x] = \exp(x\beta^*)E[\exp(\varepsilon)|x]. \quad (2\text{-}15)$$

In the first case above, when the distribution of (ε | x) is assumed to be unspecified, we use the Duan smearing estimator in the estimation of the conditional mean of the outcome variable,

$$E\left[y \mid x\right] = \exp(x\hat{\beta}^*)\frac{\sum_{i=1}^{N}\exp(\hat{e}_i)}{n} \quad (2\text{-}16)$$

where $\hat{e}_i = \ln(y_i) - x_i\hat{\beta}^*$, to take care of the retransformation issue (Duan, 1983). In the second case, when the distribution of (ε | x) is assumed to be normally distributed, the conditional mean of the outcome variable can be estimated as

$$E\left[y \mid x\right] = \exp(x\hat{\beta}^* + 0.5\hat{\sigma}^2). \quad (2\text{-}17)$$

The log-normal model is easy to apply therefore it is frequently used in health economics and labor economics. An important drawback of this model is that it is relatively inefficient. Between the two cases for the distribution of (ε | x) presented above, the first method is less susceptible to misspecification because it does not require the specification of the distribution of (y | x). Hence, we would use this first case of unspecified distribution of (ε | x) in our experiment with the log-normal model.

Gamma generalized linear models (GLM-Gamma): In the GLM class of models (McCullagh and Nelder, 1989) (y | x) is assumed to follow a specified distribution, g(y |

x), and has a particular assumed form for its mean, E[y | x]. Estimation of the model is carried out by the FIML method. This approach offers substantial flexibility with respect to the choice of alternative specifications for g(y | x) and E[y | x]. Once these choices are made, however, the model is fixed and inflexible. We specify the distribution of (y | x) as the standard gamma and the link function, which relates the mean of the distribution function to the linear predictor, in our GLM model specification as $E[y \mid x] = \exp(x\beta)$.

2.3.3 Criteria for Evaluation and Comparison

In general, a researcher's ultimate objective is rarely only the estimation of the regression parameters, but rather the marginal or incremental effects on y of changes in some (or all) of the elements of x. There are a number of ways to characterize such effects (see Terza, 2010 for a detailed discussion). Here we focus on the expected marginal effect defined as

$$ME = E\left[\frac{\partial M(x\beta)}{\partial x_e}\right] \tag{2-18}$$

where $M(x\beta) = E[y \mid x]$. The marginal effect (2-18) is typically estimated as

$$ME = \sum_{i=1}^{n} \frac{1}{n} \frac{\partial M\left(x_i \hat{\beta}\right)}{\partial x_e} \tag{2-19}$$

where $\hat{\beta}$ denotes the estimated value of β.

For various model in our analysis the expected marginal effect (2-18) and the appropriate consistent estimator of (2-19) are:

Log-linear OLS: In the log-normal model with OLS estimation the marginal effect of interest is:

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_e}\right] = E\left[\frac{\partial \exp(x\beta)E[\exp(\varepsilon)|x]}{\partial x_e}\right] = \beta_e E[\exp(\varepsilon)|x]E[\exp(x\beta)] \,. \qquad (2\text{-}20)$$

In the first case of the log-normal model described above, where the distribution of (ε | x) assumed to be unspecified, $E[\exp(\varepsilon)|x]$ is a constant ψ and in the second case of log normal model, where (ε | x) is normally distributed, $E[\exp(\varepsilon)|x]$ is $0.5\sigma^2$. The consistent marginal effect estimator for the first case, where the distribution of (ε | x) assumed to be unspecified, then is

$$ME = \hat{\beta}_e \left(\frac{\sum_{i=1}^{n} \exp(\hat{e}_i)}{n}\right)\left(\frac{\sum_{i=1}^{n} \exp(x_i\hat{\beta})}{n}\right) \qquad (2\text{-}21)$$

where $\hat{\beta}$ is the classical OLS estimate of β and $\hat{e}_i = \ln(y_i) - x_i\hat{\beta}$. This is Duan's smearing estimator (Duan, 1983). The consistent marginal effect estimator for the second case, where (ε | x) is normally distributed, is

$$ME = \hat{\beta}_e \left(0.5\hat{\sigma}^2\right)\left(\frac{\sum_{i=1}^{n} \exp(x_i\hat{\beta})}{n}\right). \qquad (2\text{-}22)$$

GLM-Gamma: In the GLM-Gamma estimator the marginal effect of interest is

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_e}\right] = E\left[\frac{\partial \exp(x\beta)}{\partial x_e}\right] = \beta_e E[\exp(x\beta)] \qquad (2\text{-}23)$$

because the expected conditional mean of the outcome is assumed to be $E[y \mid x] = \exp(x\beta)$ in the GLM-Gamma model. The appropriate consistent marginal effect estimator of (2-23) is

$$ME = \hat{\beta}_e \left(\frac{\sum_{i=1}^{n} \exp(x_i \hat{\beta})}{n} \right) \tag{2-24}$$

where the $\hat{\beta}$ is the GLM estimate.

GGM: The marginal effect in the generalized gamma model is

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_e} \right] = E\left[\frac{\partial \exp(x\beta^* + C(\sigma,\kappa))}{\partial x_e} \right] = \beta_e^* E[\exp(x\beta^* + C(\sigma,\kappa))] \tag{2-25}$$

where $C(\sigma,\kappa) = \left(\frac{\sigma}{\kappa}\right)\ln(\kappa^2) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right)$ as given in equation (2-4).

We can write the consistent marginal effect estimator as

$$ME = \hat{\beta}_e^* \left(\frac{\sum_{i=1}^{n} \exp(x_i \hat{\beta}^* + \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\ln(\hat{\kappa}^2) + \ln\left(\Gamma\left\{\left(\frac{1}{\hat{\kappa}^2}\right) + \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\hat{\kappa}^2}\right\}\right))}{n} \right) \tag{2-26}$$

where $\hat{\beta}^*$, $\hat{\sigma}$ and $\hat{\kappa}$ are the GG parameter estimates.

The mean squared error of the marginal effect is used as a criterion for the statistical efficiency comparisons and it is measured as

$$MSE_{kd} = \left[\sum_{j=1}^{500} \frac{(ME_{jkd} - ME_d)^2}{500} \right] \tag{2-27}$$

where ME is defined as the true average marginal effect, the “k” indicates the sample size, the “j” indicates the jth repetition in the simulation (j = 1,…,500), and the “d” indicates the sampling design. To obtain the “true” value of ME for a particular sampling design, we simulated a large sample of 5 million observations and calculated the analog

to equation (2-19) for this super sample. To measure bias relative to the true value of marginal effect, we calculate the average percentage absolute bias as

$$APAB_{kd} = \left[\sum_{j=1}^{500} \frac{1}{500} \left(\frac{ABS(ME_{jkd} - ME_d)}{ABS(ME_d)} \right) \right] \times 100\% \quad (2\text{-}28)$$

where ME, “k”, “j”, and “d” defined as given above. Since our model is a non-linear model, we calculated the ME for the three quartile values of the endogenous regressor x_e. We are interested in both comparisons among the various estimation techniques and the differences between the estimates that are corrected for endogeneity versus those that are not.

2.3.4 Simulation Results

In the following set of results we would like to highlight the efficiency as well as the bias performance of the GGE model as compared to other popular models discussed above. The GG estimator is known to provide significant efficiency gains compared to log-linear models and GLM models (Manning et al., 2005). The objective here is to show that the GGE model maintains the statistical efficiency advantages of the conventional GG approach in the presence of endogeneity. To assess this, we focus on mean squared error (MSE) of the estimated marginal effect of the endogenous variable x_e as defined in (2-27). Furthermore, we also intent to show that the GGE model provides consistent and unbiased estimates across various sampling designs. This would be accomplished using average percentage absolute bias of the marginal effect as defined in (2-28).

The results of the simulation study with respect to MSE are presented in Table 2-1 and Table 2-2. Here the number of observations is 10,000 and 500, respectively, and

the number of repetitions for each simulation is 500. In these tables we are interested in comparing MSE among the various estimators defined in Section 2.3.2 for the endogeneity corrected models.[10] Since our model is a nonlinear model, the MSE is calculated at different quartiles of the endogenous variable x_e. To highlight the differences between OLS versus GGE, and GLM versus GGE, we have included relative efficiency gain percentage measures in Tables 2-1 and 2-2. These percentage measures indicate the relative efficiency of the GGE estimator compared to the OLS and the GLM estimators, and is defined by $\left((\text{MSE}_m - \text{MSE}_{GGE})/\text{MSE}_m\right)*100$, where m = OLS, GLM. In the log-normally distributed data, in Table 2-1 with sample size 10,000, OLS and GGE have similar MSE values and both are lower than the GLM results (GGE is 37% efficient relative to GLM). Similarly for the gamma distributed samples, the GLM and GGE estimators are equally efficient and both are more efficient than the OLS estimator (GGE is 16% efficient relative to OLS). For the case of the Weibull, exponential and generalized gamma distributed samples we see similar patterns, wherein the GGE estimator has lower or equal MSE compared to other estimators (the efficiency gain from GGE goes up to 60%). In Table 2-2, for the sample size 500, we observe that for the Weibull, the exponential and the GG distributed data, the GGE estimator provides efficiency gains compared to the OLS and the GLM estimators. The efficiency gain of the GGE estimator as compared to the OLS estimator ranges from 35% to 80%, and the efficiency gain of the GGE estimator as compared to the GLM estimator ranges from 4% to 50%. For the log normally distributed data, in this table, the

[10] For the models corrected for endogeneity, the values of the mean squared errors are significantly lower than the cases when they are not corrected (goes up to 1078.9250) (not included in Table 2-1 and 2-2). Endogeneity induces bias in the marginal effect of the endogenous variable, resulting in higher mean squared errors.

OLS estimator is slightly more efficient (1.8%) compared to the GGE estimator. Similarly for gamma distributed data the GLM estimator is slightly more efficient (0.45%) compared to the GGE estimator. We should be cautious in interpreting the results from the sample with 500 observations since the sample size is very small. Overall the findings of this experiment confirm that the GGE estimator provides efficiency gains relative to the OLS and GLM estimators in the presence of endogenous regressors.

Next, we shift our focus to the average marginal effect of the endogenous variables. Table 2-3 shows the average percent absolute bias of the estimated average marginal effect, as defined in (2-28), obtained from the various sampling designs for different number of observations, 10,000, 100,000, and 500,000, with 500 replications. We have reported both the results from the estimators corrected for endogeneity using 2SRI, and the results from the uncorrected versions of the estimators. The importance of correcting for endogeneity in the presence of an endogenous regressor is reinforced by the fact that the average percentage absolute bias ranges from 60% to 93% (for the values of x_e in the third quartile) when we do not correct for endogeneity. On the other hand, average percentage absolute bias is significantly lower when we do apply 2SRI to account for endogeneity.

Furthermore, the asymptotic properties of the included estimators predict that the average percentage absolute bias should tend to zero as the number of observations increases. This trend can be readily noted in Table 2-3 in the cases where we correct for endogeneity as the number of observations increases from 10,000 to 100,000 to 500,000. Whereas, when we do not correct for endogeneity, the average percentage absolute bias consistently remains high.

Next we focus only on the endogeneity corrected versions of the estimators, comparing their performance across the various sampling designs detailed in Section 2.3.1.2. In the case of log-normally distributed data, we expect the OLS estimator to provide low average percentage absolute bias, and this is confirmed by our findings in Table 2-3. The important point to note is that the performance of the OLS model is matched by our flexible GGE model, whereas GLM fares worse than both of these estimators.

Since we use the gamma distribution in our GLM model, we expect gamma and exponentially distributed data to provide low average percentage absolute bias when estimated with GLM and this is corroborated in our results in Table 2-3. The crucial point to note here is that the GGE model, due to its flexible nature, provides similar or slightly better results than the GLM estimator across the different sample sizes and different quartile values of x_e.

Next we look at the results obtained using the Weibull and the generalized gamma distributed data. For the case of Weibull distributed data, we used two different settings of the parameter σ in order to vary the amount of skewness in the data. Here we find that in the case with mild skewness, GGE performs slightly better than GLM and OLS for most of the sample sizes and quartile values. This advantage of the GGE model is magnified for the case in which the data is more skewed. We attribute this behavior of the GGE model to its flexible nature. For the generalized gamma distributed data, we chose three different settings of parameter σ in order to test our estimators. For $\sigma = 0.5$, OLS performs significantly worse than GGE and GLM, while GGE performs

noticeably better than GLM. For $\sigma = 2.0$ and $\sigma = 5.0$, the performance of GGE is still better than the other two estimators.

Table 2-4 presents the average percentage absolute bias results from small sample simulations, i.e., N = 500, 1,000, 2,500, and 5,000. Here, although the main findings are similar to the large sample results a word of caution is in order for the smallest sample sizes, 500, and 1,000, in specific sample designs where there is extreme skewness. For the Weibull distributed data with $\sigma = 3.0$, and the generalized gamma distributed data with $\kappa / \sigma = 0.5$, the average percentage absolute bias ranges from 216% to 70% for various estimators for the sample sizes 500 and 1,000. The models that are not corrected for endogeneity lead to less average percentage absolute bias in these specific sampling designs. Both OLS and GLM estimators, however, generate considerably higher average percentage absolute bias compared to GGE in the endogeneity corrected models. For the smallest sample sizes, where there is extreme skewness, none of the suggested estimators behave well.

Overall, it can be noted that due to its flexible nature, GGE consistently provides lower average percentage absolute bias across different data types, sampling sizes and quartiles, while the performance of OLS and GLM estimators vary depending on data types and parameter settings. Results here reinforce our expectation that the use of the proposed GGE model has advantages over the OLS and the GLM estimators when we encounter endogenous regressors with skewed outcomes.

Table 2-5 summarizes the results of the nested model selection tests from the GGE regressions in terms of the proportion of the replicated data sets of size 10,000 for which a particular null model (H_o:) is rejected at the 5% level of significance. In order to

demonstrate that the GGE model is useful for model selection from among the nested alternatives, these model selection tests should have two important features. First, they should manifest the same empirical size (i.e., likelihood of type I error) as the theoretical size of the test, in this case 5%. Second, the empirical power of the test [1-Pr(Type II error)] for a particular non-null sampling design should be extremely high – i.e., near 1. These two criteria appear to be supported by the results in Table 2-5. For example, for standard gamma generated data, the GGE estimator fails to reject the null hypotheses that $\kappa = \sigma$ in 94.5% of the replications, implying that the inherent data distribution is standard gamma. Similar correct predictions are also obtained for the Weibull and the log-normally distributed samples. For the GG distributed data, the null hypotheses $\kappa = \sigma$ (Standard gamma), $\kappa = 0$ (Log-normal), $\ln(\sigma) = 0$ and $\kappa = 1$ (Exponential) or $\kappa = 1$ (Weibull) are rejected in all repetitions (at the 95% confidence interval). Finally, in the exponential sampling design, the GGE model selection tests require further analysis since the exponential is a special case of both the standard gamma where the $\kappa = \sigma = 1$ and the Weibull distribution where $\kappa = 1$. For this reason, in ~95% of the replications, the GGE selection test fails to reject the null hypotheses that $\kappa = \sigma$ and $\kappa = 1$ respectively. These results show that GGE correctly predicts and selects the inherent data distribution when the regressors are endogenous.

2.4 Prescription Drug Use and Hospital Cost Offsets

One of the main characteristics of healthcare utilization/expenditure data is skewness. The population generally consists of healthy individuals who have very small or zero healthcare expenditures, or healthcare utilization. A very small percentage of the remaining unhealthy population incurs very large costs or needs excessive inpatient,

outpatient or emergency care. As a result, the expenditure distribution is typically skewed. Another noticeable characteristic of healthcare data which can lead to inconsistent estimation is endogeneity of one or more of the regressors (continuous, count, or dummy variable). This is commonly caused by the presence of unobservable confounding variables.

Stuart et al. (2009) examine the effect of outpatient prescription drug utilization on the cost of inpatient hospitalization for Medicare beneficiaries. In this paper they use a two-part model specification correcting for endogeneity by applying 2SRI (Terza et al., 2008a). We re-estimate this model using the newly developed GGE estimator in the second part of the two-part model, whereas they implemented NLS in the second part of the model. The GGE estimator introduces efficiency gains by taking full account of skewness and other higher-order moments of the hospital expenditure distribution.

The literature suggests that hospital expenditures make up the highest percentage of total personal health care spending in all age groups including elderly (65+) (refer Heffler et al., 2005). Heffler et al. (2005) indicates that, for the elderly, hospital spending as a percentage of total expenditure dropped from 43 percent in 1987 to 37 percent in 2004. In that same period, the health expenditures shifted towards other services such as prescription drug utilization, nursing homes etc. They also observe that prescription drug use has continued to rise since 1987. The observed concurrent increase in prescription drug use and decrease in hospital spending among the elderly leads one to suspect a possible causal relationship – increased prescription drug use might promote reduced inpatient costs. Stuart et al. (2009) analyzes this relationship in a two-part

regression framework that controls for confounding factors, e.g., unknown health status known only to the patient and the prescribing doctor.

Stuart et al. (2009) provide three possible explanations for an inverse relationship between prescription drug use and hospital costs. First, a high percentage of the prescription drug use by elderly Medicare beneficiaries prevents common chronic conditions. The preventive effect of prescription drug use would decrease hospitalization for such illnesses. Second, even if prescription drug utilization does not prevent hospitalization, it may lower the cost of hospitalization significantly by decreasing length-of-stay. Third, people with higher prescription drug utilization are more likely to adhere to the prescribed dosages which enhance medication effectiveness and reduce the likelihood of inpatient stays.

2.4.1 The Econometric Model

Here we implement the two-part model with 2SRI comprising the following three steps. In the first step (also the first stage of 2SRI), an auxiliary regression, as in equation (2-12), is estimated using NLS, where the dependent variable is prescription drug usage (a count variable measured as the number of prescription drug fills). The auxiliary regression function is defined as

$$x_e = \exp(w\alpha) + x_u \tag{2-29}$$

where x_e denotes prescription drug usage. In the second step (still the first part of the two-part model) we estimate the probability of hospital utilization using a conventional probit specification – the dependent variable (y*) is a binary variable taking the value one if hospital expenditure is positive (non-zero) and zero otherwise. We obtain

estimates of the parameters of the first part of the two-part model (β_1) by maximizing the following probit log-likelihood function with respect to $\beta_1 = [\beta_{o1} \quad \beta_{e1} \quad \beta_{u1}]$

$$\sum_{i=1}^{n}\left\{y^{*}\ln\Phi(x_{oi}\beta_{o1}+x_{ei}\beta_{e1}+\hat{x}_{ui}\beta_{u1})+(1-y^{*})\ln\left[1-\Phi(x_{oi}\beta_{o1}+x_{ei}\beta_{e1}+\hat{x}_{ui}\beta_{u1})\right]\right\} \qquad (2\text{-}30)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\hat{x}_{ui}$ is the residual of the NLS estimation from the first step in (2-29), i = 1, …, n, and n denotes the size of the full sample. In the third and final step, the new GGE method is used to estimate the parameters of the second part of the two part model – $\beta_2^{*} = [\beta_{o2}^{*} \quad \beta_{e2}^{*} \quad \beta_{u2}^{*}]'$, σ and κ. Specifically, we maximize the following log-likelihood function

$$\sum_{i=1}^{n_y}\ln\left[f(y_i;\kappa,\mu_i,\sigma)\right] \qquad (2\text{-}31)$$

with respect to σ, κ and β_2*, where $f(\cdot)$ is defined as in (2-1),

$$\mu_i = x_{oi}\beta_{o2}^{*}+x_{ei}\beta_{e2}^{*}+\hat{x}_{ui}\beta_{u2}^{*}, \qquad (2\text{-}32)$$

y denotes the hospital expenditures for the subsample of individuals who had at least one inpatient stay, i = 1, …, n_y, and n_y is the size of this subsample of hospitalized individuals. The conventional standard errors and t-statistics as an output by Stata® can not be used in a three-stage estimator, they must be corrected. The derivation of average marginal effect and the corrected standard errors are presented in Appendices B and C, respectively.

Following Mullahy (1998) and using the argument surrounding equations (2-3) through (2-5), it can be shown that the conditional mean of hospital expenditures (including the zeros and the positive values) is

$$E[y \mid x] = \Phi(x\beta_1)\exp(x\beta_2) \qquad (2\text{-}33)$$

where x is defined as in (2-10) and β_2 is the same as β_2^* except for the constant term, the first element of β_{o2} (i.e., β_{o21}) which becomes $\beta_{o21} = \beta^*_{o21} + C(\sigma,\kappa)$, with $C(\sigma, \kappa)$ defined as in (2-4) and β^*_{o21} being the first element (constant term) of β^*_{o2}. From (2-18) we get that the expected marginal effect as

$$ME = E\left[\beta_{e1}\varphi(x\beta_1)\exp(x\beta_2^*) + \beta^*_{e2}\Phi(x\beta_1)\exp(x\beta_2^*)\right] \quad (2\text{-}34)$$

where $\varphi(\cdot)$ is the standard normal probability density function. The appropriate estimator of (2-34) is

$$ME = \frac{1}{n}\sum_{i=1}^{N}\left\{\hat{\beta}_{e1}\varphi\left(x_{oi}\hat{\beta}_{o1} + x_{ei}\hat{\beta}_{e1} + \hat{x}_{ui}\hat{\beta}_{u1}\right)\exp\left(x_{oi}\beta^*_{o2} + x_{ei}\beta^*_{e2} + \hat{x}_{ui}\beta^*_{u2} + C(\sigma,\kappa)\right)\right.$$

$$\left.+\hat{\beta}^*_{e2}\Phi\left(x_{oi}\hat{\beta}_{o1} + x_{ei}\hat{\beta}_{e1} + \hat{x}_{ui}\hat{\beta}_{u1}\right)\exp\left(x_{oi}\beta^*_{o2} + x_{ei}\beta^*_{e2} + \hat{x}_{ui}\beta^*_{u2} + C(\sigma,\kappa)\right)\right\}. \quad (2\text{-}35)$$

2.4.2 Data Source and Variables

The data is from the 1999 and 2000 Medicare Current Beneficiary Survey (MCBS). It includes information on health status, health care use and expenditures, health insurance coverage, and socioeconomic and demographic characteristics of a nationally representative sample of Medicare beneficiaries. Stuart et al. (2009) created a subsample of Medicare beneficiaries who were enrolled for 24 months with continuous Medicare Part A and Part B coverage. Moreover each Medicare beneficiary has to have continuous drug coverage during the study period of 24 months or no coverage at all.[11] The overall data set includes 3,101 observations with 20 percent of the sample having positive hospital costs.

[11] For more information about the data set restrictions and properties refer to Stuart et al. (2009).

The dependent variable in the first part of the two-part model (step two of three in our estimation) is the probability of any hospital stay, and in the second part of the model (step three of three in our estimation) the dependent variable is total hospital expenditures which are obtained from Medicare Part A claims in the second year of the study period. Figures 2-1 and 2-2 show the histogram and Kernel density estimate of hospital expenditures for the overall data, including the zero expenditure, and the subsample of positive hospital expenditures, respectively. In both cases, hospital expenditures are highly positively skewed as expected. The independent variable of interest is the number of prescription drug fills in the second part of the study period. The other control variables are demographic variables such as age, sex, educational attainment, marital status, residence (urban vs. rural and the census region – northeast, midwest, south, and west), annual income; the DCG/HCC (Diagnostic Cost Group/Hierarchical Coexisting Conditions) risk adjuster[12], and Medicare entitlement status [aged with no prior disability, aged and previously disabled, Social Security Disability Insurance (SSDI) disabled]. In the first stage of the 2SRI estimation (step one of our estimation) we will include four instrumental variables to correct for the potential endogeneity of prescription drug usage.[13] The instruments correlate with the endogenous policy variable – prescription drug utilization – through the prescription drug coverage and do not correlate with the dependent variable. These four variables are as follows. (1) The percent of the work force in the respondent's state that is unionized. A unionized work force has a higher probability of having coverage and

[12] See Pope et al. (2004) for more information on DCG/HCC model.

[13] Stuart et al. (2009) utilizes prescription drug coverage as the instrumental variable in their analysis. We incorporate other instrumental variables in our study and replicate their analysis using the new instrumental variables.

indirectly more prescription drug usage. (2) The average premium for Medigap Plan H, I, and J in the state. Coverage for prescription drugs can be higher in those states with lower average premium. (3) A variable indicating if the state has a pharmaceutical assistance plan for low income elders/or disabled Medicare beneficiaries. (4) The state per capita income. The wealthy states can be associated with greater Medicare supplemental policies, including drug coverage. Table 2-6 summarizes the descriptive statistics for these variables.

2.4.3 Estimation Results

Table 2-7 presents the estimated average marginal effect as given in equation (2-35) from the two-part model (with various estimators in the second part).[14] Results for both the endogeneity corrected models using 2SRI and the uncorrected models are included. The first row shows the replicated results of Stuart et al. (2009) using the new instrumental variables where we find that the average marginal effect of prescription drug fill is a statistically significant -$140.48, i.e., a one unit increase in prescription drug fills leads to a $140.48 decrease in hospital expenditures. Here NLS is used to estimate the second part of the model. The next three rows show that the results from using OLS for a semi-log model, GLM with a gamma function and our GGE[15] model in the second part are -$87.19, -$109.32 and-$89.30 respectively.

All of the endogeneity corrected average marginal effect estimates are negatively signed. This is consistent with our expectation that appropriate usage of prescription

[14] We do not include the estimated coefficients from the three stages since we are primarily interested in the average marginal effect.

[15] Note that although our sample size is small, N = 600, we are not dealing with one of the extreme skewness cases mentioned in the simulation analysis. The nested model selection tests suggest that our data is log-normally distributed and the log-normal sampling design is not one of the severely skewed distributions.

drugs will offset inpatient hospital costs. For all the estimation procedures that we implemented, failure to correct for endogeneity leads to results that are substantially divergent from the 2SRI estimates. The average marginal effect in all four uncorrected estimations is positive and ranges from $15.69 to $17.78 – contrary to expectation. Secondly, the Wald test results obtained from our GGE model fails to reject that the dependent variable is log-normally distributed. The result of this model specification test is corroborated by the estimated average marginal effects from OLS and GGE, which are -$87.29 and -$89.30 respectively. Thirdly, the GGE estimator along with the OLS has the lowest standard errors for the estimated marginal effect.[16] The precision is born out again in the actual data analysis. Finally, we must point out that although the estimated effects have the expected negative sign in both, NLS and GGE, there is a significant difference (approximately $50) between the Stuart et al. (2009) model and our model.

2.5 Summary, Discussion and Conclusion

Health utilization and expenditures data is known to suffer from skewness and endogeneity. To address skewness, a FIML estimator based on a flexible distributional form (the GG) that offers increased efficiency relative to alternative estimators was suggested by Manning et al. (2005). In this chapter we have proposed a model that retains the precision advantages of the GG estimator but also includes a 2SRI component to account for endogeneity (the GGE estimator).

In order to test our model, we conducted extensive simulation analysis using endogenous regressors and outcomes generated by various distributions. We

[16] See Appendix C for detailed derivation of the asymptotic standard errors of the marginal effect.

compared the results obtained by GGE to other commonly used models – OLS and GLM with Gamma. Our findings suggest that GGE consistently has low average percentage absolute bias across different data distributions and parameter configurations. On the other hand, OLS and GLM provide low average percentage absolute bias only for certain specific data distributions and parameter settings. This versatility of GGE is a testament to its flexible nature and its ability to correct for endogeneity using the 2SRI model. We confirmed that the statistical efficiency inherent in the GG estimator transforms to the GGE model both in our simulation analysis and the real data analysis. In our simulation analysis, we were also able to verify the asymptotic properties of the GGE estimator. The distribution identification tests in our simulations showed that the GGE model was able to identify inherent nested models with high accuracy.

Finally, the application of the GGE model to health expenditures data (Stuart et al., 2009) predicted that an increase in the prescription drug fill by one unit leads to a $89.30 decrease in hospital expenditures. This result is similar to the result obtained using the OLS model on this data. Note that the Wald test result using the GGE model fails to reject that the dependent variable is log-normally distributed, which makes above two findings consistent with each other. The marginal effect predicted by the GGE model is 35% lower than that predicted by the previously published NLS model (Stuart et al., 2009).

Table 2-1. For sample size 10,000, mean squared error of the marginal effect with percent relative efficiency gain

Data	Estimator	1st Quartile of x_e 2SRI	2nd Quartile of x_e 2SRI	3rd Quartile of x_e 2SRI
Log-Normal, $\sigma_\varepsilon^2 = 1$	OLS for ln(y)	0.5815 (0.382%)	0.8669 (0.348%)	1.2692 (0.320%)
	GLM-Gamma	0.9252 (37.389%)	1.3790 (37.355%)	2.0215 (37.419%)
	GGM	0.5793	0.8639	1.2651
Gamma, $\gamma = 2.0$	OLS for ln(y)	0.1201 (16.134%)	0.1803 (15.989%)	0.2653 (15.890%)
	GLM-Gamma	0.1007 (0.002%)	0.1514 (0.005%)	0.2231 (0.008%)
	GGM	0.1007	0.1514	0.2231
Weibull, $\sigma = 0.5$	OLS for ln(y)	0.0644 (35.553%)	0.0972 (35.015%)	0.1436 (34.551%)
	GLM-Gamma	0.0448 (7.204%)	0.0675 (6.417%)	0.0996 (5.687%)
	GGM	0.0415	0.0632	0.0940
Weibull, $\sigma = 3.0$	OLS for ln(y)	99.2321 (39.006%)	149.5710 (39.736%)	226.7172 (40.964%)
	GLM-Gamma	125.8915 (51.923%)	184.9323 (51.259%)	275.1588 (51.357%)
	GGM	60.5254	90.1380	133.8454
Exponential	OLS for ln(y)	0.3075 (38.448%)	0.4640 (38.726%)	0.6878 (39.055%)
	GLM-Gamma	0.1900 (0.388%)	0.2851 (0.260%)	0.4198 (0.168%)
	GGM	0.1893	0.2843	0.4191
Generalized Gamma, $\kappa/\sigma = 0.5$, $\gamma = 0.5$	OLS for ln(y)	34.8443 (60.072%)	50.0370 (58.997%)	72.8109 (58.668%)
	GLM-Gamma	18.4038 (24.404%)	27.1557 (24.447%)	40.0981 (24.949%)
	GGM	13.9125	20.5169	30.0939
Generalized Gamma, $\kappa/\sigma = 2.0$, $\gamma = 0.5$	OLS for ln(y)	0.1490 (57.411%)	0.2221 (57.268%)	0.3252 (57.182%)
	GLM-Gamma	0.0700 (9.292%)	0.1047 (9.297%)	0.1535 (9.302%)
	GGM	0.0635	0.0949	0.1393
Generalized Gamma, $\kappa/\sigma = 5.0$, $\gamma = 0.5$	OLS for ln(y)	0.0304 (51.370%)	0.0459 (50.942%)	0.0677 (50.640%)
	GLM-Gamma	0.0203 (27.383%)	0.0309 (27.104%)	0.0458 (26.914%)
	GGM	0.0148	0.0225	0.0334

Note: The values given in parentheses are the percent relative efficiency gains that measure the relative efficiency of GGE estimator compared to the OLS and GLM estimators. The percent relative efficiency gain is defined by $\left(\left(MSE_m - MSE_{GGE}\right)/MSE_m\right)*100$, where m = OLS, GLM.

Table 2-2. For sample size 500, mean squared error of the marginal effect with percent relative efficiency gain

Data	Estimator	1st Quartile of x_e 2SRI	2nd Quartile of x_e 2SRI	3rd Quartile of x_e 2SRI
Log-Normal, $\sigma_\varepsilon^2 = 1$	OLS for ln(y)	11.0042 (-1.883%)	15.7595 (-1.778%)	22.0935 (-1.646%)
	GLM-Gamma	17.9952 (37.698%)	25.7775 (37.776%)	36.7286 (38.857%)
	GGM	11.2114	16.0397	22.4570
Gamma, $\gamma = 2.0$	OLS for ln(y)	2.7292 (27.163%)	3.8404 (25.650%)	5.2698 (24.582%)
	GLM-Gamma	1.9789 (-0.452%)	2.8384 (-0.596%)	3.9472 (-0.689%)
	GGM	1.9879	2.8553	3.9744
Weibull, $\sigma = 0.5$	OLS for ln(y)	1.3322 (38.303%)	1.8927 (37.026%)	2.6193 (35.841%)
	GLM-Gamma	0.9128 (9.954%)	1.3144 (9.317%)	1.8417 (8.751%)
	GGM	0.8219	1.1919	1.6805
Weibull, $\sigma = 3.0$	OLS for ln(y)	4889.0390 (63.492%)	4267.7430 (50.076%)	5798.5690 (44.484%)
	GLM-Gamma	3590.7890 (50.293%)	3822.8940 (44.267%)	6430.4850 (49.939%)
	GGM	1784.8660	2130.6130	3219.1520
Exponential	OLS for ln(y)	7.1370 (41.557%)	9.6358 (38.669%)	13.0916 (37.060%)
	GLM-Gamma	4.1719 (0.019%)	5.9140 (0.073%)	8.2485 (0.104%)
	GGM	4.1711	5.9097	8.2399
Generalized Gamma, $\kappa/\sigma = 0.5$, $\gamma = 0.5$	OLS for ln(y)	2103.2300 (80.647%)	1568.9410 (70.760%)	2027.5790 (67.791%)
	GLM-Gamma	572.4497 (28.896%)	654.1119 (29.866%)	1014.3330 (35.616%)
	GGM	407.0346	458.7557	653.0694
Generalized Gamma, $\kappa/\sigma = 2.0$, $\gamma = 0.5$	OLS for ln(y)	3.1888 (58.239%)	4.3580 (56.109%)	5.9124 (54.804%)
	GLM-Gamma	1.4409 (7.578%)	2.0386 (6.174%)	2.8122 (4.980%)
	GGM	1.3317	1.9128	2.6722
Generalized Gamma, $\kappa/\sigma = 5.0$, $\gamma = 0.5$	OLS for ln(y)	0.5603 (54.001%)	0.8076 (53.064%)	1.1379 (51.400%)
	GLM-Gamma	0.3577 (27.933%)	0.5199 (27.088%)	0.7444 (25.709%)
	GGM	0.2577	0.3790	0.5530

Table 2-3. For the large samples, average percentage absolute bias of the marginal effect

			1st Quartile of x_e		2nd Quartile of x_e		3rd Quartile of x_e	
Data	Estimator	N	(1) 2SRI (%)	(2) No IV (%)	(3) 2SRI (%)	(4) No IV (%)	(5) 2SRI (%)	(6) No IV (%)
Log-Normal, $\sigma_\varepsilon^2 = 1$	OLS for ln(y)	10,000	8.740	61.770	9.681	74.968	10.617	89.298
		100,000	2.436	63.549	2.714	77.361	2.992	92.330
		500,000	1.152	63.696	1.283	77.669	1.415	92.861
	GLM-Gamma	10,000	10.960	61.765	12.137	75.021	13.312	89.429
		100,000	3.345	63.651	3.728	77.505	4.112	92.523
		500,000	1.440	63.749	1.604	77.741	1.769	92.954
	GGM	10,000	8.716	61.783	9.653	74.983	10.587	89.315
		100,000	2.432	63.530	2.709	77.331	2.988	92.285
		500,000	1.218	63.693	1.350	77.641	1.482	92.805
Gamma, $\gamma = 2.0$	OLS for ln(y)	10,000	6.589	62.859	7.319	76.341	8.044	90.986
		100,000	2.153	63.621	2.400	77.449	2.648	92.434
		500,000	0.870	63.656	0.969	77.613	1.068	92.786
	GLM-Gamma	10,000	6.035	62.839	6.709	76.304	7.378	90.928
		100,000	1.837	63.619	2.043	77.445	2.250	92.427
		500,000	0.775	63.626	0.865	77.573	0.952	92.734
	GGM	10,000	6.029	62.846	6.702	76.313	7.371	90.940
		100,000	1.837	63.619	2.043	77.444	2.250	92.426
		500,000	0.776	63.628	0.865	77.575	0.953	92.737
Weibull, $\sigma = 0.5$	OLS for ln(y)	10,000	5.459	62.953	6.074	76.455	6.685	91.120
		100,000	1.580	63.451	1.759	77.226	1.940	92.151
		500,000	0.767	63.645	0.851	77.606	0.934	92.783
	GLM-Gamma	10,000	4.568	62.949	5.084	76.437	5.597	91.085
		100,000	1.334	63.460	1.486	77.239	1.640	92.167
		500,000	0.638	63.634	0.707	77.591	0.775	92.765
	GGM	10,000	4.342	62.812	4.844	76.239	5.341	90.817
		100,000	2.232	63.425	2.459	77.194	2.686	92.112
		500,000	0.567	63.070	0.632	76.871	0.697	91.870
Weibull, $\sigma = 3.0$	OLS for ln(y)	10,000	30.858	62.866	34.264	77.425	37.835	93.543
		100,000	9.386	63.180	10.424	76.925	11.463	91.835
		500,000	4.519	63.669	5.018	77.634	5.518	92.820
	GLM-Gamma	10,000	35.014	64.052	38.741	79.312	42.671	96.289
		100,000	11.095	63.395	12.355	77.317	13.619	92.440
		500,000	4.803	63.308	5.332	77.214	5.862	92.335
	GGM	10,000	24.100	63.464	26.700	77.467	29.373	92.930
		100,000	7.590	63.139	8.450	76.889	9.311	91.799
		500,000	3.374	63.927	3.755	77.697	4.138	92.757

Table 2-3. Continued

Data	Estimator	N	1st Quartile of x_e (1) 2SRI (%)	(2) No IV (%)	2nd Quartile of x_e (3) 2SRI (%)	(4) No IV (%)	3rd Quartile of x_e (5) 2SRI (%)	(6) No IV (%)
Exponential	OLS for ln(y)	10,000	10.445	62.924	11.608	76.527	12.775	91.329
		100,000	3.098	63.374	3.450	77.131	3.802	92.037
		500,000	1.486	63.637	1.653	77.596	1.820	92.771
	GLM-Gamma	10,000	8.162	62.997	9.056	76.571	9.949	91.329
		100,000	2.569	63.409	2.864	77.179	3.158	92.099
		500,000	1.172	63.615	1.300	77.571	1.428	92.743
	GGM	10,000	8.152	62.934	9.051	76.494	9.949	91.236
		100,000	2.523	63.403	2.819	77.151	3.115	92.047
		500,000	1.121	63.622	1.250	77.557	1.378	92.707
Generalized Gamma, $\kappa/\sigma = 0.5$, $\gamma = 0.5$	OLS for ln(y)	10,000	36.967	61.470	40.589	75.822	44.394	91.744
		100,000	11.317	62.711	12.582	76.431	13.851	91.325
		500,000	4.974	63.884	5.550	77.957	6.127	93.271
	GLM-Gamma	10,000	26.417	62.903	29.227	77.258	32.124	93.075
		100,000	8.513	62.969	9.471	76.703	10.429	91.599
		500,000	3.750	63.601	4.176	77.565	4.601	92.750
	GGM	10,000	23.697	63.021	26.231	77.177	28.817	92.714
		100,000	7.321	62.958	8.141	76.693	8.961	91.586
		500,000	3.227	63.571	3.596	77.534	3.965	92.717
Generalized Gamma, $\kappa/\sigma = 2.0$, $\gamma = 0.5$	OLS for ln(y)	10,000	9.151	62.604	10.136	76.044	11.118	90.650
		100,000	2.834	63.364	3.161	77.122	3.487	92.028
		500,000	1.252	63.739	1.396	77.729	1.541	92.940
	GLM-Gamma	10,000	6.359	62.909	7.053	76.402	7.744	91.059
		100,000	1.906	63.429	2.126	77.203	2.346	92.128
		500,000	0.870	63.675	0.969	77.642	1.068	92.826
	GGM	10,000	6.038	62.882	6.695	76.373	7.349	91.028
		100,000	1.814	63.356	2.024	77.117	2.235	92.026
		500,000	0.822	63.581	0.916	77.526	1.010	92.686
Generalized Gamma, $\kappa/\sigma = 5.0$, $\gamma = 0.5$	OLS for ln(y)	10,000	3.915	62.823	4.363	76.253	4.806	90.832
		100,000	1.142	63.482	1.278	77.263	1.414	92.195
		500,000	0.511	63.705	0.570	77.681	0.628	92.875
	GLM-Gamma	10,000	3.255	62.847	3.636	76.279	4.012	90.861
		100,000	0.897	63.494	1.003	77.280	1.110	92.216
		500,000	0.414	63.678	0.462	77.645	0.508	92.828
	GGM	10,000	2.845	62.677	3.175	76.084	3.501	90.638
		100,000	0.825	63.266	0.919	76.998	1.014	91.873
		500,000	0.347	63.394	0.387	77.286	0.425	92.386

Table 2-4. For the small samples, average percentage absolute bias of the marginal effect

			1st Quartile of x_e		2nd Quartile of x_e		3rd Quartile of x_e	
Data	Estimator	N	(1) 2SRI (%)	(2) No IV (%)	(3) 2SRI (%)	(4) No IV (%)	(5) 2SRI (%)	(6) No IV (%)
Log-Normal, $\sigma_\varepsilon^2 = 1$	OLS for ln(y)	500	37.783	60.643	41.321	74.114	44.282	85.732
		1,000	23.861	61.503	26.876	76.014	29.720	90.594
		2,500	17.114	61.793	18.815	76.099	20.495	91.122
		5,000	11.747	62.492	13.021	75.792	14.321	90.576
	GLM-Gamma	500	48.551	61.641	53.220	75.731	57.289	88.153
		1,000	31.118	60.692	34.989	75.492	38.705	90.500
		2,500	21.925	61.706	24.104	76.209	26.258	91.496
		5,000	14.927	61.965	16.539	75.197	18.185	89.921
	GGM	500	38.209	61.062	41.779	74.572	44.748	86.232
		1,000	23.961	61.767	26.987	76.308	29.845	90.918
		2,500	17.101	61.891	18.799	76.196	20.471	39.645
		5,000	11.784	62.521	13.060	75.820	14.364	90.603
Gamma, $\gamma = 2.0$	OLS for ln(y)	500	31.335	61.277	34.138	74.605	36.476	85.946
		1,000	20.046	61.818	22.745	76.216	25.238	90.620
		2,500	14.100	62.537	15.576	76.942	17.027	92.049
		5,000	9.168	63.200	10.170	76.624	11.195	91.538
	GLM-Gamma	500	26.375	61.569	28.854	74.685	30.913	85.745
		1,000	18.156	62.080	20.651	76.485	22.929	90.878
		2,500	12.567	62.640	13.852	77.036	15.129	92.122
		5,000	8.253	63.275	9.166	76.703	10.104	91.616
	GGM	500	26.430	61.703	28.925	74.869	30.995	85.980
		1,000	18.118	62.017	20.606	76.397	22.880	90.761
		2,500	12.543	62.883	13.823	77.432	15.096	92.723
		5,000	8.274	63.277	9.190	76.706	10.130	91.621
Weibull, $\sigma = 0.5$	OLS for ln(y)	500	24.107	61.406	26.297	74.154	28.197	84.819
		1,000	16.515	62.656	18.757	77.016	20.783	91.339
		2,500	11.157	63.113	12.332	77.625	13.508	92.840
		5,000	7.198	63.381	8.002	76.816	8.821	91.736
	GLM-Gamma	500	20.139	61.587	22.042	74.181	23.752	84.654
		1,000	14.027	62.683	15.953	76.940	17.641	91.127
		2,500	9.296	62.991	10.250	77.414	11.212	92.518
		5,000	5.883	63.200	6.532	76.556	7.190	91.380
	GGM	500	19.398	61.544	21.265	74.104	22.960	84.541
		1,000	13.391	62.501	15.231	76.685	16.836	90.790
		2,500	8.951	62.898	9.867	77.285	10.790	92.347
		5,000	5.890	62.866	6.532	76.099	7.183	90.779

Table 2-4. Continued

Data	Estimator	N	1st Quartile of x_e (1) 2SRI (%)	1st Quartile of x_e (2) No IV (%)	2nd Quartile of x_e (3) 2SRI (%)	2nd Quartile of x_e (4) No IV (%)	3rd Quartile of x_e (5) 2SRI (%)	3rd Quartile of x_e (6) No IV (%)
Weibull, $\sigma = 3.0$	OLS for ln(y)	500	174.135	116.031	172.859	142.537	182.334	172.394
		1,000	101.121	86.337	111.586	110.611	125.538	138.336
		2,500	66.614	67.601	72.859	86.281	80.146	107.268
		5,000	43.326	63.716	48.000	79.191	53.135	96.915
	GLM-Gamma	500	160.658	111.766	164.120	140.805	177.166	174.705
		1,000	105.551	85.606	113.430	109.821	125.954	138.026
		2,500	74.288	67.939	79.617	85.865	86.070	106.019
		5,000	52.077	61.773	56.457	76.897	61.483	94.298
	GGM	500	121.808	91.341	127.821	114.549	137.313	138.841
		1,000	75.819	72.453	84.836	93.184	95.300	116.018
		2,500	52.713	63.002	57.271	79.360	62.134	97.268
		5,000	35.294	61.407	38.010	75.648	41.333	91.767
Exponential	OLS for ln(y)	500	48.557	63.166	52.357	77.666	55.660	90.502
		1,000	31.862	61.911	36.088	77.058	40.148	92.453
		2,500	21.732	62.660	24.061	77.486	26.359	93.149
		5,000	14.365	63.348	15.970	76.982	17.617	92.171
	GLM-Gamma	500	38.354	61.325	41.779	74.930	44.726	86.672
		1,000	24.855	62.285	28.249	77.048	31.438	91.914
		2,500	17.303	62.247	19.065	76.715	20.788	91.923
		5,000	11.341	62.853	12.565	76.238	13.815	91.117
	GGM	500	38.279	61.293	41.694	74.868	44.640	86.576
		1,000	24.960	62.292	28.359	77.064	31.565	91.938
		2,500	17.311	62.162	19.095	76.611	20.844	91.800
		5,000	11.419	62.840	12.691	76.223	13.998	91.100
Generalized Gamma, $\kappa / \sigma = 0.5$, $\gamma = 0.5$	OLS for ln(y)	500	216.083	126.006	206.200	152.897	214.710	185.599
		1,000	114.183	88.418	124.841	114.146	141.059	144.522
		2,500	77.942	65.292	84.611	83.302	92.881	103.588
		5,000	49.045	62.777	53.992	78.469	59.547	96.566
	GLM-Gamma	500	134.372	89.954	140.609	110.875	152.430	132.945
		1,000	82.625	72.716	92.260	94.069	103.881	117.789
		2,500	54.857	61.421	59.606	77.536	64.760	95.174
		5,000	37.119	63.859	41.183	78.916	45.563	95.998
	GGM	500	112.709	81.574	118.408	100.969	127.005	120.607
		1,000	70.589	67.607	78.887	86.512	88.167	106.962
		2,500	49.639	60.133	54.220	75.562	59.091	92.290
		5,000	31.170	63.034	34.561	77.604	38.156	94.064

Table 2-4. Continued

Data	Estimator	N	1st Quartile of x_e (1) 2SRI (%)	(2) No IV (%)	2nd Quartile of x_e (3) 2SRI (%)	(4) No IV (%)	3rd Quartile of x_e (5) 2SRI (%)	(6) No IV (%)
Generalized Gamma, $\kappa/\sigma = 5.0$, $\gamma = 0.5$	OLS for ln(y)	500	16.717	61.011	18.310	73.337	19.783	83.523
		1,000	11.655	62.968	13.277	77.264	14.649	91.478
		2,500	8.186	62.903	9.022	77.244	9.884	92.248
		5,000	4.810	63.191	5.353	76.517	5.900	91.304
	GLM-Gamma	500	13.253	61.327	14.546	73.631	15.831	83.771
		1,000	9.681	62.794	11.028	76.972	12.116	91.045
		2,500	6.794	63.020	7.464	77.381	8.174	92.403
		5,000	3.774	63.202	4.202	76.520	4.632	91.294
	GGM	500	11.506	61.256	12.689	73.496	13.965	83.565
		1,000	8.478	62.151	9.666	76.110	10.590	89.943
		2,500	5.961	63.020	6.519	77.410	7.134	92.466
		5,000	3.103	62.878	3.454	76.107	3.808	90.776
Generalized Gamma, $\kappa/\sigma = 2.0$, $\gamma = 0.5$	OLS for ln(y)	500	41.263	59.452	44.599	72.748	47.428	84.260
		1,000	26.344	62.653	29.923	77.651	33.307	92.795
		2,500	19.020	62.051	21.053	76.433	23.058	91.547
		5,000	11.998	62.917	13.329	76.323	14.691	91.236
	GLM-Gamma	500	28.237	59.696	30.783	72.272	32.943	82.829
		1,000	18.178	62.454	20.633	76.871	22.872	91.270
		2,500	13.341	62.467	14.729	76.804	16.105	91.828
		5,000	7.976	63.299	8.871	76.723	9.782	91.635
	GGM	500	27.108	59.582	29.663	72.060	31.868	82.511
		1,000	17.803	62.332	20.216	76.690	22.429	91.019
		2,500	12.620	62.433	13.923	76.755	15.220	91.758
		5,000	7.732	63.312	8.593	76.741	9.471	91.657

Table 2-5. Nested model selection tests from the GGE estimator for sample size 10,000

Data	Proportion Significant at 5% Gamma, $\sigma = \kappa$	Lognormal, $\kappa = 0$	Weibull, $\kappa = 1$	Exponential, $\ln(\sigma) = 0$, $\kappa = 1$
Log-Normal, $\sigma_\varepsilon^2 = 1$	1.0	0.056	1.0	1.0
Gamma, $\gamma = 2.0$	0.055	1.0	1.0	1.0
Weibull, $\sigma = 0.5$	1.0	1.0	0.034	1.0
Exponential	0.05	1.0	0.058	0.05
Generalized Gamma, $\kappa / \sigma = 2.0$, $\gamma = 0.5$	1.0	1.0	1.0	1.0

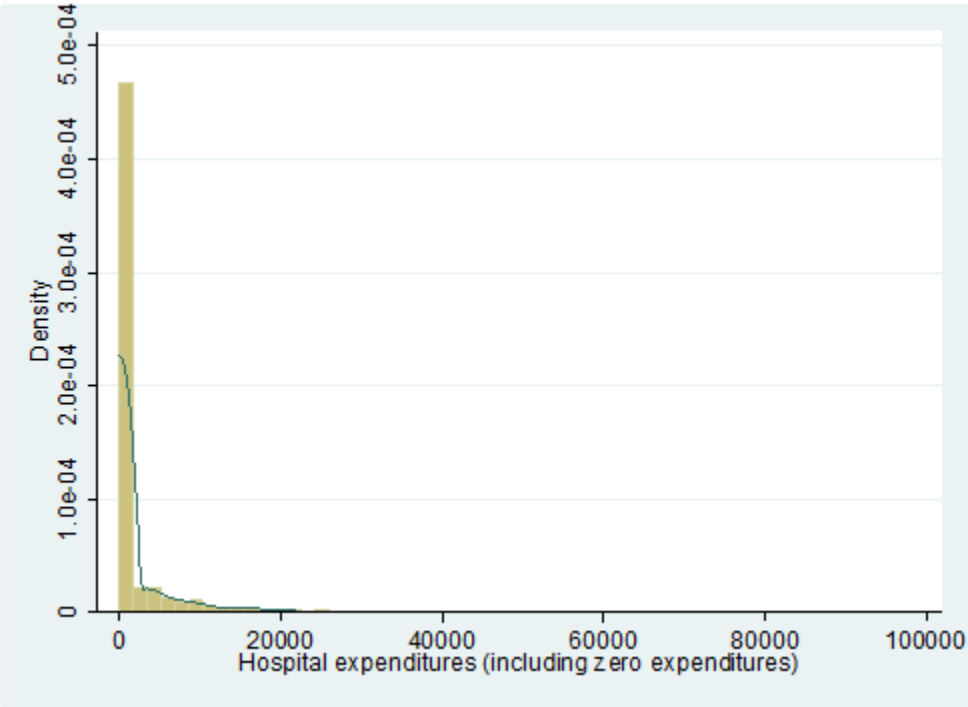

Figure 2-1. Histogram and Kernel density estimate of hospital expenditures (Overall)

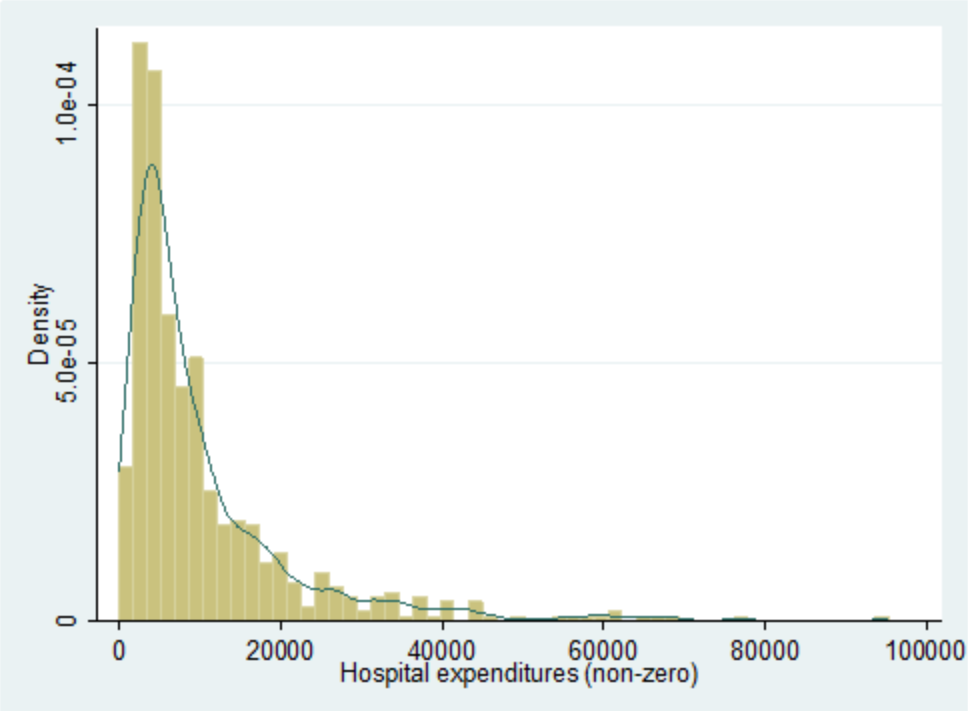

Figure 2-2. Histogram and Kernel density estimate of hospital expenditures (Non-zero)

Table 2-6. Descriptive statistics of the study sample – prescription drug use and hospital cost offsets

Variables	Hospitalized (N = 624)	Overall Sample (N = 3,101)
Dependent variables		
Hospitalized	NA	20.1%
Hospital expenditures	$10,425 (11,627)	$2,098 (6,682)[†]
Independent variables		
Number of prescription fills	40 (31)	30 (27)
Medicare entitlement status		
Aged/no prior disability	72.8%	76.7%
SSDI disabled (<65)	15.7%	16.2%
Aged/previously disabled (>65)	11.5%	7.1%
Age		
<65	18.7%	20.0%
65-69	9.3%	8.5%
70-74	19.4%	24.3%
75-79	18.9%	18.8%
80+	33.7%	28.4%
Female	55.3%	57.1%
Marital status		
Married	43.1%	48.4%
Single/widowed/divorced/separated	56.9%	51.6%
Educational attainment		
<High school graduate	37.0%	35.7%
High school graduate	63.0%	64.3%
Urban residence	67.9%	66.1%
Census region		
Northeast	20.5%	20.6%
Midwest	22.3%	21.3%
South	41.2%	40.5%
West	16.0%	17.6%
Annual income		
≤ $10,000	32.5%	26.9%
$10,001 - $20,000	25.5%	26.3%
$20,001 - $30,000	18.0%	19.9%
> $30,000	24.0%	26.9%
DCG/HCC risk adjuster		
Ybase	$8,416 (6,512)	$5,925 (4,911)
Instrumental variables		
% State workforce unionized	13.8%	14.0%
Mean annual state-level Medigap premium (H,I,J plans)	$2,108 (400)	$2,089 (417)
State has pharmaceutical assistance plans	31.9%	32.1%
State per capita income	$28,230 (3,653)	$28,428 (3,659)

[†]The values in the parenthesis are standard deviations.

Table 2-7. The estimation results of the real data analysis – prescription drug use and hospital cost offsets

	Average Marginal Effect			
Type of Estimation	2SRI	S.E.	No-correction	S.E.
NLS	-140.48	8.323	15.69	0.831
OLS for ln(y)	-87.29	4.057	17.78	1.846
GLM (Gamma)	-109.32	5.537	16.34	1.681
GG	-89.30	4.538	16.32	1.774

CHAPTER 3
THE GENERALIZED GAMMA ESTIMATOR WITH A FLEXIBLE-FORM CONDITIONAL MEAN REGRESSION SPECIFICATION

3.1 Introduction and Background

The most widely used models designed to take account of skewed data in health economics and health services research assume that the conditional mean is an exponential function of the regressors, e.g., generalized linear models (GLM) and log-linear models estimated via OLS. A misspecification of the conditional mean (e.g., assuming it has an exponential form when it does not) can lead to biased estimates of targeted marginal effects. It has been shown that the introduction of a more flexible conditional mean function can serve to alleviate such problems. This model, which incorporates the inverse of the classical Box-Cox transformation (Box and Cox, 1964), was first used in a regression context by Wooldridge (1992). This model encompasses exponential, linear and the power specifications as special cases. More recently, Kenkel and Terza (2001), Terza et al. (2008a) and Terza et al. (2008b) used a variant of the inverse Box-Cox (IBC) transformation with nonlinear least squares (NLS). Since NLS is not a full information maximum likelihood (FIML) estimator, it may not be efficient in all cases. It is not, however, prone to misspecification bias. The application of this IBC transformation to GLM was explored by Basu and Rathouz (2005). Their model, like Wooldridge's original model, allows for possible conditional heteroskedasticity, is implemented using an extension to the estimating equations in GLM. Since this also is not a FIML estimator it is not as efficient as a FIML model.

In Chapter 2 we introduced the generalized gamma (GG) estimator which is known for the flexibility it allows in the conditional distribution of the outcomes. In particular, it allows for skewness and subsumes many different distributions as special

cases. Being a FIML estimator, GG is fully asymptotically efficient. Here we combine the distributional flexibility and asymptotic precision of the GG model with the conditional mean regression flexibility of the IBC specification and obtain a new model called the generalized gamma with inverse Box-Cox transformation (GG-IBC). This composite model, with additional flexibility in its parametric structure, is consistent, relatively robust, and efficient relative to alternatives such as the NLS and the GG model.

In this chapter, using extensive simulation analyses, we show that when the data is obtained from a distribution with a non-exponential conditional mean: (1) The GG-IBC estimator is consistent while GG is not; (2) The GG-IBC estimator is consistent for the average marginal effect; and (3) The GG-IBC estimator is more precise than the NLS applied to an IBC conditional mean regression specification (NLS-IBC) and the GG estimator. We also used the GG-IBC model in a re-examination of Mullahy's (1997) model of the effect of mothers' cigarette smoking during pregnancy on newborn birthweight (ignoring the potential endogeneity of smoking). For the purpose of comparison, we also estimated the model using the NLS-IBC and the GG estimators.

The rest of this chapter is organized as follows. In Section 3.2 we begin by detailing the IBC transformation, and discussing how it can be combined with the GG conditional distribution specification to obtain the GG-IBC estimator. Section 3.3 summarizes the simulation analysis, briefly outlining the sampling designs, other estimators used in our comparisons, and our evaluation metrics (i.e., average percentage absolute bias and mean squared errors). The section concludes with a summary of the simulation results. Section 3.4 presents the real data application where we use GG-IBC to estimate the effect of mothers' cigarette smoking during pregnancy

on infant birthweight. Finally in Section 3.5 we present our concluding remarks for this chapter.

3.2 Inverse Box-Cox Transformation in the GG Framework

In this section we extend the GG model by specifying the conditional mean regression model using the IBC transformation. This transformation allows for more flexibility in the functional form of the regression model. Note that this is in contrast with popularly used estimators like GLM, log normal estimators etc. which rigidly assume an exponential conditional mean.

3.2.1 Inverse Box-Cox Transformation

First introduced by Box and Cox (1964), the Box-Cox transformation is an application of a power transformation to the outcome variable y that brings about linearity in the parameters of the regression model. They suggested that the outcome variable y be transformed according to following equation

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} \text{ , if } \lambda \neq 0 \\ \log(y) \text{ , if } \lambda = 0 \end{cases} . \qquad (3\text{-}1)$$

In their paper they were interested in the inference on the transformation parameter λ.

Wooldridge (1992) observed that, in general, the conditional mean of y given the regressors is of prime interest and that a transformation of the outcome variable as suggested by Box and Cox (1964) does not provide any direct help in estimation and inference in the conditional mean regression context. He instead used the Box-Cox transformation as a device for generalizing the functional form of the conditional mean. This generalization subsumes many special cases, such as the exponential form, the

classical linear model, and the power specifications. This IBC transformation form of the conditional mean is given as

$$E(y \mid x) = \begin{cases} [1+\lambda(x\beta)]^{1/\lambda}, & \lambda \neq 0 \\ \exp(x\beta) & , \lambda = 0 \end{cases}. \tag{3-3}$$

The interpretation of the transformation of $x\beta$ in (3-3) as the "inverse" of the transformation (of y) used in (3-1) should be noted. The advantage of introducing more flexibility to the conditional mean is that it leads to a more robust estimator. The transformation given in equation (3-3) is modified and used in the following section.

3.2.2 Generalized Gamma with a Flexible Form Conditional Mean Function

The GG model encompasses FIML models based on a variety of distributions such as standard gamma, log normal, exponential, Weibull, etc.[1] At the same time, in the GG framework the parameter μ, which serves as the foundation for the formulation of the conditional mean regression, has a very restrictive form, as is seen in equation (2-2). By specifying μ using the IBC transformation, we seek to generate a more robust GG model (i.e., one that is less susceptible to misspecification bias).

In the GG framework, we respecify (2-2) using the IBC transformation function in the following way

$$\mu = \ln\left(k\left(x\beta^*, \tau\right)\right) \tag{3-4}$$

where

$$k(a,\tau) = \begin{cases} \left(\left((\tau/2)a+1\right)^2\right)^{1/\tau} & \tau \neq 0 \\ \exp(a) & \tau = 0 \end{cases} \tag{3-5}$$

[1] We have described the GG model in detail in Chapter 2.

and τ is a scalar parameter such that $-\infty < \tau < \infty$. Equation (3-5) is a variant of the transformation used by Wooldridge (1992) given in equation (3-3). The conditional mean of the outcome, implied by (3-4) will assume different specifications according to varying values of the parameter τ. These can be listed as follows:

Case 1 (τ = 0): As τ approaches zero, equation (3-5) approaches $k(a,\tau) = \exp(a)$ and subsequently $\mu = x\beta^*$. The case for $\tau = 0$ is the conventional GG specification where the conditional mean of the outcome takes on the following exponential linear index form

$$E[y \mid x] = \exp\left(x\beta^* + C(\sigma,\kappa)\right) \tag{3-6}$$

as shown in equation (2-3).

Case 2 (τ = 2): For $\tau = 2$, equation (3-5) becomes $k(a,\tau) = a + 1$ and $\mu = \ln(x\beta^* + 1)$. Substituting this into equation (2-3) yields the following classical linear conditional mean regression model

$$E[y \mid x] = \exp\left(\ln(x\beta^* + 1) + C(\sigma,\kappa)\right) = x\beta \tag{3-7}$$

where $\beta = \beta^* \exp(C(\sigma,\kappa))$.

Case 3 ($\tau \neq$ 0, 2): Here $\mu = \ln\left(k(x\beta^*,\tau)\right)$ and substituting this into equation (2-3) yields the following expression for the conditional mean

$$E[y \mid x] = \exp\left(\ln\left(k(x\beta^*,\tau)\right) + C(\sigma,\kappa)\right). \tag{3-8}$$

The GG-IBC estimator is generated by substituting (3-4) into the GG probability density function in equation (2-1). It is estimated by using the maximum likelihood (ml) procedure in Stata®, and the parameters to be estimated are β^*, σ, κ, and τ. The

vector β^* is comprised of the regression coefficient parameters, σ and κ are the basic parameters of the GG distribution, and τ is the transformation parameter in the IBC equation.

3.3 Simulation Analysis

The simulation analysis presented here aims to verify three primary statistical properties of the GG-IBC estimator. First, we validate the consistency of the GG-IBC method based on the estimated values of the marginal effect of a variable of interest using the average percentage absolute bias as the evaluation criteria. The average percentage absolute bias of the marginal effect is evaluated for increasing sample sizes from 10,000 to 500,000. If the estimator is consistent, as theory predicts, the average percentage absolute bias of the marginal effect estimates should diminish as the sample size increases. Similar analyses are conducted for the relevant alternative methods, GG and NLS-IBC. Second, using these same results, we analyze the relative unbiasedness (consistency) of the GG-IBC vis-à-vis GG and NLS-IBC. Finally, we seek to verify the asymptotic efficiency of the GG-IBC estimator using the mean squared error of the marginal effect as our evaluation metric. For this we present the percent relative efficiency gain values which measure the relative efficiency of GG-IBC estimator compared to the GG and the NLS-IBC estimators.

In order to accomplish these objectives we use various sampling designs to generate the data. A variety of conditional distributions are used including the standard gamma, exponential, Weibull, GG, log-normal and beta, with different conditional mean specifications. Note that different conditional mean functions here correspond to different values of the τ parameter. The criteria for evaluation include the average

percentage absolute bias and the mean squared error of the estimated marginal effect[2], i.e., the expected derivative of the conditional mean function with respect to a continuous regressor.

3.3.1 Sampling Designs

3.3.1.1 The observable variables

In this simulation analysis our focus is on continuous regressors, although this analysis is also applicable to binary and count regressors. There are two observable confounders, x_p and x_o, distributed uniformly in the [0, 4] interval. Without loss of generality, we choose x_p as the policy variable of interest. Both regressor coefficients, β_p and β_o, are chosen to be equal to 0.5 and the intercept is set to zero. For each repetition of the simulation, j (the total number of repetitions is 500), and for each sampling design, the same observable variables are used. The simulation analysis is implemented in Stata/Mata 10®.

3.3.1.2 The outcome variable

In each instance the distribution of the outcome variable chosen to be one of – standard gamma, Weibull, exponential, lognormal, GG, or beta. Then the value of parameter τ is picked to generate data for a particular conditional mean function. Our objectives in implementing these varied sampling designs are threefold. The first objective is to analyze the GG-IBC estimator in the context of distributions that are within the GG family. The second objective is to observe if the GG-IBC estimator maintains its accuracy for a distribution outside the GG family but which has an

[2] We can also estimate the treatment effect for a binary regressor and the incremental effect for a count regressor.

exponential conditional mean. Finally, we want to analyze the behavior of the GG-IBC estimator across varying τ values.

The various distributions used for data generation are detailed below:

Standard gamma: The first sampling design for the outcome variable is the standard gamma, a special case of GG where $\kappa = \sigma$. In order to generate an outcome variable with alternative conditional mean functions, the IBC transformation is implemented in the Stata® rgamma(a,b) command, where "a" is the scale parameter and "b" is the shape parameter. By assigning varying values to the τ parameter, the Stata® command rgamma(1/|κ|^2,exp(ln(((τ/2)*xβ+1)^(2/τ)))|κ|^2) is used to generate data in the simulation analysis. A right skewed outcome variable is formed with the shape parameter value $1/|\kappa^2| = 2.0$, i.e., $\kappa \cong 0.71$.

Log-normal: The second data generation process used is the log-normal distribution. It is a special case of the GG distribution where the parameter κ goes to zero and σ can take any value. For the log-normal sampling design, in the first step we generate a normally distributed error term e using Stata® command rnormal(a,b), where "a" is the mean and "b" is the standard deviation. In the second step, we substitute this error term into the following regression function $y = \exp\left(\ln\left(k\left(x\beta^*,\tau\right)\right) + e\right)$, where vector x includes the policy variable of interest x_p and the other observable variable x_o. The Stata® command used for generation of this outcome variable is exp(ln(((τ/2)*xβ+1)^(2/τ))+e). The parameter σ is set to 0.2.

Weibull: The third data generation process used is the Weibull. The Inverse Transform Method (Rubinstein, 1981) is applied to implement the Weibull sampling

design with the insertion of IBC transformation, $y = \left(-\ln(U)\right)^{\sigma} \exp(\mu)$ in place of μ in equation (3-4). Note that U is a uniformly distributed random variable. The Stata® command used for generation of the outcome variable is exp(ln(((τ/2)*xβ+1)^(2/τ)))*(-ln(runiform()))^σ, where the value of τ governs the form of the conditional mean function. The value chosen for the parameter σ is 0.5.

Exponential: The next data generating process chosen is the exponential distribution. The exponential is a special case of standard gamma where $\kappa = \sigma = 1.0$. It can be generated using the rgamma(1,b) command in Stata®. Again the shape parameter is replaced by equation (3-4) and the full Stata® command utilized to generate this outcome variable is rgamma(1,exp(ln(((τ/2)*xβ+1)^(2/τ)))) with varying τ values.

Generalized gamma (GG): This is the last sampling design chosen from the GG class of distributions. This data generating process utilizes the standard gamma generator as given in equation (2-14) (Tadikamalla, 1979). The parameter μ is replaced by equation (3-4) to generate varying conditional mean functions. The Stata® command used for generating the outcome variable using this sampling design is d*((rgamma(a,1))^(1/c))*(exp(ln(((τ/2)*xβ+1)^(2/τ)))). The selected parameter values are $1/|\kappa|^2 = 0.5$, i.e., $\kappa \cong 1.41$, and $\kappa/\sigma = 2.0$, i.e., $\sigma \cong 0.71$.

Beta: The last data generating process is based on a variant of the beta distribution, which is outside the GG class of distribution. It does, however, have an exponential conditional mean function. The support of this distribution is the unit interval and it has two parameters α and γ. The mean for the beta distribution is defined as

$E[z] = \frac{\alpha}{\alpha+\gamma}$, where z is the beta distributed random variable. We extend the support of this random variable to the positive part of the real line and generate a new random variable $y = \theta z$, where $0 < \theta < \infty$ and $0 < y < \infty$. The mean of the new distribution then becomes $E[y] = \theta\left(\frac{\alpha}{\alpha+\gamma}\right)$. This distribution can be used to generate a wide variety of distributional shapes – e.g., right skewed, left skewed, uniform, symmetric, modes at the extremes, etc.

To develop a conditional (on x) version of the random variable y, we can specify

$$\theta = k(x\beta^*, \tau), \tag{3-9}$$

where $k(x\beta^*, \tau)$ is defined as in equation (3-5). This leads to the conditional mean function of y as

$$E[y \mid x] = k(x\beta^*, \tau)\left(\frac{\alpha}{\alpha+\gamma}\right) = k(x\beta^*, \tau)\exp(C) \tag{3-10}$$

where $C = \ln\left(\frac{\alpha}{\alpha+\gamma}\right)$.

The beta sampling design is implemented using the (((τ/2)*xβ+1)^(2/τ))*rbeta(α,γ) command. In our experiments various values for the parameters α and γ were tried but some of them failed to converge with one or more of the estimation methods that we detail in Section 3.3.2. Here we have reported results for $\alpha = 2$ and $\gamma = 2$.

3.3.2 Estimators Used in Evaluation and Comparison

The GG-IBC estimator is a FIML estimator. It is programmed in Stata® using maximum likelihood programming via the ml command. The main difference between the GG and the GG-IBC estimators is the inclusion of an additional parameter τ (the

IBC parameter) in the latter case. The parameter τ defines the form of the conditional mean which could be a function other than $x\beta$ in GG-IBC. We compare GG-IBC estimator with two other estimators. The first one is the GG estimator which is selected to analyze the effect of introducing a flexible conditional mean form in an estimator. Note that both the GG and the GG-IBC estimators are estimated with FIML. The second estimator is the NLS with IBC transformation. This NLS-IBC estimator, like GG-IBC, also takes the flexible conditional mean form into account. But since NLS-IBC is not a FIML estimator it does not take higher order moments into account which may lead to decreased precision. The NLS-IBC estimator is chosen in order to compare the GG-IBC with an estimator that also has a flexible conditional mean form but is not a FIML estimator. In the following we describe the three chosen estimators in detail.

Generalized gamma (GG): The GG estimator is the natural alternative estimator for comparison with GG-IBC. Since it is also a FIML estimator like GG, it is based on the full conditional moment assumptions. In GG, the parameter μ is assumed to be equal to $x\beta^*$, which as we have seen in equation (2-3) implies that $E[y \mid x] = \exp(x\beta^* + C(\sigma,\kappa))$. The main objective here is to observe how the GG estimator works when the data has a different conditional mean form than the one assumed by the GG estimator. The streg command in Stata® is used to implement this model.

Nonlinear least squares with IBC (NLS-IBC): The objective of this chapter is to evaluate the usefulness of using a flexible conditional mean function with a FIML estimator. So far in the literature the flexible conditional mean form has only been used with estimators that are not FIML estimators. NLS-IBC (Kenkel and Terza, 2001;

Wooldridge, 1992) is one such estimator with which we want to compare our model. The conditional mean function of the NLS-IBC model can be defined as

$$E[y \mid x] = \exp\left(\ln\left(k\left(x\beta,\tau\right)\right)\right) = k\left(x\beta,\tau\right) \tag{3-11}$$

where the elements of β are the same as those of β*, given in equation (3-8), but scaled by the factor $\exp((\tau/2)C)$, and with the scaled constant term shifted by $(2/\tau)[\exp((\tau/2)C)-1]$.[3]

The comparison between the GG-IBC and the NLS-IBC estimators will highlight the usefulness of using a FIML based estimation even when flexible conditional mean forms are used. The NLS estimator with the exponential conditional mean assumption with IBC transformation was implemented in Stata® using the nl command.

Generalized gamma with IBC (GG-IBC): This is the proposed GG-IBC estimator which was detailed in Section 3.2.2. Note that both the GG-IBC estimator and the NLS-IBC estimator offer flexibility with respect to the conditional mean, $E[y \mid x]$. Therefore, both are robust to misspecification bias. When the data distribution is from the GG family, GG-IBC estimator is consistent irrespective of the conditional mean form. The NLS-IBC estimator on the other hand is consistent irrespective of the data distribution since it does not make any assumptions about it. The GG-IBC estimator is implemented in Stata® using maximum likelihood program and ml commands.[4]

3.3.3 Criteria for Evaluation and Comparison

Health economic models are generally nonlinear by nature and the researchers are usually interested in estimating the marginal effects of one or more of the

[3] See Appendix D for the derivation of the scaling terms.

[4] For more information in programming maximum likelihood estimators in Stata® see Gould et al. (2003).

regressors. The coefficient estimates, $\hat{\beta}$, in a model do not represent the marginal effects of the regressors in a nonlinear setting. The marginal effect for a continuous policy variable, x_p, is defined as the derivative of the conditional mean of the outcome variable with respect to x_p.

For the GG-IBC model, for each case summarized in Section 3.2.2, there is a different conditional mean function as given in equations (3-6), (3-7) and (3-8). The marginal effects for these cases are defined as[5]

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right] = E\left[\exp\left(x\beta^* + C(\sigma,\kappa)\right)\beta_p^*\right], \tag{3-12}$$

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right] = E\left[\beta_p^* \exp\left(C(\sigma,\kappa)\right)\right], \text{ and,} \tag{3-13}$$

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right] = E\left[((\tau/2)x\beta^* + 1)^{(2/\tau)-1}\beta_p^* \exp\left(C(\sigma,\kappa)\right)\right], \tag{3-14}$$

respectively.

The marginal effects as given in (3-12), (3-13) and (3-14) can be consistently estimated using

$$ME = \hat{\beta}_p^* \left(\frac{\sum_{i=1}^{n} \exp\left(x_i\hat{\beta}^* + C(\hat{\sigma},\hat{\kappa})\right)}{n} \right), \tag{3-15}$$

$$ME = \hat{\beta}_p^* \exp\left(C(\hat{\sigma},\hat{\kappa})\right), \text{ and,} \tag{3-16}$$

[5] Appendix E summarizes the derivation of these marginal effects.

$$ME = \hat{\beta}_p^* \left(\frac{\sum_{i=1}^{n} ((\hat{\tau}/2)x_i\hat{\beta}^* + 1)^{(2/\hat{\tau})-1} \exp\left(C\left(\hat{\sigma},\hat{\kappa}\right)\right)}{n} \right), \quad (3\text{-}17)$$

respectively.

Both the bias and the precision of the marginal effect as estimated from the GG-IBC model are of interest. We used the following two metrics in our comparison of models:

The first is the average percentage absolute bias of the estimated marginal effect which is defined as

$$APAB_{kd\tau} = \left[\sum_{j=1}^{500} \frac{1}{500} \left(\frac{ABS(ME_{jkd\tau} - ME_{d\tau})}{ABS(ME_{d\tau})} \right) \right] \times 100\%, \quad (3\text{-}18)$$

where ME is defined as the "true" average marginal effect, "j" is the number of repetitions in the simulation, j = 1,…,500, "k" is the number of observations used in the simulation analysis, k = 10,000, 50,000, 100,000, 250,000 and 500,000, and "d" is the sampling design. The true value of the marginal effect is derived separately for each sampling design and τ value. To obtain the true values of (3-12), (3-13) and (3-14) we generated a large sample of 5 million observations using Monte Carlo simulation and calculated the true values from this super sample average.

The second metric for comparison is the mean squared error of the estimated marginal effect which is defined as

$$MSE_{kd\tau} = \left[\sum_{j=1}^{500} \frac{(ME_{jkd\tau} - ME_{d\tau})^2}{500} \right] \quad (3\text{-}19)$$

where "j", "k" and "d" are defined as above.

3.3.4 Simulation Results

This simulation analysis aims to examine the efficiency gains achieved with the use of a FIML estimator as well as the bias performance attained through the introduction of a flexible conditional mean form. As mentioned before in Section 2.3.4, Manning et al. (2005) showed that the GG estimator provides higher efficiency compared to the log-linear and the GLM models. Here we want to investigate whether this efficiency advantage of the GG model is preserved even after incorporation of the above described IBC transformation. As before (Section 2.3.4), the mean squared error of the marginal effect, as defined in equation (3-19), will be our metric of evaluation and comparison as we examine the statistical efficiency of various models. Along with efficiency, the consistency of the GG-IBC estimator is also of interest since we have introduced additional flexibility through the IBC transformation. Here we use the average percentage absolute bias of the marginal effect, as defined in equation (3-18), as our metric for evaluation and comparison.

Table 3-1 summarizes the average percentage absolute bias of the marginal effect of the policy variable x_p for the GG sampling design. Here we present results from increasing sample sizes, 10,000, 50,000, 100,000, 250,000 and 500,000. Since our model is a nonlinear model, the marginal effects are calculated using three different quartile values of the policy variable x_p. The objective here is to examine the consistency of the GG-IBC estimator relative to other estimators. As the number of observations increases, the average percentage absolute bias for the GG-IBC estimator goes to zero, for instance, for samples of size 10,000, it is 1.882%, it goes down to 0.820% for samples of size 50,000 and further down to 0.576% for samples of size

100,000 and so on. Its noteworthy that the average percentage absolute bias for the GG estimator remains high even as the sample size is increased, e.g., for samples of size 10,000 it is 9.005%, goes down a little to 8.752% for sample size 50,000 but increases to 8.796% for samples of size 100,000 etc. It is evident that the GG estimator is not a consistent estimator since it does not take into account the non-exponential functional form of the conditional mean used to generate the outcome variable. The NLS-IBC estimator on the other hand is a consistent estimator as can be seen in Table 3-1, although it has slightly higher average percentage absolute bias compared to the GG-IBC estimator. As was discussed earlier, the NLS-IBC estimator is consistent.

Next we examine the bias performance of the GG-IBC estimator as compared to the GG and the NLS-IBC estimators. Table 3-2 presents the average percentage absolute bias of the estimated marginal effect for the samples of size10,000 for the various sampling designs detailed in Section 3.3.1.2. Note that for every sampling design, the GG-IBC estimator has lower average percentage absolute bias as compared to the GG and the NLS-IBC estimators. The GG estimator is not consistent since it does not correct for the conditional mean functional form. For example, for the standard gamma distributed samples, the GG estimator generates an 8.914% bias whereas the GG-IBC has a bias of 1.803% for the first quartile of variable x_p. For the second quartile of x_p the percentage biases from GG and GG-IBC are closer to each other due to the fact that the functional form of the two percentage bias functions are closest to each other at the mean value of x_p. For the third quartile of x_p, the percentage bias from the GG goes up to 19.662% whereas the percentage bias from the GG-IBC is still low at 2.910%. Similar results are observed for the other sampling

designs too – Weibull, exponential, log-normal, generalized gamma, and beta. Although NLS-IBC is a consistent estimator, the GG-IBC estimator generates slightly less percentage bias. For example, for Weibull distributed data at the first quartile of x_p, the average percentage absolute bias is 1.703% for the NLS-IBC whereas it is 1.263% for the GG-IBC. The values observed at other quartiles are similar in nature. This trend is maintained for the other sampling designs and similar results are observed in comparisons of the GG-IBC and the NLS-IBC estimators.

For examining the precision of estimators, mean squared error is the most widely used metric of evaluation and comparison. Table 3-3 presents the mean squared error of the estimated marginal effect of the policy variable x_p for the samples of size 10,000 for various sampling designs. Note that since the marginal effects generated by the three included estimators are small, the computed mean squared errors are quite small. In view of this, to highlight the efficiency gain attained by the use of GG-IBC compared to the GG and the NLS-IBC estimators, we used a percentage measure in our evaluation which indicates the efficiency of the GG-IBC estimator relative to that of the alternative estimators. It is defined as $\left(\left(MSE_m - MSE_{GG-IBC}\right)/MSE_m\right)*100$ where m = GG, NLS-IBC. The values next to the mean squared errors, in the parentheses, represent this percentage efficiency gain. Since the GG-IBC estimator is a FIML estimator it is expected to provide efficiency gains compared to NLS-IBC and this is indeed what we observe. For the standard gamma distributed data, GG-IBC provides 53% efficiency gain for the first quartile, 23% for the second quartile and 30% for the third quartile of x_p compared to NLS-IBC. Similarly, for the rest of the sampling designs, the efficiency gain from GG-IBC compared to NLS-IBC ranges from 9% to an

impressive 91%. The percentage efficiency gain from GG-IBC compared to GG ranges from 43% to 99% for different sampling designs.

Most commonly, researchers are interested in the marginal effects of various policy variables, but in a nonlinear model setting the parameter estimates by themselves do not represent the marginal effects. However, in Table 3-4 we present the parameter estimates estimated by various methods for the sake of completeness and also to highlight the consistency in the estimates of the crucial parameter $\hat{\tau}$ as obtained from the GG-IBC and NLS-IBC estimators. It can be noted that the parameter estimates obtained form the GG-IBC estimator are consistent for data distributed as gamma, Weibull, exponential, log-normal and GG. Also the τ parameter is estimated correctly in all the cases. For the beta sampling design, since the beta distribution is not in the GG class of distributions, the parameter estimates for β_p, β_o and the constant are slightly different, yet the τ parameter estimates appear to be the same. However, the GG estimates of β_p, β_o and the constant are biased, due to the conditional mean misspecification bias. In the case of the NLS-IBC estimator, the conditional mean is defined as $\exp(\mu)$, where μ is given as $\ln\left(k\left(x\beta,\tau\right)\right)$, without the constant shifter C. For the gamma and the exponential sampling designs this constant shifter is defined as zero (see equation (3-11) and Appendix D for details.). As a result of this, in the gamma and the exponential sampling designs, we can compare the NLS-IBC parameter estimates with the GG-IBC parameter estimates. For other sampling designs – Weibull, log-normal, GG and beta, the NLS-IBC estimator internally adjusts the parameter estimates to compensate for this constant shifter.

Although in Tables 3-1, 3-2, 3-3 and 3-4 the τ parameter is chosen to be unity, the results are similar for other values of τ, such as τ = 0.5, 1.0, 1.5, 3 ,4. Recall that the parameter τ controls the form of the conditional mean function. Table 3-5 presents the average percentage absolute bias of the marginal effect of the policy variable x_p for various τ values. Here, as an example, the sampling design is chosen to be based on the GG distribution and the sample size is set to 10,000. Note that the percentage bias for the GG estimator is 6.169% for $\tau = 0.5$ and it goes up to 16.703% for $\tau = 4.0$ for the first quartile of x_p. The GG-IBC estimator, on the other hand, consistently generates lower percentage bias compared to GG for all τ values. For example, for $\tau = 1.5$ the percentage bias from GG is 11.106% and for GG-IBC it is 2.348% etc. Finally the percentage bias for NLS-IBC is slightly higher than the percentage bias for GG-IBC for all τ values. These results show that the GG-IBC estimator generates less bias regardless of the conditional mean form, i.e., regardless of the τ values.

3.4 The Effect of Cigarette Smoking on Birthweight

Here we demonstrate the practicality of the GG-IBC model via a real data application. We also compare the results obtained using the GG-IBC model with the GG and the NLS-IBC models mentioned earlier. We revisit Mullahy's (1997) model wherein he considers the effect of mothers' cigarette smoking during pregnancy on infant birthweight. This data set has 1,388 observations and the outcome variable and the covariates are from the Child Health Supplement to the 1988 National Health Interview Survey (Mullahy, 1997).[6] The outcome variable is defined as the birthweight in pounds. Figure 3-1 shows the histogram and the Kernel density of the outcome variable. The

[6] Refer to Mullahy (1997) for additional information on the data set.

policy variable of interest is the number of cigarettes smoked during pregnancy. Additional covariates used are child's sex, race, and birth order. The definitions of all of the variables in the analysis are given in Table 3-6 and their descriptive statistics are given in Table 3-7.

Our objectives here are two-fold. First, we want to verify that the results obtained by our estimator are empirically meaningful, i.e., there is a negative impact of smoking on infant birthweight. Second, we want to study the practical applicability of our GG-IBC model, especially since it is a FIML method with multiple parameters as unknowns. For the purpose of this chapter, we will ignore the potential endogeneity of smoking and use our GG-IBC model. We will revisit the effects of endogeneity in conjunction with flexible conditional mean form for this application in Chapter 4.

3.4.1 Model

We first estimate Mullahy's (1997) birthweight model assuming that the conditional mean regression for y follows the IBC formulation given in equation (3-11) of the text. Here we use the NLS-IBC estimator. The metric of our interest, the marginal effect (ME) in the NLS-IBC context can be defined as

$$ME = E\left[\frac{\partial k(x_p\beta_p + x_o\beta_o, \tau)}{\partial x_p}\right] = E\left[\frac{\partial \psi^{(2/\tau)}}{\partial x_p}\right] \tag{3-20}$$

where

$$k(x\beta, \tau) = \psi^{(2/\tau)} \tag{3-21}$$

$$\psi = \frac{\tau}{2}(x\beta) + 1 \tag{3-22}$$

and

$$x\beta = x_p\beta_p + x_o\beta_o \tag{3-23}$$

where x_p is the policy variable of interest in this application – the number of cigarette smoked during pregnancy – and x_o is the vector of observable covariates. By combining equations (3-20) to (3-23), we obtain the following expression for the ME:

$$ME = E\left[\frac{2}{\tau}\psi^{(2/\tau - 1)}\left(\frac{\tau}{2}\beta_p\right)\right] = E\left[\left(\psi^2\right)^{\left(\frac{1}{\tau} - \frac{1}{2}\right)}\beta_p\right] \tag{3-24}$$

The estimator of the ME of cigarette smoking is the following sample analog to (3-24)

$$ME = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{\left(\hat{\psi}_i^2\right)^{\left(\frac{1}{\hat{\tau}} - \frac{1}{2}\right)}\hat{\beta}_p\right\} \tag{3-25}$$

where $\hat{\psi}_i = \frac{\hat{\tau}}{2}\left(x_{pi}\hat{\beta}_p + x_{oi}\hat{\beta}_o\right) + 1$, n_s is the size of the subsample of smokers, and the "^s" denote estimates. Note that the marginal effect is estimated for smokers only.

The average effect of smoking cessation on birthweight is the expected (i.e., average) effect from forcing all smokers to quit and it pertains to the smokers only. For the NLS-IBC model the cessation effect is given by

$$CE = E\left[k(x_p\beta_p + x_o\beta_o, \tau) - k(x_o\beta_o, \tau)\right] \tag{3-26}$$

and this is estimated as

$$CE = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{k(x_{pi}\hat{\beta}_p + x_{oi}\hat{\beta}_o, \hat{\tau}) - k(x_{oi}\hat{\beta}_o, \hat{\tau})\right\} \tag{3-27}$$

where n_s is the size of the subsample of smokers.

We next estimate Mullahy's (1997) birthweight model assuming that the conditional density function of y follows the GG-IBC formulation given in equation (3-8) of the text. The metric of our interest as before, the ME in the GG-IBC context can be defined as

$$ME = E\left[\frac{\partial k(x_p\beta_p^* + x_o\beta_o^*, \tau)\times \exp(C)}{\partial x_p}\right] = E\left[\frac{\partial\left(\psi^{*(2/\tau)}\times \exp(C)\right)}{\partial x_p}\right] \tag{3-28}$$

where

$$k(x_p\beta_p^* + x_o\beta_o^*, \tau) = \psi^{*(2/\tau)} \tag{3-29}$$

$$\psi^* = \frac{\tau}{2}\left(x\beta^*\right) + 1 \tag{3-30}$$

and

$$x\beta^* = x_p\beta_p^* + x_o\beta_o^* \tag{3-31}$$

where x_p is the policy variable in this application – the number of cigarette smoked during pregnancy – and x_o is the vector of observable covariates and the constant C is

$$C = \left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right). \tag{3-32}$$

Therefore, we obtain the following expression for ME,

$$ME = E\left[\frac{2}{\tau}\psi^{*(2/\tau - 1)}\left(\frac{\tau}{2}\beta_p^*\right)\exp(C)\right] = E\left[\left(\psi^{*2}\right)^{\left(\frac{1}{\tau} - \frac{1}{2}\right)}\beta_p^*\exp(C)\right]. \tag{3-33}$$

The estimator of the ME of cigarette smoking is the following sample analog to (3-33)

$$ME = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{\left(\hat{\psi}_i^{*2}\right)^{\left(\frac{1}{\hat{\tau}} - \frac{1}{2}\right)}\hat{\beta}_p^*\exp(\hat{C})\right\} \tag{3-34}$$

where

$$\hat{\psi}_i^* = \frac{\hat{\tau}}{2}\left(x_{pi}\hat{\beta}_p^* + x_{oi}\hat{\beta}_o^*\right) + 1 \tag{3-35}$$

$$\hat{C} = \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\ln\left(\hat{\kappa}^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\hat{\kappa}^2}\right) + \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\hat{\kappa}^2}\right\}\right) \tag{3-36}$$

and the "^s" denote estimates.

Here we can also derive the average effect of smoking cessation on birthweight. The appropriate equation for average effect of smoking cessation is defined as

$$CE = E\left[\left(k(x_p\beta_p^* + x_o\beta_o^*, \tau) - k(x_o\beta_o^*, \tau)\right)\exp(C)\right] \tag{3-37}$$

and it is estimated as

$$CE = \sum_{i=1}^{n_s} \frac{1}{n_s}\left\{\left(k(x_{pi}\hat{\beta}_p^* + x_{oi}\hat{\beta}_o^*, \hat{\tau}) - k(x_{oi}\hat{\beta}_o^*, \hat{\tau})\right)\exp(\hat{C})\right\} \tag{3-38}$$

where n_s is the size of the subsample of smokers.

Finally, we estimated the model using the standard GG estimator. The expressions for ME and its estimators, along the same lines as above, are given by equations (3-39) and (3-40) respectively below.

$$ME = E\left[\frac{\partial \exp(x_p\beta_p^* + x_o\beta_o^* + C(\sigma,\tau))}{\partial x_p}\right] = E\left[\exp(x_p\beta_p^* + x_o\beta_o^* + C(\sigma,\tau))\beta_p^*\right] \tag{3-39}$$

where $C(\sigma,\kappa)$ given in equation (3-32).

$$ME = \sum_{i=1}^{n_s} \frac{1}{n_s}\exp(x_{pi}\hat{\beta}_p^* + x_{oi}\hat{\beta}_o^* + C(\hat{\sigma},\hat{\kappa}))\hat{\beta}_p^* \tag{3-40}$$

where $C(\hat{\sigma},\hat{\kappa})$ given in equation (3-36). The expressions for CE and its estimators, along the same lines as above, are given by equations (3-41) and (3-42) respectively below.

$$CE = E\left[\left(\exp(x_p\beta_p^* + x_o\beta_o^*) - \exp(x_o\beta_o^*)\right)\exp(C)\right] \tag{3-41}$$

$$CE = \sum_{i=1}^{n_s} \frac{1}{n_s}\left\{\left(\exp(x_{pi}\hat{\beta}_p^* + x_{oi}\hat{\beta}_o^*) - \exp(x_{oi}\hat{\beta}_o^*, \hat{\tau})\right)\exp(\hat{C})\right\}. \tag{3-42}$$

3.4.2 Results

The results from the GG-IBC, NLS-IBC and GG estimators are given in Table 3-8. The marginal effects as in (3-25), (3-32) and (3-40) estimated for the NLS-IBC, the GG-IBC and the GG models, respectively, are presented in the first row of Table 3-8. For these estimates the sample was restricted (after regression estimation) to those who smoked during pregnancy (i.e., CIGSPREG > 0). This subset was comprised of 212 individuals. The second row presents the average effect of smoking cessation on birthweight for the 212 mothers who smoked during pregnancy. The quantities here are computed using the expressions given in equations (3-27), (3-38) and (3-42) for the NLS-IBC, the GG-IBC and the GG models, respectively.

Foremost, it can be noted that the ME of the policy variable of our interest, as estimated by all the three methods, is negative. This result is also consistent with those obtained by Mullahy (1997). For the case of GG-IBC, we find that the birthweight decreases by 0.489 ounces for every additional cigarette smoked per day during the pregnancy by the mother. Secondly, from the cessation effect, we see that among the smokers, if we force them all to quit – i.e., decreased their observed level of daily smoking from the current level to zero – the birthweight goes up. For the GG-IBC method, we find this increase to be 6.913 ounces.

In addition to this quantitative evaluation, we also found that the GG-IBC estimator is very well behaved in terms of its convergence as compared to the NLS-IBC model, especially with a relatively small dataset like the one used in our application. We found that the NLS-IBC fails to converge with our Stata® implementation wherein the parameter τ is estimated as part of a comprehensive (i.e., including all parameters simultaneously) optimization routine. We were only able to get it to work when the

parameter τ is estimated separately using a line search method in a Gauss implementation of NLS-IBC. Since the results we obtained are fairly similar across the two methods – GG-IBC and NLS-IBC – the ease of use suggests that GG-IBC may be more useful in a practical setting. It should also be noted that since our extensive simulations with various sampling designs and sizes showed that the GG-IBC is a comparatively more precise estimator than the NLS-IBC, we see it as the preferred estimator.

3.5 Summary, Discussion and Conclusion

In Chapter 2, we addressed two commonly observed characteristics of health economics data, namely skewness and endogeneity. In addition to these issues, health economics data can come from distributions which may have a non-exponential conditional mean. This could potentially be a problem if the model used to estimate the data impose the exponential conditional mean assumption, as this can lead to conditional mean misspecification bias. In this chapter we propose a novel model for data estimation, called GG-IBC, which specifically targets this issue.

Note that we have retained the GG model since it is a FIML estimator that has flexible distributional form. We modify the conditional mean expression in this model, via IBC transformation, in order to introduce conditional mean flexibility to handle the aforementioned misspecification bias problem. The composite model presented here retains the precision benefits of the GG model while providing consistent estimates.

In order to test our model, we conducted extensive simulation analyses using various sampling designs and sizes. We compared the results obtained using our model to two commonly implemented alternatives – the NLS-IBC and the GG models. We showed that the GG-IBC model is a consistent estimator. This is true not only for data

obtained from the family of the GG distributions, but, GG-IBC remains consistent even when the data is obtained from a distribution not in the GG family – e.g., the beta distribution. As compared to the NLS-IBC and the GG models, we found that the proposed model consistently has lower average percentage absolute bias for the marginal effects across different distributions. We also confirmed that the GG-IBC model retains the statistical efficiency of the GG model.

Finally, we tested the proposed model on a real data application. We revisited Mullahy's (1997) model of the effect of mothers' daily cigarette smoking during pregnancy on infant birthweight. We showed that, consistent with Mullahy (1997) our model predicts a decrease in the birthweight as the smoking during pregnancy increases. We also found that our model, unlike NLS-IBC, is well behaved with respect to convergence when the number of observations is small, which suggests that the GG-IBC can be quite suitable for real data scenarios.

Table 3-1. For generalized gamma distributed data, average percentage absolute bias of the marginal effect ($\tau = 1$) (in percentages)

Estimator	N	1st quartile of x_p	2nd quartile of x_p	3rd quartile of x_p
GG	10,000	9.005	3.610	19.472
	50,000	8.752	3.128	18.658
	100,000	8.796	3.173	18.743
	250,000	8.634	3.372	19.039
	500,000	8.537	3.543	19.161
GG-IBC	10,000	1.882	2.226	3.089
	50,000	0.820	0.991	1.368
	100,000	0.576	0.691	0.968
	250,000	0.354	0.438	0.617
	500,000	0.261	0.313	0.435
NLS-IBC	10,000	2.609	2.492	3.733
	50,000	1.069	1.121	1.688
	100,000	0.831	0.791	1.219
	250,000	0.489	0.521	0.776
	500,000	0.354	0.348	0.544

Table 3-2. For various sampling designs, average percentage absolute bias of the marginal effect ($\tau = 1$) (N=10,000) (in percentages)

Data	Estimator	1st quartile of x_p	2nd quartile of x_p	3rd quartile of x_p
Gamma	GG	8.914	3.618	19.662
$\gamma = 2.0$	NLS-IBC	2.247	2.184	3.525
	GG-IBC	1.803	2.041	2.910
Weibull	GG	8.942	3.801	19.972
$\sigma = 0.5$	NLS-IBC	1.703	1.667	2.655
	GG-IBC	1.263	1.542	2.294
Exponential	GG	8.699	4.407	20.196
	NLS-IBC	3.318	3.221	5.174
	GG-IBC	2.511	3.086	4.585
Log-normal	GG	8.818	4.242	20.773
$\sigma_\varepsilon = 0.2$	NLS-IBC	0.716	0.644	0.974
	GG-IBC	0.532	0.605	0.871
Generalized	GG	9.005	3.610	19.472
gamma $\kappa / \sigma = 2.0$	NLS-IBC	2.609	2.492	3.733
$\gamma = 0.5$	GG-IBC	1.882	2.226	3.089
Beta	GG	8.642	3.667	19.776
$\alpha = 2,\ \gamma = 2$	NLS-IBC	1.371	1.310	2.068
	GG-IBC	0.889	1.044	1.479

Table 3-3. For various sampling designs mean squared error of the marginal effect with percent relative efficiency gain ($\tau = 1$) (N=10,000)

Data	Estimator	1st quartile of x_p	2nd quartile of x_p	3rd quartile of x_p
Gamma	GG	0.0062978	0.0018094	0.050493
$\gamma = 2.0$		(94.1%)	(63.9%)	(96.6%)
	NLS-IBC	0.0006095	0.0007401	0.0023845
		(38.7%)	(11.78%)	(28.6%)
	GG-IBC	0.0003734	0.0006531	0.001703
Weibull	GG	0.0053677	0.0021317	0.0409466
$\sigma = 0.5$		(97.1%)	(86.0%)	(98.0%)
	NLS-IBC	0.0002827	0.0003511	0.0010945
		(45.9%)	(15.3%)	(24.9%)
	GG-IBC	0.0001530	0.0002975	0.0008215
Exponential	GG	0.0062624	0.0028202	0.0552056
		(87.7%)	(46.3%)	(92.4%)
	NLS-IBC	0.0013399	0.0016627	0.0052693
		(42.4%)	(9.0%)	(20.7%)
	GG-IBC	0.0007723	0.0015134	0.0041792
Log-normal	GG	0.0066033	0.0048269	0.0648765
$\sigma_\varepsilon = 0.2$		(99.5%)	(98.8%)	(99.8%)
	NLS-IBC	0.0000643	0.0000676	0.0001969
		(46.0%)	(12.4%)	(17.9%)
	GG-IBC	0.0000347	0.0000592	0.0001616
Generalized	GG	0.0041126	0.0011594	0.0317138
gamma		(93.2%)	(58.0%)	(96.3%)
$\kappa / \sigma = 2.0$	NLS-IBC	0.0005173	0.000628	0.0018073
$\gamma = 0.5$		(45.7%)	(22.5%)	(35.3%)
	GG-IBC	0.0002808	0.000487	0.0011685
Beta	GG	0.0014405	0.0003661	0.0124458
$\alpha = 2,\ \gamma = 2$		(98.4%)	(88.4%)	(99.1%)
	NLS-IBC	0.0000587	0.0000683	0.0002226
		(61.8%)	(37.6%)	(50.9%)
	GG-IBC	0.0000224	0.0000426	0.0001093

Note: The values given in parentheses are the percent relative efficiency gains that measure the relative efficiency of GG-IBC estimator compared to the GG and the NLS-IBC estimators. The percent relative efficiency gain is defined by $\left(\left(MSE_m - MSE_{GG-IBC}\right) / MSE_m\right) * 100$, where m = GG, NLS-IBC.

Table 3-4. Parameter estimates ($\tau = 1$) (N=10,000)

Data	Parameter	GG	NLS-IBC	GG-IBC
Gamma $\gamma = 2.0$	$\beta_1 = 0.5$	0.2590	0.5011	0.5014
	$\beta_2 = 0.5$	0.2591	0.5015	0.5015
	$\beta_0 = 0$	0.3040	-0.0001	-0.0010
	$\kappa \cong 0.71$	0.7002	-	0.7051
	$\ln(\sigma) \cong -0.35$	-0.3430	-	-0.3470
	$\tau = 1$	-	0.9962	1.0030
Weibull $\sigma = 0.5$	$\beta_1 = 0.5$	0.2590	0.4718	0.4986
	$\beta_2 = 0.5$	0.2591	0.4719	0.4990
	$\beta_0 = 0$	0.3041	-0.1201	0.0002
	$\kappa = 1$	0.9870	-	0.9999
	$\ln(\sigma) \cong -0.69$	-0.6843	-	-0.6934
	$\tau = 1$	-	0.9975	0.9927
Exponential	$\beta_1 = 0.5$	0.2599	0.5035	0.4984
	$\beta_2 = 0.5$	0.2601	0.5040	0.4991
	$\beta_0 = 0$	0.3002	-0.0089	-0.0009
	$\kappa = 1$	0.9960	-	0.9999
	$\ln(\sigma) = 0$	0.0023	-	-0.0003
	$\tau = 1$	-	0.9853	0.9844
Log-normal $\sigma_\varepsilon = 0.2$	$\beta_1 = 0.5$	0.2604	0.5047	0.5001
	$\beta_2 = 0.5$	0.2602	0.5047	0.5000
	$\beta_0 = 0$	0.3053	0.0206	-0.0003
	$\kappa \to 0$	0.0344	-	-0.0012
	$\ln(\sigma) \cong -1.61$	-1.5732	-	-1.6096
	$\tau = 1$	-	0.9981	0.9999
Generalized gamma $\kappa / \sigma = 2.0$ $\gamma = 0.5$	$\beta_1 = 0.5$	0.2587	0.4496	0.5021
	$\beta_2 = 0.5$	0.2591	0.4510	0.5028
	$\beta_0 = 0$	0.3040	-0.2186	-0.0014
	$\kappa \cong 1.41$	1.4041	-	1.4166
	$\ln(\sigma) \cong -0.35$	-0.3410	-	-0.3480
	$\tau = 1$	-	1.0071	1.0045
Beta $\alpha = 2$, $\gamma = 2$	$\beta_1 = 0.5$	0.2581	0.3537	0.4101
	$\beta_2 = 0.5$	0.2579	0.3539	0.4098
	$\beta_0 = 0$	-0.1042	-0.5865	-0.3611
	κ	1.8677	-	2.0110
	$\ln(\sigma)$	-1.1282	-	-1.1877
	$\tau = 1$	-	1.0017	0.9993

Table 3-5. For generalized gamma distributed sampling design with various τ parameter values, average percentage absolute bias of the marginal effect (N=10,000) (in percentages)

	Estimator	1st quartile of x_p	2nd quartile of x_p	3rd quartile of x_p
$\tau = 0.5$	GG	6.169	2.564	13.847
	NLS-IBC	2.411	2.044	3.261
	GG-IBC	1.460	1.763	2.600
$\tau = 1.0$	GG	9.005	3.610	19.472
	NLS-IBC	2.609	2.492	3.733
	GG-IBC	1.882	2.226	3.089
$\tau = 1.5$	GG	11.106	4.322	22.774
	NLS-IBC	2.953	2.958	4.201
	GG-IBC	2.348	2.675	3.554
$\tau = 3.0$	GG	15.084	5.881	28.145
	NLS-IBC	4.234	4.328	5.558
	GG-IBC	3.725	4.013	4.963
$\tau = 4.0$	GG	16.703	6.787	30.211
	NLS-IBC	5.130	5.221	6.471
	GG-IBC	4.666	5.008	6.040

Table 3-6. The variable definitions from the birthweight analysis

Variable	Definition
The outcome variable	
BIRTHWEIGHT	the infant's birthweight measured in lbs.
Policy Variable (x_p)	
CIGSPREG	the number of cigarettes smoked per day during pregnancy
The observable confounders (x_o)	
PARITY	the birth order
WHITE	= 1 if white, 0 otherwise
MALE	= 1 if male, 0 otherwise

Table 3-7. Descriptive statistics for the birthweight sample (N=1,388)

Variable	Mean	Minimum	Maximum
The outcome variable			
BIRTHWEIGHT	7.42	1.44	16.94
Policy variable (x_p)			
CIGSPREG	2.09	0	50
The observable confounders (x_o)			
PARITY	1.63	1	6
WHITE	%78	-	-
MALE	%52	-	-

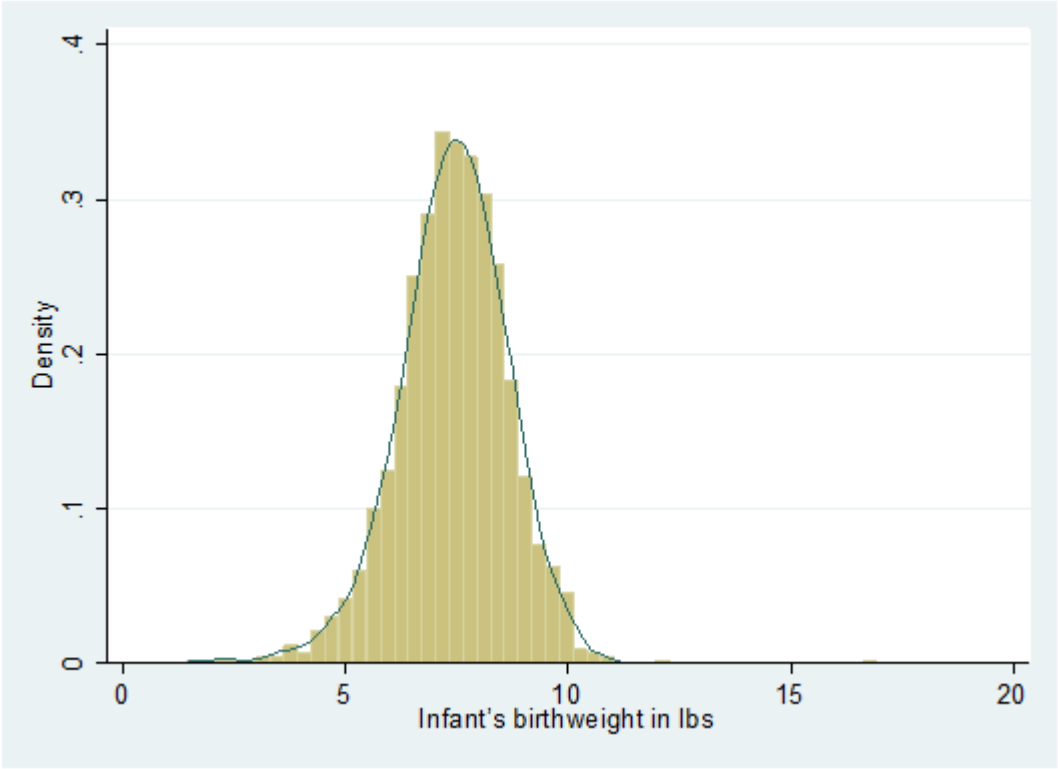

Figure 3-1. Histogram and Kernel density estimate of infant's birthweight in lbs

Table 3-8. The marginal effect and the cessation effect estimates from the birthweight analysis

	Type of estimation		
Marginal effect	NLS-IBC	GG-IBC	GG
Average marginal effect	-0.5383	-0.4894	-0.5080
Cessation effect	7.6853	6.9138	7.0811

CHAPTER 4
MODELING AND ESTIMATING FLEXIBLE-FORM HEALTH ECONOMETRIC MODELS WITH ENDOGENEITY

4.1 Introduction and Background

One of the main properties of regression models in health economics and health services research is skewness in the data on the dependent variable. In the literature the generalized gamma (GG) model is one of the suggested models to deal with such skewness. It is also very often the case in empirical health econometric modeling that one or more of the regressors of interest are endogenous. The two-stage residual inclusion (2SRI) method has been offered as a means of producing consistent estimates in the presence of regressor endogeneity. The model introduced in Chapter 2, the generalized gamma with endogeneity (GGE), offers a way to correct for both skewness and endogeneity in health econometric regression models. Though the GGE model is flexible with regard to accommodating skewness in that it subsumes various asymmetric distributions as its special cases, it is subject to classical misspecification bias because it imposes a fixed conditional mean regression form – the exponential. The objective of the current chapter is to extend the GGE model through the introduction of a flexible conditional mean form in order to lessen the chance of misspecification bias. The combined model (generalized gamma with endogeneity and inverse Box-Cox transformation (GGE-IBC)) will offer flexibility both in the distributional form (to account for skewness) as well as in the conditional mean function (to avoid misspecification bias).

To address skewness and thereby gain precision in health econometric models, Manning et al. (2005), advocated use of the GG estimator, a full information maximum likelihood (FIML) estimator based on a flexible distributional form. The substantial

potential efficiency gains afforded by this model when estimated in a 2SRI framework (we called this GG-2SRI model GGE) were demonstrated by the simulation results presented in Chapter 2. There we also verified that, in presence of endogeneity, the GGE model is indeed consistent, and that application of the GG model (ignoring endogeneity) can lead to substantial bias. In addition to the issue of flexible distributional form addressed in Chapter 2, in Chapter 3 we explored the effect of introducing flexibility in the conditional mean function of the GG model. This modification to the GG model addressed potential bias in the estimates of the targeted marginal effects which can result from assuming an incorrect conditional mean form. We introduced flexibility in the conditional mean function using the inverse Box-Cox (IBC) transformation, which was first applied in a regression context by Wooldridge (1992). We showed that our model, called GG-IBC, combines the efficiency of a FIML estimator with the bias-related benefits obtained due to the use of the IBC transformation. We also demonstrated the usefulness of the GG-IBC model as compared to the GG and non-linear least squares with IBC (NLS-IBC) models through our simulation and real data analyses.

In this chapter we bring together the benefits obtained by our models introduced in Chapter 2 and Chapter 3. Here we present a new model for health econometric data analysis, called generalized gamma with endogeneity and IBC transformation (GGE-IBC), which is both robust to the presence of endogeneity in regressors and flexible enough to accommodate exponential, linear and power specifications in the conditional mean function. Through extensive simulation and real data experiments, we verify that

the GGE-IBC possesses the useful properties inherited from the GGE and the GG-IBC models – viz., consistency and statistical efficiency.

The remainder of this chapter is organized as follows. In Section 4.2 the GGE-IBC model/method, which incorporates the IBC transformation into the GG model and implements the 2SRI method, is described. In Section 4.3 the simulation analysis is summarized, including the sampling designs that are used, the estimators included in the comparisons, the definition of the marginal effect in the GGE-IBC context, the criteria for evaluation and comparison and summary of the simulation results. Section 4.4 describes the real data analysis, where we apply the GGE-IBC model to Mullahy's (1997) data and estimate the effect of mothers' daily cigarette smoking on infant birthweight. Finally, in Section 4.5 we conclude summarizing the contributions and findings in this chapter.

4.2 Integrating IBC and Endogenous Confounders into the GG Model

In Chapters 2 and 3 we addressed two different forms of flexibility in health econometric regression modeling with nonnegative outcomes – the need for flexibility in the distribution of the dependent variable conditional on the regressors as a means of accommodating skewness; and flexibility in the conditional mean regression specification to avoid misspecification bias. We allow for distributional flexibility in Chapter 2 by incorporating the GG model which subsumes various skewed distributions like log-normal, standard gamma, Weibull and exponential as its special cases. The choice of these distributions is controlled by the parameters κ and σ in the GG model. In Chapter 3, we accommodate conditional mean flexibility, through the use of the IBC transformation in the GG modeling framework. Our objective in this section is to build a

model which combines both of these flexibility features and is consistent in the presence of endogenous regressors.

In Chapter 2 we proposed the GGE model to account for potential endogeneity in the regressors. In that model the μ parameter is defined as $\mu = x_o\beta_o^* + x_e\beta_e^* + x_u\beta_u^*$, where x_o is the observable confounders, x_e is the endogenous regressors and x_u is the unobservable confounders. In Chapter 3 in the GG-IBC model, without endogeneity, this is replaced by $\mu = \ln\left(k\left(x\beta^*,\tau\right)\right)$, where the $k\left(a,\tau\right)$ function is the IBC transformation. In the present chapter since we are accounting for endogeneity we define the μ parameter as a combination of these two and we can rewrite it as

$$\mu = \ln\left(k\left(x_o\beta_o^* + x_e\beta_e^* + x_u\beta_u^*,\tau\right)\right) \tag{4-1}$$

where x_o, x_e and x_u are as defined above. The formulation of the parameter μ given in (4-1) implies the following form for the conditional mean

$$E\left(y \mid x_o, x_e, x_u\right) = \exp\left(\ln\left(k\left(x_o\beta_o^* + x_e\beta_e^* + x_u\beta_u^*,\tau\right)\right) + C\right) \tag{4-2}$$

where C is define as

$$C = \left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right) \tag{4-3}$$

$k\left(a,\tau\right)$ is defined as

$$k(a,\tau) = \begin{cases} \left(\left((\tau/2)a + 1\right)^2\right)^{1/\tau} & \tau \neq 0 \\ \exp\left(a\right) & \tau = 0 \end{cases} \tag{4-4}$$

and τ is a scalar parameter such that $-\infty < \tau < \infty$. The parameter τ controls the conditional mean form and a detailed discussion of the IBC transformation and its special cases can be found in Section 3.2.2.

With this modified conditional mean form, we now proceed as prescribed by the 2SRI method. To correct for the endogeneity which is caused by the presence of the unobservable confounders, we define one, possibly non-linear, auxiliary equation as in equation (2-12) for each of the endogenous variables. Using these equations, we regress the endogenous variables on some or all of the observable regressors and appropriate number of instrumental variables. The estimated residuals (see equation (2-13) for details) from these regressions are used as proxy variables for the unobservable confounders in the main GG regression. The composite model resulting from incorporation of all the aforementioned steps is called GGE-IBC. We have implemented this estimator in Stata/Mata 10® using maximum likelihood estimation via the ml procedure.

4.3 Simulation Analysis

There are three main objectives in this simulation analysis. The first objective is to examine the consistency properties of our model GGE-IBC, as well as those of the NLS-IBC-2SRI and the GGE – the models with which we compare our results. For the class of the GG distributed data, the GGE-IBC estimator should be consistent irrespective of the conditional mean form. Here we seek to verify whether this consistency of GGE-IBC is carried to data distributions not belonging to the GG family of distributions, e.g., the beta distribution. We would accomplish this by observing the behavior of GGE-IBC in experiments with increasing sample sizes – 10,000, 50,000, 100,000 and 250,000. Note that parallel to the argument in Chapter 3, the NLS-IBC-

2SRI model (part of the class of IBC models) should also be consistent for all the data distributional cases due to the use of a flexible conditional mean form. The GGE model, which is a special case of the IBC class of models, is subject to misspecification bias in the conditional mean form because it imposes a particular value of τ ($\tau = 0$) and should be inconsistent when either the data distributional form is not from the GG family or the conditional mean does not have an exponential form. The second objective is to study the bias in the estimated marginal effects of the targeted policy variable obtained using the GGE-IBC estimator and compare it with biases corresponding to the GGE and the NLS-IBC-2SRI based estimators. The latter of these is the NLS-IBC estimator used in Chapter 3 modified to account for endogeneity. The final objective is to observe the precision of the GGE-IBC marginal effect estimator in comparison to the estimators based on GGE and NLS-IBC-2SRI. For all of the above mentioned cases, the data was generated using Monte-Carlo simulations based on various sampling designs. Details of which are provided below.

4.3.1 Sampling Designs

4.3.1.1 The observable and unobservable confounders, instrumental variables and endogenous variable

As mentioned in Section 4.2, in the presence of endogeneity, the regressor vector x is partitioned into three types of variables – the observable confounders, the endogenous variables and the unobservable confounders. In our simulations we use only one variable as the observable confounder. It is generated as a uniformly distributed variate in the interval [0, 4]. We use one variable for the unobservable confounder too and it is generated uniformly in the range [-0.55, 0.55]. To incorporate endogeneity in our simulation analysis, we can define the endogenous variables as a

function of the observable confounders, unobservable confounders and instrumental variables. In this simulation analysis we use only one endogenous regressor (hence only one instrumental variable is also used). For simplicity, the endogenous variable is defined as a linear combination of the observable confounder, the unobservable confounder and the instrumental variable as

$$x_e = x_o\alpha_o + w^+\alpha_{w^+} + x_u, \tag{4-5}$$

where x_e is the endogenous variable, x_o is the observable confounder, w^+ is the instrumental variable and x_u is the unobservable confounder. The coefficients α_o and α_{w^+} are set to 0.3 in the simulation analysis. The instrumental variable w^+ is generated with a uniform distribution in the interval [0, 2]. The simulation analysis is repeated 500 times for four different sample sizes – 10,000, 50,000, 100,000 and 250,000.

4.3.1.2 The outcome variable

The outcome variable y is generated using six different distributions. The linear index $x\beta^*$, which used to generate the conditional means of these distributions, is defined as

$$x\beta^* = x_o\beta_o^* + x_e\beta_e^* + \hat{x}_u\beta_{\hat{u}}^* \tag{4-6}$$

where $\hat{x}_u$ is the residual from the auxiliary equation. The coefficients β_o^* and β_e^* are set to 0.5 and $\beta_{\hat{u}}^*$ is 1.0.

Since the data generation process used here is very similar to that used in Chapter 3, we refer the readers to Section 3.3.1.2 for details. The only difference is the inclusion of the observable confounder (proxied by the first-stage residual – $\hat{x}_u$) and explicit notation distinguishing the three types of regressors (endogenous variable,

observable confounders, and unobservable confounders) in equation (4-6). We have used the standard gamma, log-normal, Weibull, exponential, generalized gamma and beta distributions with $\tau = 1$ in the IBC transformation. The parameter values for κ and σ are the same as those used in Chapter 3.

4.3.2 Estimators to be Evaluated and Compared

The GGE-IBC estimator is a FIML estimator where parameter μ is defined using the IBC transformation, $\mu = \ln\left(k\left(x\beta^*, \tau\right)\right)$. For the estimation of GGE-IBC we implemented a maximum likelihood program using ml command in Stata®. For comparison we included the two other estimators detailed below:

Generalized gamma with endogeneity (GGE): The GGE model was defined in Chapter 2. The GGE model is in the class of exponential conditional mean models where it is assumed that parameter $\mu = x\beta^*$. Although this model has distributional flexibility it lacks the flexibility in the conditional mean. We use this model as one of our comparison models in order to assess the positive consequences of the implementation of the IBC transformation when data is derived from a distribution with non-exponential conditional mean. Note that like the model proposed in this chapter, GGE-IBC, GGE is also a FIML estimator. It is estimated using streg command in Stata®. This model is described in detail in Section 2.2.

Nonlinear least squares with inverse Box-Cox transformation and endogeneity correction (NLS-IBC-2SRI): We have picked a NLS based model for comparison since, in contrast to the GGE model, it is a non-FIML based estimator. As the name suggests, we have modified this estimator to account for both endogeneity

and conditional mean flexibility. The conditional mean function for the NLS-IBC-2SRI is defined as

$$E[y \mid x_o, x_e, x_u] = \exp\left(\ln\left(k\left(x_o\beta_o + x_e\beta_e + x_u\beta_u, \tau\right)\right)\right) \tag{4-7}$$

where $k(x\beta,\tau)$ is given by equation (4-4). The elements of coefficient vector β are a scaled version of the coefficient vector used with the GGE-IBC estimator β^*, where β is scaled by $\exp((\tau/2)C)$, and the constant term is shifted by the expression $(2/\tau)[\exp((\tau/2)C)-1]$.[1] The NLS-IBC-2SRI estimator offers flexibility with respect to the conditional mean, therefore it is robust to misspecification bias, but it offers no correction for skewness since it makes no assumption about the distribution of the outcome variable. Like GGE-IBC, NLS-IBC-2SRI is also a consistent estimator since they both use the 2SRI for endogeneity correction and they are both flexible with respect to the conditional mean. The NLS-IBC-2SRI estimator is implemented using the nl command in Stata®.

4.3.3 Criteria for Evaluation and Comparison

Along the same lines as Chapters 2 and 3, our ultimate objective is to examine the expected marginal effect of a targeted policy variable as given in equation (2-18). In general, the marginal effect of a potentially endogenous policy variable x_e is defined as[2]

$$ME = E\left[\frac{\partial E[y \mid x_o, x_e, x_u]}{\partial x_e}\right]. \tag{4-8}$$

[1] See Appendix D for detailed description of this transformation.

[2] See Terza (2010) for details.

For the GG-IBC model, in which there is no endogeneity problem, the marginal effect as it pertains to the three cases of interest in the GG-IBC framework, is presented in equations (3-12), (3-13) and (3-14). For the GGE-IBC model, in the presence of endogeneity, with the vector of regressors defined as $x = \left[x_o \quad x_e \quad x_u \right]$, the marginal effect expressions analogous to equations (3-12), (3-13) and (3-14) are

$$ME = E\left[\frac{\partial E[y \mid x_o, x_e, x_u]}{\partial x_e} \right] = E\left[\exp\left(x\beta^* + C\left(\sigma,\kappa\right)\right)\beta_e^* \right], \tag{4-9}$$

$$ME = E\left[\frac{\partial E[y \mid x_o, x_e, x_u]}{\partial x_e} \right] = E\left[\beta_e^* \exp\left(C\left(\sigma,\kappa\right)\right) \right], \text{ and,} \tag{4-10}$$

$$ME = E\left[\frac{\partial E[y \mid x_o, x_e, x_u]}{\partial x_e} \right] = E\left[((\tau / 2)x\beta^* + 1)^{(2/\tau)-1}\beta_e^* \exp\left(C\left(\sigma,\kappa\right)\right) \right], \tag{4-11}$$

respectively.

The consistent marginal effect estimators for (4-9), (4-10) and (4-11), are:

$$ME = \hat{\beta}_e^* \left(\frac{\sum_{i=1}^{n} \exp\left(x_i\hat{\beta}^* + C\left(\hat{\sigma},\hat{\kappa}\right)\right)}{n} \right), \tag{4-12}$$

$$ME = \hat{\beta}_e^* \exp\left(C\left(\hat{\sigma},\hat{\kappa}\right)\right), \text{ and,} \tag{4-13}$$

$$ME = \hat{\beta}_e^* \left(\frac{\sum_{i=1}^{n} ((\hat{\tau} / 2)x_i\hat{\beta}^* + 1)^{(2/\hat{\tau})-1} \exp\left(C\left(\hat{\sigma},\hat{\kappa}\right)\right)}{n} \right), \tag{4-14}$$

respectively. Note that these estimators are analogous to (3-15), (3-16) and (3-17) in Chapter 3.

In the NLS-IBC-2SRI framework the appropriate analogues to (4-9), (4-10) and (4-11) are

$$ME = E\left[\exp(x\beta)\beta_e\right], \tag{4-15}$$

$$ME = \beta_e, \text{ and,} \tag{4-16}$$

$$ME = E\left[((\tau/2)x\beta + 1)^{(2/\tau)-1}\beta_e\right], \tag{4-17}$$

and the corresponding consistent estimators are

$$ME = \hat{\beta}_e\left(\frac{\sum_{i=1}^{n}\exp\left(x_i\hat{\beta}\right)}{n}\right), \tag{4-18}$$

$$ME = \hat{\beta}_e, \text{ and,} \tag{4-19}$$

$$ME = \hat{\beta}_e\left(\frac{\sum_{i=1}^{n}((\hat{\tau}/2)x_i\hat{\beta} + 1)^{(2/\hat{\tau})-1}}{n}\right), \tag{4-20}$$

where $\hat{\beta}$ is $\hat{\beta}^*$ scaled by the factor $\exp\left((\tau/2)C(\sigma,\kappa)\right)$, and with the scaled constant term shifted by $(2/\tau)\left[\exp\left((\tau/2)C(\sigma,\kappa)\right)-1\right]$.

In our simulation analysis we are interested in the bias and the precision of the marginal effect estimator of the potentially endogenous policy variable. As in the simulation analyses conducted in Chapters 2 and 3, the metrics for comparison of the bias and precision are the average percentage absolute bias and the mean squared error of the estimated marginal effect, respectively. The former is defined in equation (3-18) and the latter in equation (3-19).

4.3.4 Simulation Results

In this simulation analysis we are interested in the bias and precision performance of the GGE-IBC model as compared to two possible alternative models, GGE and NLS-IBC-2SRI. In Chapter 2 we showed that in the presence of endogeneity the GGE model provides efficiency gains. In Chapter 3 we showed that when the data is derived from a distribution with non-exponential conditional mean, the GG-IBC model is more precise compared to the GG and the NLS-IBC alternatives. Here we investigate whether these efficiency gains prevail in the presence of both endogeneity and differing conditional mean forms when the estimation is carried out using the proposed GGE-IBC model. The metrics for comparison are defined in Section 4.3.3.

In Chapter 3 we showed the consistency of the GG-IBC model. Here our objective is to examine the consistency properties of the GGE-IBC model in the presence of endogeneity. Table 4-1 presents the average percentage absolute bias of the estimated marginal effect of the endogenous variable x_e for different sampling designs for various sample sizes, 10,000, 50,000, 100,000 and 250,000. The percentage biases are calculated for the three quartile values of x_e. For each sampling design we can observe that the average percentage absolute bias from the GGE-IBC estimator decreases as the number of observations in a sample increases. For instance, for the standard gamma distributed data, the average percentage absolute bias for the sample size 10,000 is 9.507% and it decrease to 4.265% for the sample size 50,000 etc. Similarly for all other sampling designs – Weibull, exponential, lognormal, GG, and beta, the average percentage absolute bias for GGE-IBC gets smaller as the number of observations increases. These results are as expected in light of the theoretical

consistency of the GGE-IBC estimator. Although throughout Table 4-1, the average percentage absolute bias results from the NLS-IBC-2SRI model are higher than average percentage absolute bias results from the GGE-IBC model, the findings validate that the NLS-IBC-2SRI model is also a consistent estimator. In the case of standard gamma distributed data, the percentage bias decreases from 14.189% to 6.585% as the sample size increase from 10,000 to 50,000 etc. Similarly the average percentage absolute bias from the GGE model decreases as the sample size increases but at a slower pace. The marginal effect estimates are expected to be biased both in the presence of endogeneity and when the conditional mean functional form is incorrectly specified.

In Chapter 2 we showed that 2SRI endogeneity correction in the GGE model compensates for the first problem – endogeneity – and generates unbiased estimates. In Chapter 3 we showed that the introduction of the IBC transformation in the GG-IBC model rectifies the second problem – conditional mean misspecification bias – and leads to consistent estimates. In this next experiment we examine the bias performance of the GGE-IBC method which simultaneously accounts for skewness, corrects for endogeneity and is robust to conditional mean misspecification bias. All of the alternative estimators included in the simulation comparisons are corrected for endogeneity, so the focus in the first part of the analysis is on the importance of conditional mean flexibility. Both the GGE-IBC model and the NLS-IBC-2SRI model afford this type of flexibility. The GGE estimator does not. Table 4-2 presents the average percentage absolute bias of the marginal effect of x_e for various sampling designs for samples of size 10,000. It can be readily seen that the percentage bias from

the GGE-IBC estimator is consistently lower than the percentage bias from the other estimators. Although the average percentage absolute bias difference between GGE and GGE-IBC is relatively small in the first quartile of x_e, it increases significantly in the second and third quartiles. For example, in the Weibull distributed data, the average percentage absolute bias values from the GGE estimator are 7.961 %, 13.196% and 19.82% for the first, second and third quartile of x_e, respectively, whereas, the corresponding average percentage absolute bias values from the GGE-IBC estimator are 7.132%, 7.619%, and 7.983% – consistently lower. Similar findings can be observed for all of the other sampling designs. It is noteworthy that the average percent absolute bias from the NLS-IBC-2SRI model is consistently higher than that of the GGE-IBC model for all sampling designs. Table 4-2 confirms that the bias performance of the GGE-IBC model is consistently better than the GGE and the NLS-IBC-2SRI estimators.

As mentioned before, the NLS-IBC-2SRI model is not a FIML based model, as a result we would expect efficiency gains from the GGE-IBC model as compared to NLS-IBC-2SRI. In order to verify this, in Table 4-3 we present the mean squared error of the marginal effect of x_e for various sampling designs at a sample size of 10,000. The terms in parentheses represent the percentage efficiency gain from the GGE-IBC model compared to the GGE and the NLS-IBC-2SRI models and it is defined as $\left(\left(MSE_m - MSE_{GGE-IBC}\right)/MSE_m\right)*100$ where m = GGE, NLS-IBC-2SRI. As expected, the mean squared error from the GGE-IBC estimator is lower as compared to its non-FIML counterpart the NLS-IBC-2SRI estimator for all the sampling designs. For example, for the standard gamma distributed data, GGE-IBC provides approximately 54% efficiency gain compared to NLS-IBC-2SRI in every quartile of x_e. This statistical efficiency is

observed for all other sampling designs and the efficiency gain of GGE-IBC compared to NLS-IBC-2SRI ranges from 52% to 77%. Furthermore, GGE-IBC also provides efficiency gains compared to the GGE model where these efficiency gains range from 5.7% to 85% (except for the lognormal distributed data for the first quartile of x_e, the efficiency loss of GGE-IBC compared to GGE is 27%, however, for the second and third quartiles the efficiency gain of GGE-IBC compared to GGE are 61% and 83%, respectively). All in all, the findings confirm that the GGE-IBC model is more efficient compared to the GGE and the NLS-IBC-2SRI models.

Finally, to confirm the consistency of the estimation of the parameter τ in the GGE-IBC and the NLS-IBC-2SRI models and to show the differences in parameter estimates between the GGE and the GGE-IBC we examine the parameter values obtained by these methods. Table 4-4 presents the parameter estimates from various sampling designs for the sample size 10,000. Recall that the parameter τ was fixed to unity. It can be observed that the parameter τ is estimated consistently in both the GGE-IBC and the NLS-IBC-2SRI models. As noted before, the estimates of the elements of β from NLS-IBC-2SRI are scaled (and shifted) versions of the elements of β^*. For the specific cases of the exponential and the standard gamma, where $C(\sigma,\kappa)=0$, this scaling factor disappears. For other distributions used in the sampling designs, the true β values can be calculated using the scaling factor. For example, for the GG distributed sampling design, the true β values are $\beta_o \cong 0.45$, $\beta_p \cong 0.45$, $\beta_{\hat{u}} \cong 0.89$ and the constant $\beta_0 = -0.21$. The parameter estimates from the GG distributed data also confirm the consistency of the NLS-IBC-2SRI model.

4.4 The Effect of Cigarette Smoking on Birthweight in Presence of Endogeneity

Here we demonstrate the usability of the GGE-IBC model for real data in presence of endogeneity. We would also compare the results obtained with the GGE and the NLS-IBC-2SRI methods described above. Our application is the same as the one we used in Section 3.4 of Chapter 3 – the effect of cigarette smoking on birthweight, Mullahy (1997) – but here we take the effect of endogeneity into account. Like Chapter 3, we are interested in the average marginal effect of smoking an additional cigarette per day by the mother on infant birthweight. But here we take into account the potential endogeneity of the cigarette smoking variable due to unobservable confounders like the health consciousness of the mother, which may affect both her smoking and the newborn's birthweight. While correcting for potential endogeneity, we would use paternal schooling, maternal schooling, family income and the per pack state excise tax on cigarettes, as our instrumental variables (the same ones used by Mullahy, 1997). The definitions of all of the variables used in the analysis are given in Table 4-5 and their descriptive statistics are given in Table 4-6. The histogram and the Kernel density of the outcome variable are given in Figure 3-1.

The objectives of this study are threefold. First, we want to verify that the results obtained by our estimator, GGE-IBC, are empirically meaningful, i.e., there is a negative impact of mothers' smoking on infant birthweight. Second, we would like to compare the results obtained in this chapter, where we explicitly take endogeneity into account, with the results obtained in Chapter 3, where presence of endogeneity was ignored. Finally, as in Chapter 3, we would like to examine the practical applicability of the GGE-IBC model as compared to the NLS-IBC-2SRI model.

4.4.1 Model

We first estimate Mullahy's (1997) birthweight model assuming that the conditional mean regression for y follows the IBC formulation given in equation (4-7) of the text. In addition, we assume that

$$x_e = w\alpha + x_u \tag{4-21}$$

where $w = [x_o \quad w^+]$, and w^+ is a vector of identifying instrumental variables – variables that are correlated with x_e, but not with y or x_u. To account for the potential endogeneity of x_e, we implement the 2SRI estimator of Terza et al. (2008a). In the first stage we estimate α by applying ordinary least squares (OLS) to (4-21), and use the result $\hat{\alpha}$ to compute $\hat{x}_u$ as the OLS residual for (4-21). In the second stage we estimate $\beta' = [\beta_e \quad \beta_o \quad \beta_u]$ and τ, the parameters of the model in equation (4-7) of the text, by applying NLS to

$$y = k(x_e\beta_e + x_o\beta_o + \hat{x}_u\beta_u, \tau) + v \tag{4-22}$$

where k(. , .) is defined as in equation (4-4) of the text, v is the random error term and $\hat{x}_u$ is the first-stage residual.

We use the parameter estimates obtained via NLS-IBC-2SRI to estimate the marginal effect (ME) of cigarette smoking during pregnancy on infant birthweight as

$$ME = E\left[\frac{\partial k(x_e\beta_e + x_o\beta_o + x_u\beta_u, \tau)}{\partial x_e}\right] = E\left[\frac{\partial \psi^{(2/\tau)}}{\partial x_e}\right] \tag{4-23}$$

where

$$k(x\beta, \alpha, \tau) = \psi^{(2/\tau)}, \tag{4-24}$$

$$\psi = \frac{\tau}{2}(x\beta) + 1 \tag{4-25}$$

and

$$x\beta = x_e\beta_e + x_o\beta_o + (x_e - w\hat{\alpha})\beta_u \tag{4-26}$$

where the x vector is partitioned as x = [x_e x_o x_u] and x_u is substituted by $(x_e - w\hat{\alpha})$. Therefore, the ME can be written as

$$ME = E\left[\frac{2}{\tau}\psi^{\left(\frac{2}{\tau}-1\right)}\left(\frac{\tau}{2}\beta_e\right)\right] = E\left[\left(\psi^2\right)^{\left(\frac{1}{\tau}-\frac{1}{2}\right)}\beta_e\right]. \tag{4-27}$$

The estimator of the ME of cigarette smoking is the following sample analog to (4-27) is given by

$$ME = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{\left(\hat{\psi}_i^2\right)^{\left(\frac{1}{\hat{\tau}}-\frac{1}{2}\right)}\hat{\beta}_e\right\} \tag{4-28}$$

where $\hat{\psi}_i = \frac{\hat{\tau}}{2}\left(x_{ei}\hat{\beta}_e + x_{oi}\hat{\beta}_o + \hat{x}_{ui}\hat{\beta}_u\right) + 1$, n_s is the size of the subsample of smokers, and the "^s" denote estimates. The ME is estimated for smokers only.

The average effect of smoking cessation on birthweight for the mothers who smoked during pregnancy for the NLS-IBC-2SRI model is defined as

$$CE = E\left[k(x_e\beta_e + x_o\beta_o + x_u\beta_u, \tau) - k(x_o\beta_o + x_u\beta_u, \tau)\right] \tag{4-29}$$

which can be estimated as

$$CE = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{k(x_{ei}\hat{\beta}_e + x_{oi}\hat{\beta}_o + \hat{x}_{ui}\hat{\beta}_u, \hat{\tau}) - k(x_{oi}\hat{\beta}_o + \hat{x}_{ui}\hat{\beta}_u, \hat{\tau})\right\} \tag{4-30}$$

where n_s is the size of the subsample of smokers.

Next, we estimate Mullahy's (1997) birthweight model assuming that the conditional density function of y follows the GGE-IBC model (Section 4-2) formulation given in equation (4-2) of the text. In addition, we assume that the auxiliary equation (4-21) holds. Here again, to account for the potential endogeneity of x_e, we implement the 2SRI estimator of Terza et al. (2008a). In the first stage, like for NLS-IBC-2SRI, we

estimate α by applying ordinary least squares (OLS) to (4-21), and use the result $\hat{\alpha}$ to compute $\hat{x}_u$ as the OLS residual for (4-21). In the second stage we estimate β' = [β_e β_o β_u], σ, κ and τ, the parameters of the model in equation (4-2) of the text, by applying the GGE-IBC described above with $\hat{x}_u$ included as an additional regressor.

We use the parameter estimates obtained via GGE-IBC to estimate the ME of cigarette smoking during pregnancy on infant birthweight as

$$ME = E\left[\frac{\partial k(x_e\beta_e^* + x_o\beta_o^* + x_u\beta_u^*, \tau)\times \exp(C)}{\partial x_e}\right] = E\left[\frac{\partial\left(\psi^{*(2/\tau)}\exp(C)\right)}{\partial x_e}\right] \quad (4\text{-}31)$$

where

$$k(x_e\beta_e^* + x_o\beta_o^* + x_u\beta_u^*, \tau) = \psi^{*(2/\tau)}, \quad (4\text{-}32)$$

$$\psi^* = \frac{\tau}{2}(x\beta^*) + 1, \text{ and,} \quad (4\text{-}33)$$

$$x\beta^* = x_e\beta_e^* + x_o\beta_o^* + (x_e - w\hat{\alpha})\beta_u^* \quad (4\text{-}34)$$

where the x vector is partitioned as x = [x_e x_o x_u] and the x_u is substituted by $(x_e - w\hat{\alpha})$ and C is defined as

$$C = \left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right). \quad (4\text{-}35)$$

Therefore, ME is given as

$$ME = E\left[\frac{2}{\tau}\psi^{*\left(\frac{2}{\tau} - 1\right)}\left(\frac{\gamma}{2}\beta_e^*\right)\exp(C)\right] = E\left[\left(\psi^{*2}\right)^{\left(\frac{1}{\tau} - \frac{1}{2}\right)}\beta_e^*\exp(C)\right]. \quad (4\text{-}36)$$

The estimator of the ME of cigarette smoking is the following sample analog to (4-36)

$$ME = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{\left(\hat{\psi}_i^{*2}\right)^{\left(\frac{1}{\hat{\tau}} - \frac{1}{2}\right)}\hat{\beta}_e^*\exp(\hat{C})\right\} \quad (4\text{-}37)$$

where

$$\hat{\psi}_i^* = \frac{\hat{\tau}}{2}\left(x_{ei}\hat{\beta}_e^* + x_{oi}\hat{\beta}_o^* + \hat{x}_{ui}\hat{\beta}_u^*\right) + 1, \tag{4-38}$$

$$\hat{C} = \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\ln\left(\hat{\kappa}^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\hat{\kappa}^2}\right) + \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\hat{\kappa}^2}\right\}\right), \tag{4-39}$$

and the "^s" denote estimates.

For the GGE-IBC model, the cessation effect is given by

$$CE = E\left[\left(k(x_e\beta_e^* + x_o\beta_o^* + x_u\beta_u^*, \tau) - k(x_o\beta_o^* + x_u\beta_u^*, \tau)\right)\exp(C)\right] \tag{4-40}$$

which can be estimated as

$$CE = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{\left(k(x_{ei}\hat{\beta}_e^* + x_{oi}\hat{\beta}_o^* + \hat{x}_{ui}\hat{\beta}_u^*, \tau) - k(x_{oi}\hat{\beta}_o^* + \hat{x}_{ui}\hat{\beta}_u^*, \tau)\right)\exp(\hat{C})\right\} \tag{4-41}$$

where n_s is the size of the subsample of smokers.

We will also use the GGE model, defined by us in Chapter 2, to estimate the ME. The ME and the ME estimator of the cigarette smoking for the GGE estimator are given in equations (2-25) and (2-26), respectively. The cessation effect and its estimator for the same model are given by the following equations, respectively:

$$CE = E\left[\left(\exp(x_e\beta_e^* + x_o\beta_o^* + x_u\beta_u^*) - \exp(x_o\beta_o^* + x_u\beta_u^*)\right)\exp(C)\right], \tag{4-42}$$

$$CE = \sum_{i=1}^{n_s}\frac{1}{n_s}\left\{\left(\exp(x_{ei}\hat{\beta}_e^* + x_{oi}\hat{\beta}_o^* + \hat{x}_{ui}\hat{\beta}_u^*) - \exp(x_{oi}\hat{\beta}_o^* + \hat{x}_{ui}\hat{\beta}_u^*)\right)\exp(\hat{C})\right\}. \tag{4-43}$$

4.4.2 Results

Table 4-7 summarizes the results from the GGE-IBC, the NLS-IBC-2SRI and the GGE estimators described above. The marginal effects as defined in equations (4-28), (4-37) and (2-26) for the NLS-IBC-2SRI, the GGE-IBC and the GGE, respectively, restricting the sample (after estimation) to those who smoked during pregnancy (i.e.,

CIGSPREG > 0) – this subset had 212 individuals, are presented in the first row of Table 4-7. The average effect of smoking cessation on birthweight for these 212 mothers who smoked during pregnancy, as captured by the expressions in equations (4-30), (4-41) and (4-43) for the NLS-IBC-2SRI, the GGE-IBC and the GGE models, respectively, are presented in the second row of Table 4-7.

The first observation that can be made from this table is that the estimates of ME obtained for all the models are negative, indicating that the birthweight decreases for each additional cigarette smoked per day during pregnancy. For our GGE-IBC model, the weight of a newborn goes down by 1.143 ounces for each additional cigarette smoked per day. Recall that the same figure for the GG-IBC model in Chapter 3, which does not correct for endogeneity, was estimated to be 0.489 ounces. This marked difference in these values can be attributed to the relatively unbiased estimates obtained by the GGE-IBC, a model which accounts for endogeneity. We have shown through extensive simulations in Chapter 2 that models that correct for endogeneity using 2SRI are unbiased compared to those which do not.

Next, from the cessation effect given in the second row of Table 4-7, it can be noted that if all the smokers quit smoking with immediate effect, the birthweight goes up. For the GGE-IBC method, we find this increase to be 18.491 ounces. The same figure for the GG-IBC model in Chapter 3 was obtained to be 6.913 ounces. We expect the results obtained from the GGE-IBC model to be unbiased compared those obtained from GG-IBC because of the same reason as described above.

As in Chapter 3, we would again like to point out that from a practical applicability point of view, GGE-IBC behaves better than the NLS-IBC-2SRI model. The later model

was found to have convergence issues in cases with small number of observations, like our real data application.

4.5 Summary, Discussion and Conclusion

In Chapter 2 we address the two common problems seen in health outcomes data, skewness and endogeneity. In Chapter 3 we take into account another important issue that plagues health econometric models, the bias that surfaces with the misspecification of the conditional mean. Here we are interested in providing solutions to all of these problems at the same time by proposing a novel model. The GGE-IBC model offers a simultaneous remedy for all of these problems since it is a FIML estimator that has a flexible distributional form, flexible conditional mean form and a built in mechanism for endogeneity correction.

Like Chapters 2 and 3, we first tested our model via extensive simulation analyses and then on a real data application. In our simulation analyses, we used datasets of various sizes generated using various distributions. We compared our results with those obtained from popular alternative models like the GGE and the NLS-IBC-2SRI – both detailed above. Through our simulation analysis we found that the GGE-IBC model is consistent and does not suffer from the conditional mean misspecification bias. We empirically show that our model unfailingly generates lower average percentage absolute bias for the marginal effect estimates compared to the GGE and the NLS-IBC-2SRI models. Like the other estimators introduced by us in this dissertation, the GGE-IBC model retains the precision properties of the GG model because it is a FIML-based estimator.

Finally, we tested the proposed GGE-IBC method by revisiting Mullahy's (1997) model of the effect of mothers' daily cigarette smoking during pregnancy on infant

birthweight. Note that this is same dataset we used in Chapter 3 but here our model explicitly recognizes and corrects for endogeneity while this was not the case in Chapter 3. We found that the ME of smoking obtained when endogeneity is corrected for in the GGE-IBC is twice that of the ME of smoking obtained by the GG-IBC model. The conclusions drawn in Chapter 3 about the practical usefulness of the GG-IBC model relative to the NLS-IBC model remains true for the GGE-IBC model as compared to the NLS-IBC-2SRI model.

Table 4-1. For various sample sizes, average percentage absolute bias of the marginal effect ($\tau = 1$) (in percentages)

Data	Estimator	N	1st quartile of x_e	2nd quartile of x_e	3rd quartile of x_e
Gamma $\gamma = 2.0$	GGE	10,000	10.338	15.655	22.167
		50,000	4.196	7.895	14.030
		100,000	3.876	9.323	16.294
		250,000	2.958	8.678	15.419
	NLS-IBC-2SRI	10,000	14.189	15.051	15.784
		50,000	6.585	6.966	7.323
		100,000	4.570	4.878	5.151
		250,000	2.885	3.045	3.192
	GGE-IBC	10,000	9.508	10.146	10.676
		50,000	4.265	4.524	4.763
		100,000	2.916	3.119	3.292
		250,000	1.854	1.965	2.071
Weibull $\sigma = 0.5$	GGE	10,000	7.961	13.296	19.820
		50,000	3.150	6.802	13.182
		100,000	2.691	8.058	14.769
		250,000	1.996	0.764	14.095
	NLS-IBC-2SRI	10,000	10.809	11.475	12.031
		50,000	5.017	5.301	5.558
		100,000	3.511	3.720	3.902
		250,000	2.073	2.174	2.268
	GGE-IBC	10,000	7.132	7.620	7.983
		50,000	3.175	3.334	3.470
		100,000	2.385	2.539	2.660
		250,000	0.014	0.014	0.015
Exponential	GGE	10,000	14.473	20.159	26.897
		50,000	7.436	10.484	16.171
		100,000	7.757	11.714	18.824
		250,000	6.682	10.931	17.769
	NLS-IBC-2SRI	10,000	20.107	21.434	22.588
		50,000	9.420	9.960	10.463
		100,000	6.721	7.118	7.471
		250,000	4.056	4.281	4.485
	GGE-IBC	10,000	14.186	15.155	15.999
		50,000	6.757	7.127	7.459
		100,000	5.141	5.449	5.714
		250,000	2.795	2.973	3.134

Table 4-1. Continued

Data	Estimator	N	1st quartile of x_e	2nd quartile of x_e	3rd quartile of x_e
Log-normal $\sigma_\varepsilon = 0.5$	GGE	10,000	7.932	13.989	21.124
		50,000	4.587	10.941	22.347
		100,000	4.631	11.501	19.531
		250,000	8.133	19.785	37.046
	NLS-IBC-2SRI	10,000	10.048	10.692	11.233
		50,000	5.169	5.470	5.747
		100,000	3.590	3.813	4.008
		250,000	2.226	2.350	2.465
	GGE-IBC	10,000	8.125	7.995	7.828
		50,000	3.643	3.853	4.089
		100,000	2.669	2.681	2.690
		250,000	2.018	1.819	1.803
Generalized gamma $\kappa / \sigma = 2.0$ $\gamma = 0.5$	GGE	10,000	9.588	15.009	21.641
		50,000	4.165	7.236	13.006
		100,000	3.102	7.970	14.627
		250,000	2.115	7.291	13.731
	NLS-IBC-2SRI	10,000	14.702	15.697	16.557
		50,000	7.046	7.464	7.855
		100,000	4.890	5.202	5.477
		250,000	2.956	3.133	3.295
	GGE-IBC	10,000	9.419	10.065	10.576
		50,000	4.457	4.725	4.970
		100,000	3.069	3.276	3.454
		250,000	1.943	2.057	2.155
Beta $\alpha = 2,\ \gamma = 2$	GGE	10,000	4.966	10.272	16.877
		50,000	1.960	5.089	11.287
		100,000	1.364	6.605	13.043
		250,000	0.879	6.045	12.217
	NLS-IBC-2SRI	10,000	9.460	10.059	10.551
		50,000	4.195	4.416	4.622
		100,000	2.928	3.135	3.313
		250,000	1.874	0.020	2.096
	GGE-IBC	10,000	4.517	4.965	5.354
		50,000	1.787	1.893	1.994
		100,000	1.888	2.031	2.117
		250,000	0.947	0.992	1.005

Table 4-2. For various sampling designs, average percentage absolute bias of the marginal effect ($\tau = 1$) (N=10,000) (in percentages)

Data	Estimator	1st quartile of x_e	2nd quartile of x_e	3rd quartile of x_e
Gamma	GGE	10.338	15.655	22.167
$\gamma = 2.0$	NLS-IBC-2SRI	14.189	15.051	15.784
	GGE-IBC	9.508	10.146	10.676
Weibull	GGE	7.961	13.296	19.820
$\sigma = 0.5$	NLS-IBC-2SRI	10.809	11.475	12.031
	GGE-IBC	7.132	7.620	7.983
Exponential	GGE	14.473	20.159	26.897
	NLS-IBC-2SRI	20.107	21.434	22.588
	GGE-IBC	14.186	15.155	15.999
Log-normal	GGE	7.932	13.989	21.124
$\sigma_\varepsilon = 0.5$	NLS-IBC-2SRI	10.048	10.692	11.233
	GGE-IBC	8.125	7.995	7.828
Generalized gamma	GGE	9.588	15.009	21.641
$\kappa / \sigma = 2.0$	NLS-IBC-2SRI	14.702	15.697	16.557
$\gamma = 0.5$	GGE-IBC	9.419	10.065	10.576
Beta	GGE	4.966	10.272	16.877
$\alpha = 2,\ \gamma = 2$	NLS-IBC-2SRI	9.460	10.059	10.551
	GGE-IBC	4.517	4.965	5.354

Table 4-3. For various sampling designs mean squared error of the marginal effect with percent relative efficiency gain ($\tau = 1$) (N=10,000)

Data	Estimator	1st quartile of x_e	2nd quartile of x_e	3rd quartile of x_e
Gamma	GGE	0.01408	0.0329229	0.0683225
$\gamma = 2.0$		(32.5%)	(63.3%)	(78.3%)
	NLS-IBC-2SRI	0.0210954	0.0265183	0.0324618
		(54.9%)	(54.4%)	(54.2%)
	GGE-IBC	0.0095073	0.0120842	0.0148529
Weibull	GGE	0.005123	0.0143617	0.0329675
$\sigma = 0.5$		(20.3%)	(63.9%)	(80.8%)
	NLS-IBC-2SRI	0.0092588	0.011632	0.0141949
		(55.9%)	(55.5%)	(55.4%)
	GGE-IBC	0.0040854	0.0051798	0.0063262
Exponential	GGE	0.0287104	0.0541753	0.1045475
		(11.1%)	(41.3%)	(63.3%)
	NLS-IBC-2SRI	0.0416459	0.0529894	0.0656875
		(38.7%)	(40.0%)	(41.5%)
	GGE-IBC	0.0255283	0.0318172	0.0384056
Log-normal	GGE	0.0079024	0.0249051	0.0584094
$\sigma_\varepsilon = 0.2$		(-27.1%)	(61.2%)	(83.4%)
	NLS-IBC-2SRI	0.0137877	0.0174717	0.0214402
		(27.1%)	(44.7%)	(54.7%)
	GGE-IBC	0.0100471	0.0096598	0.0097130
Generalized	GGE	0.0061122	0.0156814	0.034095
gamma		(5.7%)	(53.1%)	(73.5%)
$\kappa / \sigma = 2.0$	NLS-IBC-2SRI	0.0148921	0.0189276	0.0233592
$\gamma = 0.5$		(61.3%)	(61.1%)	(61.3%)
	GGE-IBC	0.0057623	0.0073595	0.0090409
Beta	GGE	0.000633	0.0025785	0.0069139
$\alpha = 2,\ \gamma = 2$		(14.3%)	(70.6%)	(85.2%)
	NLS-IBC-2SRI	0.0024413	0.0030791	0.003761
		(77.8%)	(75.4%)	(72.8%)
	GGE-IBC	0.0005428	0.0007583	0.0010232

Note: The values given in parentheses are the percent relative efficiency gains that measure the relative efficiency of GGE-IBC estimator compared to the GGE and the NLS-IBC-2SRI estimators. The percent relative efficiency gain is defined by $\left(\left(\mathrm{MSE}_m - \mathrm{MSE}_{\mathrm{GGE-IBC}}\right) / \mathrm{MSE}_m\right) * 100$, where m = GGE, NLS-IBC-2SRI.

Table 4-4. Parameter estimates (N=10,000) ($\tau = 1$)

Data	Parameters	GGE	NLS-IBC-2SRI	GGE-IBC
Gamma $\gamma = 2.0$	$\beta_p = 0.5$	0.3241	0.5062	0.5088
	$\beta_o = 0.5$	0.3033	0.4997	0.5005
	$\beta_{\hat{u}} = 1.0$	0.6087	0.9909	0.9912
	$\beta_0 = 0$	0.1168	-0.0038	-0.0068
	$\kappa \cong 0.71$	0.6952	-	0.7052
	$\ln(\sigma) \cong$ -0.35	-0.3387	-	-0.3471
	$\tau = 1$	-	0.9963	1.0034
Weibull $\sigma = 0.5$	$\beta_p = 0.5$	0.3220	0.4767	0.5108
	$\beta_o = 0.5$	0.3009	0.4687	0.4966
	$\beta_{\hat{u}} = 1.0$	0.6021	0.9267	0.9822
	$\beta_0 = 0$	0.1209	-0.1209	-0.0058
	$\kappa = 1$	0.9716	-	1.0031
	$\ln(\sigma) \cong -0.69$	-0.6764	-	-0.6999
	$\tau = 1$	-	0.9910	0.9921
Exponential	$\beta_p = 0.5$	0.3263	0.5053	0.5146
	$\beta_o = 0.5$	0.2997	0.4954	0.4946
	$\beta_{\hat{u}} = 1.0$	0.5987	0.9797	0.9869
	$\beta_0 = 0$	0.1196	-0.0018	0.0008
	$\kappa = 1$	0.9915	-	0.9999
	$\ln(\sigma) = 0$	0.0007	-	-0.0026
	$\tau = 1$	-	0.9797	1.0055
Log-normal $\sigma_\varepsilon = 0.2$	$\beta_p = 0.5$	0.3262	0.5430	0.5494
	$\beta_o = 0.5$	0.3023	0.5314	0.5108
	$\beta_{\hat{u}} = 1.0$	0.6058	1.0514	0.9966
	$\beta_0 = 0$	0.1225	0.1216	-0.0152
	$\kappa \to 0$	0.0102	-	-0.0073
	$\ln(\sigma) \cong -0.69$	-0.6765	-	-0.6911
	$\tau = 1$	-	0.9958	1.0828

Table 4-4. Continued

Data	Parameters	GGE	NLS-IBC-2SRI	GGE-IBC
Generalized gamma $\kappa/\sigma = 2.0$ $\gamma = 0.5$	$\beta_p = 0.5$	0.3251	0.4609	0.5171
	$\beta_o = 0.5$	0.3005	0.4450	0.4981
	$\beta_{\hat{u}} = 1.0$	0.6017	0.8818	0.9854
	$\beta_0 = 0$	0.1180	-0.2197	-0.0076
	$\kappa \cong 1.41$	1.3879	-	1.4169
	$\ln(\sigma) \cong -0.35$	-0.3312	-	-0.3481
	$\tau = 1$	-	1.0032	1.0025
Beta $\alpha = 2,\ \gamma = 2$	$\beta_p = 0.5$	0.3153	0.3597	0.4197
	$\beta_o = 0.5$	0.3005	0.3536	0.4071
	$\beta_{\hat{u}} = 1.0$	0.6010	0.7011	0.8083
	$\beta_0 = 0$	-0.2931	-0.5909	-0.3653
	κ	1.7565	-	1.9708
	$\ln(\sigma)$	-1.0751	-	-1.1711
	$\tau = 1$	-	0.9981	1.0528

Table 4-5. The variable definitions from the birthweight analysis

Variable	Definition
The outcome variable	
BIRTHWEIGHT	the infant's birthweight measured in lbs.
Policy Variable (x_e)	
CIGSPREG	the number of cigarettes smoked per day during pregnancy
The observable confounders (x_o)	
PARITY	the birth order
WHITE	= 1 if white, 0 otherwise
MALE	= 1 if male, 0 otherwise
Instrumental Variables (w^+)	
EDFATHER	paternal schooling in years
EDMOTHER	maternal schooling in years
FAMINCOM	family income ($\times 0.001$)
CIGTAX88	per pack state excise tax on cigarettes

Table 4-6. Descriptive statistics for the birthweight sample (N=1,388)

Variable	Mean	Minimum	Maximum
The outcome variable			
BIRTHWEIGHT	7.42	1.44	16.94
Policy variable (x_e)			
CIGSPREG	2.09	0	50
The observable confounders (x_o)			
PARITY	1.63	1	6
WHITE	%78	-	-
MALE	%52	-	-
Instrumental variables (w^+)			
EDFATHER	11.32	0	18
EDMOTHER	12.93	0	18
FAMINCOM	29.03	0.5	65
CIGTAX88	19.55	2	38

Table 4-7. The 2SRI marginal effect and cessation effect estimates from the birthweight analysis

	Type of Estimation		
Marginal effect	NLS-IBC-2SRI	GGE-IBC	GGE
Average Marginal Effect	-1.1684	-1.1426	-1.1094
Cessation Effect	18.0042	18.4914	16.3405

CHAPTER 5
CONCLUSION

5.1 Summary

In this dissertation we have proposed three novel estimation techniques for modeling health outcomes data that effectively address commonly encountered issues in health economics datasets – the generalized gamma with endogeneity (GGE) (Chapter 2), the generalized gamma with inverse Box-Cox transformation (GG-IBC) (Chapter 3) and the generalized gamma with endogeneity and inverse Box-Cox transformation (GGE-IBC) (Chapter 4). In addition to the fact that all of these newly proposed estimators are designed to deal with skewness, the characteristic issues that we focus on are the endogeneity of the regressors and the possibility of the data being derived from a distribution with a non-exponential conditional mean form. The former of these, endogeneity, is common since regression models are often influenced by unobservable confounding effects. The latter is a problem with models that misspecifies their conditional mean form and can lead to undesirable effects.

Our corrective steps to rectify for the aforementioned problems are built upon the generalized gamma (GG) model. We have chosen the GG model as the base for our models since it is a full-information maximum likelihood (FIML) estimator that provides all important distributional form flexibility that is quite effective in addressing skewness in health outcomes data. Furthermore, we demonstrate that with GG model as their foundation, our models take advantage of its corresponding efficiency properties.

We present our first model, GGE, in Chapter 2 where we address the issue of endogeneity. We propose the use of two stage residual inclusion (2SRI) method with the GG model to side step the effects of endogenous regressors due to unobservable

confounders. This new model simultaneously provides robustness against both skewness and endogeneity, which to the best of our knowledge, have not been addressed together in the literature. We thoroughly evaluate the new model using extensive simulation analysis as well as real data analyses. Note that in of all our analyses presented in this dissertation, we are interested in observing the marginal effect (ME) of a targeted policy variable on the outcome variable. In our simulation analysis we compared the performance of GGE with two popular alternative models – OLS and GLM with log link function – using samples generated using various distributions and with various sample sizes. Quite interestingly, we found that though OLS and GLM models can provide low average percentage absolute bias, a metric that captures bias in the computed ME, for some cases, GGE consistently provides better results across all datasets. We also found that another benefit of using a model like the GGE with flexible distributional form is that it can be used to detect the inherent distribution underlying a given dataset. This was quite handy in our real data application, (which looks at the relationship between hospital expenditures and prescription drug usage) where the GGE model, through a Wald test, detected that the dependent variable is distributed according to the log-normal distribution.

In Chapter 3 we build a model that can accommodate a flexible conditional mean form. We accomplish this by introducing the inverse Box-Cox (IBC) transformation in the conditional mean expression of the GG model. With this modification the resultant model, GG-IBC has both distributional form flexibility as well as conditional mean form flexibility. We tested the proposed model through extensive simulation and real data analyses and compared the obtained results with the NLS-IBC and the GG models. We

found that the GG-IBC model consistently produces lower average percentage absolute bias as compared to the NLS-IBC and the GG models across different sampling distributions and sample sizes, when the data is drawn from distributions with non-exponential conditional mean forms. The application of GG-IBC to the real data application which examined the impact of smoking by pregnant women on their newborn's birthweight shows that the former has a negative effect on the latter. Quite importantly, through this real data analysis, we found that the GG-IBC model is well behaved in its convergence properties even when the dataset is small. This is at contrast to the NLS-IBC model, which has considerable convergence problems in this dataset.

Finally, in Chapter 4, we proposed the composite model GGE-IBC, which combines the two previously proposed models – GGE and GG-IBC – into a single versatile model. The GGE-IBC model, with its flexible distributional form, flexible conditional mean form and built-in mechanism to handle endogeneity, provides all the benefits previously summarized for the GGE and the GG-IBC models simultaneously. Through simulation analysis we showed that the GGE-IBC model maintains the efficiency properties of the original GG model while at the same time providing lower average percentage absolute bias as compared to alternatives such as NLS-IBC-2SRI. Note that the NLS-IBC-2SRI model is derived from NLS by appropriately modifying its conditional mean form with the IBC transformation and placing it in the 2SRI framework for handling endogeneity. In the real data experiments using GGE-IBC, we revisited the smoking-birthweight application used in Chapter 3, but this time we accounted for endogeneity in addition to allowing a flexible conditional mean in our estimation. We

found that the ME of smoking during pregnancy with GGE-IBC is twice the value obtained in Chapter 3, when endogeneity is not accounted for.

In summary, we have presented new models for health outcome data estimation based on the GG model, which provide very attractive alternatives to prevailing econometric models that are variants of conventional methods such as OLS, GLM and NLS models.

5.2 Limitations and Future Work

The research presented in this dissertation is limited in a number of respects and, therefore, there are a number of natural extensions of this work that we intend to pursue in the future. Following is a listing of a few of these possible extensions.

Foremost, in Chapter 2, we used the 2SRI model to correct for endogeneity in our GGE model. Since this is a two stage model, the correct standard errors for the parameter estimates could not be obtained directly from Stata®. Because of this we derived the expressions for the correct standard errors for parameter estimates for GGE in Appendix C. Since we were also interested in the marginal effects of the policy variables, we also derived the expressions for the correct standard errors for the marginal effects too in the same Appendix. In Chapter 3, since we did not make use of a two stage model and since all the parameters of our model, GG-IBC, were comprehensively estimated using a maximum likelihood routine, we were able to obtain the standard errors of our parameter estimates directly from Stata®. However, as before in the case of GG-IBC, we are interested in finding the correct standard errors for the marginal effects. Deriving the expressions for these correct standard errors is something to be completed as future research. Similarly, in Chapter 4, since our model, GGE-IBC, is a two stage model, it still remains to derive the expression for the correct

standard errors for the parameter estimates and the marginal effect. For the parameter estimates, we would base our derivations on asymptotic results given in White (1994) and Newey and McFadden (1994). For the case of the marginal effect, we would base our derivation on Terza (2010).

Secondly, for any given estimator, both large and small sample size analyses are of interest. This is so since the large sample analysis help validate the theoretical asymptotic properties of the estimator while the small sample size analysis provide a better understanding of the estimator's bias and precision behavior in some limited real data scenarios. In Chapter 2, for our GGE model, we presented both of these analyses in our simulation study. But for Chapter 3 and 4, where we presented the GG-IBC and the GGE-IBC models, although we presented results using real data with a small sample size, we did not include small sample size analysis in our simulation studies. Since the results from our real data application are encouraging, in the future we would like to explore the small data size properties of our models more rigorously via simulation studies.

Thirdly, in this dissertation, in both the simulations and the real data analyses, we have observed differences among the parameter estimates across the models, but we have not determined if such differences are significant in a statistical sense. A topic for future research would be the development of Hausman-Type statistical tests for this purpose.

Fourthly, the properties of the GG estimator and the GG based estimators presented in this dissertation were rigorously studied via detailed simulation for cases in which the data is distributed according to members of the GG family. Our study of data

distributions outside the GG family was limited to the beta distribution in Chapters 3 and 4. Clearly, there are many distributions that do not belong to the GG family which could be of interest. In future, we would like to explore this dimension in more detail.

Fifthly, one of the important benefits of using the GG and GG based models is that these can be used to conduct nested model tests, wherein the GG based estimators can detect which of the relevant special cases (Weibull, exponential, standard gamma and log normal) describe data the best. We detailed the performance of such a test using the GGE model in Chapter 2. In the future, we would like to explore similar nested model test applications in the GG-IBC and GGE-IBC contexts.

Sixthly, as part of this dissertation, we have developed software for our proposed estimators – GGE, GG-IBC and GGE-IBC using Stata 10®. We would like to make these available to the research community at large. For this reason, we intend to develop Stata® "ado" files for these estimators in near future.

Finally, we have explored two different real datasets as target applications for our proposed models. In the future, we plan to continue to search for good applications for our methods in empirical health economics and elsewhere.

APPENDIX A
THE FORMAL DERIVATION OF THE REPARAMETRIZATION OF THE CONDITIONAL MEAN

Standard gamma: The constant term in the conditional mean is defined in equation (2-4)

$$C(\sigma,\kappa)=\left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right)+\ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right)+\left(\frac{\sigma}{\kappa}\right)\right\}\right)-\ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right). \tag{2-4}$$

In standard gamma the shape parameters are equal, i.e., $\kappa=\sigma$. Substituting $\kappa=\sigma$ in (2-4)

$$C(\sigma,\sigma)=\underbrace{\left(\frac{\sigma}{\sigma}\right)}_{=1}\ln\left(\sigma^2\right)+\ln\left(\Gamma\left\{\left(\frac{1}{\sigma^2}\right)+\underbrace{\left(\frac{\sigma}{\sigma}\right)}_{=1}\right\}\right)-\ln\left(\Gamma\left\{\frac{1}{\sigma^2}\right\}\right) \tag{A-1}$$

$$C(\sigma,\sigma)=\ln\left(\sigma^2\right)+\ln\left(\Gamma\left\{\left(\frac{1}{\sigma^2}\right)+1\right\}\right)-\ln\left(\Gamma\left\{\frac{1}{\sigma^2}\right\}\right) \tag{A-2}$$

Note that $\Gamma(z+1)=z\Gamma(z)$. As a result we can write $\Gamma\left(\frac{1}{\sigma^2}+1\right)=\frac{1}{\sigma^2}\Gamma\left(\frac{1}{\sigma^2}\right)$ and substitute in (A-2)

$$C(\sigma,\sigma)=\ln\left(\sigma^2\right)+\ln\left(\frac{1}{\sigma^2}\Gamma\left\{\frac{1}{\sigma^2}\right\}\right)-\ln\left(\Gamma\left\{\frac{1}{\sigma^2}\right\}\right) \tag{A-3}$$

$$C(\sigma,\sigma)=\underbrace{\ln\left(\sigma^2\right)+\ln\left(\frac{1}{\sigma^2}\right)}_{=0}+\underbrace{\ln\left(\Gamma\left\{\frac{1}{\sigma^2}\right\}\right)-\ln\left(\Gamma\left\{\frac{1}{\sigma^2}\right\}\right)}_{=0}=0. \tag{A-4}$$

We derived that the constant term in the conditional mean for standard gamma is equal to zero.

Weibull: In the Weibull distribution the shape parameter κ is equal to unity. Substituting $\kappa=1$ into equation (2-4) we get

$$C(\sigma,1) = (\sigma)\underbrace{\ln(1)}_{=0} + \ln\left(\Gamma\{(1)+(\sigma)\}\right) - \underbrace{\ln\left(\Gamma\{1\}\right)}_{=0} \quad \text{(A-5)}$$

The constant term is defined as

$$C(\sigma, 1) = \ln\left(\Gamma(1+\sigma)\right). \quad \text{(A-6)}$$

Exponential: Exponential distribution is a special case of standard gamma distribution where the parameters are set equal to $\kappa = \sigma = 1$. As a result the constant term is $C(1,1) = 0$.

APPENDIX B
THE DERIVATION OF MARGINAL EFFECT OF THE ENDOGENOUS POLICY VARIABLE

The marginal effect of the endogenous policy variable is defined as

$$ME = E\left[\frac{\partial\{\Phi(x\beta_1)\exp(x\beta_2)\}}{\partial x_e}\right] \quad \text{(B-1)}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\beta_1 = [\beta_{o1} \quad \beta_{e1} \quad \beta_{u1}]$, and $\beta_2^* = [\beta_{o2}^* \quad \beta_{e2}^* \quad \beta_{u2}^*]' - \beta_2$ is the same as β_2^* except for the constant term $\beta_{o21} = \beta_{o21}^* + C(\sigma,\kappa)$, where $C(\sigma,\kappa) = \left[\left(\frac{\sigma}{\kappa}\right)\ln(\kappa^2) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right)\right]$.

According to (B-1), the marginal effect would be

$$ME = E\left[\beta_{e1}\varphi(x\beta_1)\exp\left(x\beta_2^* + C(\sigma,\kappa)\right) + \beta_{e2}^*\exp\left(x\beta_2^* + C(\sigma,\kappa)\right)\Phi(x\beta_1)\right] \quad \text{(B-2)}$$

$$ME = E\left[\left(\beta_{e1}\varphi(x\beta_1) + \beta_{e2}^*\Phi(x\beta_1)\right)\exp\left(x\beta_2^* + C(\sigma,\kappa)\right)\right] \quad \text{(B-3)}$$

where $\varphi(\cdot)$ is the standard normal probability density function.

The marginal effect can be consistently estimated by (B-4) if we have consistent estimates of α in the auxiliary regression.

$$ME = \frac{1}{n}\sum_{i=1}^{N}\left\{\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1)\exp(x_i\hat{\beta}_2) + \hat{\beta}_{e2}^*\Phi(x_i\hat{\beta}_1)\exp(x_i\hat{\beta}_2)\right\} \quad \text{(B-4)}$$

where the "^s" are the estimated values, and n is the number of observations.

APPENDIX C
STANDARD ERROR OF THE MARGINAL EFFECT OF THE ENDOGENOUS POLICY VARIABLE

Asymptotic Covariance Matrix of the Two-Part Model GGE Estimator

The following notational conventions will be maintained for a scalar function s of two vector arguments r and t (i.e., s = s(r, t) where s is a scalar and r and t are vectors):

$$\nabla_r s = \frac{\partial s}{\partial r} \tag{C-1}$$

and

$$\nabla_{rt} s = \frac{\partial^2 s}{\partial r \partial t}. \tag{C-2}$$

We also assume that the former is a row vector, and the latter is a matrix with row dimension equal to that of the first subscript on ∇ and column dimension equal to that of the second subscript.

The objective function in the first stage of the 2SRI that corresponds to equation (2-29) is:

$$q_1 = -(x_e - \exp(w\alpha))^2 \tag{C-3}$$

where $w = [1 \quad x_o \quad w^+]$ and $\alpha = \left[\alpha_0 \;\; \alpha_{x_o}' \;\; \alpha_{w^+}'\right]'$, and the objective function for the second-stage (two-part) model covering equations (2-30) and (2-31) is[1]

$$q = y^* \ln \Phi(x\beta_1) + (1 - y^*) \ln\left[1 - \Phi(x\beta_1)\right] - y^* \ln f(y;\kappa,\sigma,\mu) \tag{C-4}$$

[1] Equations (C-3) and (C-4) are actually abbreviated versions of the actual estimation objective functions $(1/n)\sum_{i=1}^{n} q_1(x_{ei}, w_i, \alpha)$ and $(1/n)\sum_{i=1}^{n} q(y_i^*, y_i, x_{ei}, w_i; \beta_1, \beta_2^*, \alpha)$, respectively. The individual subscript "I" is suppressed for notational convenience, and n is the sample size.

where $x = [x_o \;\; x_e \;\; x_u]$, $\beta_1 = [\beta_{o1}' \;\; \beta_{e1} \;\; \beta_{u1}]'$, $\mu = x\beta_2^*$, $\beta_2^* = [\beta_{o2}^{*\prime} \;\; \beta_{e2}^* \;\; \beta_{u2}^*]'$, and f( ) is defined as in (2-2). By Theorem 6.11 of White (1994) we have that

$$ACOV(\tau) = A^{-1}B(A^{-1})' \tag{C-5}$$

where $\tau = [\alpha' \;\; \beta_1' \;\; \beta_{GG}']'$ and $\beta_{GG} = [\beta_2^{*\prime} \;\; \sigma \;\; \kappa]'$

$$A = \begin{pmatrix} E[\nabla_{\alpha\alpha} q_1] & 0 & 0 \\ E[\nabla_{\beta_1\alpha} q] & E[\nabla_{\beta_1\beta_1} q] & 0 \\ E[\nabla_{\beta_{GG}\alpha} q] & 0 & E[\nabla_{\beta_{GG}\beta_{GG}} q] \end{pmatrix} \tag{C-6}$$

and

$$B = \begin{pmatrix} E[\nabla_\alpha q_1 '\nabla_\alpha q_1] & E[\nabla_\alpha q_1 '\nabla_{\beta_1} q] & E[\nabla_\alpha q_1 '\nabla_{\beta_{GG}} q] \\ E[\nabla_{\beta_1} q'\nabla_\alpha q_1] & E[\nabla_{\beta_1} q'\nabla_{\beta_1} q] & E[\nabla_{\beta_1} q'\nabla_{\beta_{GG}} q] \\ E[\nabla_{\beta_{GG}} q'\nabla_\alpha q_1] & E[\nabla_{\beta_{GG}} q'\nabla_{\beta_1} q] & E[\nabla_{\beta_{GG}} q'\nabla_{\beta_{GG}} q] \end{pmatrix}. \tag{C-7}$$

The matrices A and B can be substantially simplified. First note that, because the second-stage two-part model estimator based on (C-4) is FIML, using condition (29) in Murphy and Topel (1985), we can write

$$E[\nabla_{\beta_1\alpha} q] = -E[\nabla_{\beta_1} q'\nabla_\alpha q] \tag{C-8}$$

and

$$E[\nabla_{\beta_{GG}\alpha} q] = -E[\nabla_{\beta_{GG}} q'\nabla_\alpha q]. \tag{C-9}$$

Moreover, because estimation in each of the components of the two-part model is FIML we have

$$E[\nabla_{\beta_1} q'\nabla_{\beta_1} q] = -E[\nabla_{\beta_1\beta_1} q] \tag{C-10}$$

$$E[\nabla_{\beta_{GG}} q'\nabla_{\beta_{GG}} q] = -E[\nabla_{\beta_{GG}\beta_{GG}} q]. \tag{C-11}$$

Note also that all off-diagonal elements of B are null. In particular

$$E[\nabla_{\beta_1} q' \nabla_{\alpha} q_1] = E\left[E[\nabla_{\beta_1} q' \nabla_{\alpha} q_1] \mid x, w^+\right] = E\left[E[\nabla_{\beta_1} q' \mid x, w^+] \nabla_{\alpha} q_1\right] \tag{C-12}$$

because when x_e, x_o, x_u, and w^+ are fixed, so is $\nabla_{\alpha} q_1$. Using (13.20) on p. 393 of Wooldridge (2002), however, we have that

$$E[\nabla_{\beta_1} q' \mid x, w^+] = 0 \tag{C-13}$$

because $\nabla_{\beta_1} q$ is the score of the likelihood function for the first part of the two-part model. Therefore

$$E[\nabla_{\beta_1} q' \nabla_{\alpha} q_1] = 0\,. \tag{C-14}$$

We can similarly establish that

$$E[\nabla_{\beta_{GG}} q' \nabla_{\alpha} q_1] = 0\,. \tag{C-15}$$

Also note that we can write

$$E[\nabla_{\beta_1} q' \nabla_{\beta_{GG}} q] = E\left[E[\nabla_{\beta_1} q' \nabla_{\beta_{GG}} q] \mid x, w^+, y^*\right] = E\left[\nabla_{\beta_1} q' E[\nabla_{\beta_{GG}} q \mid x, w^+, y^*]\right] \tag{C-16}$$

because when x_e, x_o, x_u, w^+ and y^* are fixed, so is $\nabla_{\beta_1} q$. Here again, because $\nabla_{\beta_{GG}} q$ is the score of the likelihood function for the second part of the two-part model we obtain

$$E[\nabla_{\beta_1} q' \nabla_{\beta_{GG}} q] = 0\,. \tag{C-17}$$

Summarizing these results, we rewrite matrices A and B as

$$A = \begin{pmatrix} E[\nabla_{\alpha\alpha} q_1] & 0 & 0 \\ -E[\nabla_{\beta_1} q' \nabla_{\alpha} q] & E[\nabla_{\beta_1\beta_1} q] & 0 \\ -E[\nabla_{\beta_{GG}} q' \nabla_{\alpha} q] & 0 & E[\nabla_{\beta_{GG}\beta_{GG}} q] \end{pmatrix} \tag{C-18}$$

and

$$B = \begin{pmatrix} E[\nabla_{\alpha} q_1' \nabla_{\alpha} q_1] & 0 & 0 \\ 0 & -E[\nabla_{\beta_1\beta_1} q] & 0 \\ 0 & 0 & -E[\nabla_{\beta_{GG}\beta_{GG}} q] \end{pmatrix}. \tag{C-19}$$

This simplification is substantial because consistent estimates of all of the components of A and B can be obtained from conventional Stata® output for the first-stage (NLS), second-stage (Probit), and third-stage (generalized gamma).

We have derived the first element of matrix A using first stage objective function.

$$\nabla_{\alpha\alpha} q_1 = 2\left[\left(x_e - \exp(w\alpha)\right)\exp(w\alpha) - \left(\exp(w\alpha)\right)^2\right] w'w . \tag{C-20}$$

The first component of the second element of matrix A can be calculated using the score equation from probit estimation in the first part of two part model.

$$\frac{\partial q}{\partial \beta_1} = \underbrace{\frac{\partial \ln L(\beta_1)}{\partial x\beta_1}}_{\text{score}} \cdot \underbrace{\frac{\partial x\beta_1}{\partial \beta_1}}_{x} = \text{Score}_{\beta_1} \times x \tag{C-21}$$

The second component, $\nabla_\alpha q$, can be derived using score equations from probit and generalized gamma estimations.

The objective function for q is given in equation (C-4) and the parameter α enters this equation through $\hat{x}_u$ ($\hat{x}_u = x_e - \exp(w\hat{\alpha})$). We can define

$$x\beta_1 = \beta_{01} + x_e\beta_{e1} + x_o\beta_{o1} + \hat{x}_u\beta_{u1} = \beta_{01} + x_e\beta_{e1} + x_o\beta_{o1} + \left(x_e - \exp(w\hat{\alpha})\right)\beta_{u1} \tag{C-22}$$

and

$$x\beta_2^* = \beta_{02}^* + x_e\beta_{e2}^* + x_o\beta_{o2}^* + \hat{x}_u\beta_{u2}^* = \beta_{02}^* + x_e\beta_{e2}^* + x_o\beta_{o2}^* + \left(x_e - \exp(w\hat{\alpha})\right)\beta_{u2}^* . \tag{C-23}$$

Using the above equalities and the objective function for q we can derive $\nabla_\alpha q$,

$$\nabla_\alpha q = \frac{\partial \ln L(\beta_1)}{\partial x\beta_1} \cdot \frac{\partial x\beta_1}{\partial \alpha} - y^* \frac{\partial \ln L(\beta_2^*)}{\partial x\beta_2^*} \cdot \frac{\partial x\beta_2^*}{\partial \alpha}$$

$$= \text{Score}_{\beta_1}\left(-\beta_{u1}\exp(w\alpha)w\right) - y^* \text{Score}_{\beta_2^*}\left(-\beta_{u2}^*\exp(w\alpha)w\right) \tag{C-24}$$

The third element of matrix A, $\nabla_{\beta_1\beta_1} q$, can be derived from negative inverse of variance matrix from probit estimation.

The fourth element of matrix A has two parts that needs to be derived separately. For the first part we can use scores from generalized gamma estimation.

$$\frac{\partial q}{\partial \beta_2^*} = -y * \frac{\partial \ln f(\cdot)}{\partial \beta_2^*} = -y * \frac{\partial \ln f(\cdot)}{\partial x\beta_2^*} \cdot \frac{\partial x\beta_2^*}{\partial \beta_2^*} = -y * \text{Score}_{\beta_2^*} x \tag{C-25}$$

$$\frac{\partial q}{\partial \sigma} = -y * \frac{\partial \ln f(\cdot)}{\partial \sigma} = -y * \frac{\partial \ln f(\cdot)}{\partial \ln \sigma} \cdot \frac{\partial \ln \sigma}{\partial \sigma} = -y * \text{Score}_{\ln\sigma} 1/\sigma \tag{C-26}$$

$$\frac{\partial q}{\partial \kappa} = -y * \frac{\partial \ln f(\cdot)}{\partial \kappa} = -y * \text{Score}_{\kappa} . \tag{C-27}$$

For the second part we will use scores from probit and GG estimations as given in above.

$$\nabla_{\alpha} q = \frac{\partial \ln L(\beta_1)}{\partial x\beta_1} \cdot \frac{\partial x\beta_1}{\partial \alpha} - y * \frac{\partial \ln L(\beta_2^*)}{\partial x\beta_2^*} \cdot \frac{\partial x\beta_2^*}{\partial \alpha}$$

$$= \text{Score}_{\beta_1} \left(-\beta_{u1} \exp(w\alpha) w\right) - y * \text{Score}_{\beta_2^*} \left(-\beta_{u2}^* \exp(w\alpha) w\right) \tag{C-28}$$

To derive the last element of matrix A, $\nabla_{\beta_{GG}\beta_{GG}} q$, we can use the negative inverse of variance matrix from generalized gamma estimation.

The first element of matrix B, $\nabla_{\alpha} q_1 ' \nabla_{\alpha} q_1$, can be derived using the following,

$$\frac{\partial q_1}{\partial \alpha} = 2\left(x_e - \exp(w\alpha)\right) \exp(w\alpha) w . \tag{C-29}$$

As we have mentioned above, $-\nabla_{\beta_1\beta_1} q$, can be derived using the inverse of variance matrix for probit estimation.

We can use the method described above, the inverse of variance matrix for generalized gamma estimation to derive $-\nabla_{\beta_{GG}\beta_{GG}} q$.

Asymptotic Standard Error of the Marginal Effect Estimator

The marginal effect (ME) can be consistently estimated using

$$ME = \frac{1}{n}\sum_{i=1}^{N}\left\{\left(\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1) + \hat{\beta}^*_{e2}\Phi(x_i\hat{\beta}_1)\right)\exp\left(x_i\hat{\beta}^*_2 + C\left(\hat{\sigma},\hat{\kappa}\right)\right)\right\} \tag{C-30}$$

and the "^'s" denote estimates. Terza (2010) shows that the asymptotic variance of ME is

$$avar(ME) = E[\nabla_\tau me]\ ACOV(\hat{\tau})\ E[\nabla_\tau me]' + E\left[(me - ME)^2\right] \tag{C-31}$$

where

$$me = \left(\beta_{e1}\varphi(x_i\beta_1) + \beta^*_{e2}\Phi(x_i\beta_1)\right)\exp\left(x_i\beta^*_2 + C(\sigma,\kappa)\right) \tag{C-32}$$

$$\nabla_\tau me = [\nabla_\alpha me \quad \nabla_{\beta_1} me \quad \nabla_{\beta_{GG}} me]. \tag{C-33}$$

Now

$$\nabla_\alpha me = \left[\beta_{e1}(-x\beta_1)\varphi(x\beta_1)\beta_{\omega1} - \exp(w\alpha) + \beta^*_{e2}\varphi(x\beta_1)\beta_{\omega1} - \exp(w\alpha)\right]\exp(x\beta^*_2 + C(\sigma,\kappa))$$

$$+\left(\beta_{e1}\varphi(x\beta_1) + \beta^*_{e2}\Phi(x\beta_1)\right)\exp(x\beta^*_2 + C(\sigma,\kappa))\beta^*_{\omega2}\left(-\exp(w\alpha)\right)\Big\}\, w$$

$$= \Big\{\beta_{e1}\beta_{\omega1}(x\beta_1)\varphi(x\beta_1) - \beta^*_{e2}\beta_{\omega1}\varphi(x\beta_1)$$

$$-\beta^*_{\omega2}\left(\beta_{e1}\varphi(x\beta_1) + \beta^*_{\omega2}\Phi(x\beta_1)\right)\Big\}\exp(w\alpha)\exp(x\beta^*_2 + C(\sigma,\kappa))\ w \tag{C-34}$$

$$\nabla_{\beta_1} me = \left(\beta_{e1}(-x\beta_1)\varphi(x\beta_1) + \beta^*_{e2}\varphi(x\beta_1)\right)\exp(x\beta^*_2 + C(\sigma,\kappa))\ x$$

$$= \left(\beta^*_{e2} - \beta_{e1}(x\beta_1)\right)\varphi(x\beta_1)\exp(x\beta^*_2 + C(\sigma,\kappa))\ x \tag{C-35}$$

$$\nabla_{\beta^*_2} me = \left(\beta_{e1}\varphi(x\beta_1) + \beta^*_{e2}\Phi(x\beta_1)\right)\exp(x\beta^*_2 + C(\sigma,\kappa))\ x \tag{C-36}$$

and $ACOV(\hat{\tau})$ denotes the asymptotic covariance matrix of $\hat{\tau}$ given in (C-5). Following Terza (2010),

$$\sqrt{\frac{n}{a\,var(ME)}}\left(ME - ME\right) \xrightarrow{d} n(0,1) \tag{C-37}$$

where

$$\text{a}\,\text{var}(\text{ME}) = \left(\frac{\sum_{i=1}^{n}\nabla_{\tau}\text{me}_i}{n}\right)\left(n\ \text{ACOV}(\hat{\tau})\right)\left(\frac{\sum_{i=1}^{n}\nabla_{\tau}\text{me}_i}{n}\right)' + \left(\frac{\sum_{i=1}^{n}\left(\text{me}_i - \text{ME}\right)^2}{n}\right) \tag{C-38}$$

where

$$\nabla_{\tau}\text{me}_i = [\nabla_{\alpha}\text{me}_i \quad \nabla_{\beta_1}\text{me}_i \quad \nabla_{\beta_{GG}}\text{me}_i] \tag{C-39}$$

$$\nabla_{\alpha}\text{me}_i = \left\{\hat{\beta}_{e1}\hat{\beta}_{\omega 1}(x_i\hat{\beta}_1)\varphi(x_i\hat{\beta}_1) - \hat{\beta}^*_{e2}\hat{\beta}_{\omega 1}\varphi(x_i\hat{\beta}_1)\right.$$

$$\left. -\hat{\beta}^*_{\omega 2}\left(\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1) + \hat{\beta}^*_{\omega 2}\Phi(x_i\hat{\beta}_1)\right)\right\}\exp(w_i\hat{\alpha})\exp(x_i\hat{\beta}^*_2 + C(\hat{\sigma},\hat{\kappa}))\ w_i \tag{C-40}$$

$$\nabla_{\beta_1}\text{me}_i = \left(\hat{\beta}^*_{e2} - \hat{\beta}_{e1}(x_i\hat{\beta}_1)\right)\varphi(x_i\hat{\beta}_1)\exp(x_i\hat{\beta}^*_2 + C(\hat{\sigma},\hat{\kappa}))\ x_i \tag{C-41}$$

$$\nabla_{\beta^*_2}\text{me}_i = \left(\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1) + \beta^*_{e2}\Phi(x_i\hat{\beta}_1)\right)\exp(x_i\hat{\beta}^*_2 + C(\hat{\sigma},\hat{\kappa}))\ x_i \tag{C-42}$$

$$\nabla_{\ln\sigma}\text{me}_i = \left(\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1) + \hat{\beta}^*_{e2}\Phi(x_i\hat{\beta}_1)\right)\exp\left(x_i\hat{\beta}^*_2 + C(\hat{\sigma},\hat{\kappa})\right)\left(\frac{\hat{\sigma}}{\hat{\kappa}}\ln\left(\hat{\kappa}^2\right) + \Psi\left(\frac{1}{\hat{\kappa}^2} + \frac{\hat{\sigma}}{\hat{\kappa}}\right)\left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\right) \tag{C-43}$$

$$\nabla_{\kappa}\text{me}_i = \left(\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1) + \hat{\beta}^*_{e2}\Phi(x_i\hat{\beta}_1)\right)\exp\left(x_i\hat{\beta}^*_2 + C(\hat{\sigma},\hat{\kappa})\right)$$

$$\left\{-\frac{\hat{\sigma}}{\hat{\kappa}^2}\ln\left(\hat{\kappa}^2\right) + \frac{2\hat{\sigma}}{\hat{\kappa}^2} + \Psi\left(\frac{1}{\hat{\kappa}^2} + \frac{\hat{\sigma}}{\hat{\kappa}}\right)\left(-\frac{2}{\hat{\kappa}^3} - \frac{\hat{\sigma}}{\hat{\kappa}^2}\right) - \Psi\left(\frac{1}{\hat{\kappa}^2}\right)\left(-\frac{2}{\hat{\kappa}^3}\right)\right\}$$

$$= \left(\hat{\beta}_{e1}\varphi(x_i\hat{\beta}_1) + \hat{\beta}^*_{e2}\Phi(x_i\hat{\beta}_1)\right)\exp\left(x_i\hat{\beta}^*_2 + C(\hat{\sigma},\hat{\kappa})\right)$$

$$\left\{-\hat{\sigma}\ln\left(\hat{\kappa}^2\right) + 2\hat{\sigma} - \Psi\left(\frac{1}{\hat{\kappa}^2} + \frac{\hat{\sigma}}{\hat{\kappa}}\right)\left(\frac{2}{\hat{\kappa}} + \hat{\sigma}\right) + \Psi\left(\frac{1}{\hat{\kappa}^2}\right)\left(\frac{2}{\hat{\kappa}}\right)\right\}\left(\frac{1}{\hat{\kappa}^2}\right) \tag{C-44}$$

where $\Psi(\cdot)$ is the digamma function and $\text{ACOV}(\hat{\tau})$ is the estimated asymptotic covariance matrix.[2]

[2] The digamma function is the logarithmic derivative of the gamma function, $\Psi(x) = \frac{d}{dx}\ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$.

APPENDIX D
THE DERIVATION OF THE NLS-IBC MODEL PARAMETER VALUES

Let

$$k(a,\tau) = \left(\left((\tau/2)a + 1\right)^2\right)^{1/\tau} \tag{D-1}$$

and τ is a scalar parameter such that $-\infty < \tau < \infty$ and $\tau \neq 0$. If the conditional mean of the random variable y conditional on the random vector x can be expressed as

$$E[y \mid x] = k(x\beta^*, \tau)\exp(C) \tag{D-2}$$

where β* is a conformable (with x) vector of parameters and C is a scalar constant, then

$$E[y \mid x] = k(x\beta, \tau) \tag{D-3}$$

where the elements of β are the same as those of β* scaled by the factor $\exp((\tau/2)C)$, and with the scaled constant term shifted by $(2/\tau)[\exp((\tau/2)C) - 1]$.

The derivation of these scaling factors is as follows:

We substitute equation (D-1) into the equation (D-2) and rewrite (D-2) as

$$E[y \mid x] = [(\tau/2)x\beta^* + 1]^{(2/\tau)}\exp(C) \tag{D-4}$$

$$E[y \mid x] = [(\tau/2)x\beta^* + 1]^{(2/\tau)}C^{*(2/\tau)} \tag{D-5}$$

where $C^* = \exp(C)^{(2/\tau)} = \exp((\tau/2)C)$.

$$E[y \mid x] = \{[(\tau/2)x\beta^* + 1]C^*\}^{(2/\tau)} \tag{D-6}$$

$$E[y \mid x] = [(\tau/2)x\beta^*C^* + C^*]^{(2/\tau)} \tag{D-7}$$

$$E[y \mid x] = [(\tau/2)x\beta^*C^* + C^* - 1 + 1]^{(2/\tau)} \tag{D-8}$$

$$E[y \mid x] = [(\tau/2)x\beta^*C^* + (\tau/2)(2/\tau)(C^* - 1) + 1]^{(2/\tau)} \tag{D-9}$$

$$E[y \mid x] = \{(\tau/2)[x\beta^*C^* + (2/\tau)(C^* - 1)] + 1\}^{(2/\tau)} \tag{D-10}$$

$$E[y \mid x] = ((\tau / 2)x\beta + 1)^{(2/\tau)} \quad \text{(D-11)}$$

where

$$\beta = \beta^* \exp((\tau / 2)C) \quad \text{(D-12)}$$

and

$$\beta_0 = \beta_0^* + (2 / \tau)[\exp((\tau / 2)C) - 1] \quad \text{(D-13)}$$

where β_0 and β_0^* are the constant terms.

Special cases of the generalized gamma – standard gamma, exponential, Weibull and log-normal

The generalized gamma has a flexible distributional form that subsumes standard gamma, exponential, Weibull and lognormal as its special cases. We can derive the scaling factors for each of these special cases by defining constant C which is a function of the parameters σ and κ

$$C = \left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right). \quad \text{(D-14)}$$

In the case of standard gamma $\kappa = \sigma$, and $C(\kappa,\kappa) = 0$ (See Appendix A for the derivation of C for each case.). After substituting $C = 0$ into equations (D-12) and (D-13) the scaling factors for standard gamma would be

$$\beta = \beta^* \exp((\tau / 2)0) = \beta^* \quad \text{(D-15)}$$

and similarly the constant shifter is

$$\beta_0 = \beta_0^* + (2 / \tau)[\exp((\tau / 2)0) - 1] = \beta_0^*. \quad \text{(D-16)}$$

Exponential is a special case of standard gamma where $\kappa = \sigma = 1$, as a result the scaling factors for exponential are the same as in the case of standard gamma.

For the Weibul case the parameter $\kappa = 1$, and $C(\sigma,1) = \ln(\Gamma(1+\sigma))$. The scaling factors (D-12) and (D-13) can be defined as

$$\beta = \beta^* \exp((\tau/2)\ln(\Gamma(1+\sigma))), \tag{D-17}$$

$$\beta_0 = \beta_0^* + (2/\tau)[\exp((\tau/2)\ln(\Gamma(1+\sigma))) - 1], \tag{D-18}$$

respectively.

Finally for the log-normal case parameter κ goes to zero in the limit and, $C(\sigma,\kappa \to 0) = \sigma^2/2$. We can rewrite (D-12) and (D-13) as

$$\beta = \beta^* \exp((\tau/2)(\sigma^2/2)), \tag{D-19}$$

$$\beta_0 = \beta_0^* + (2/\tau)[\exp((\tau/2)(\sigma^2/2)) - 1], \tag{D-20}$$

respectively.

APPENDIX E

THE DERIVATION OF THE MARGINAL EFFECT FOR THE GG-IBC MODEL

Here we derive the marginal effect for the model where parameter μ is defined as

$$\mu = \ln(k(x\beta^*, \tau)) \tag{E-1}$$

where

$$k(a,\tau) = \begin{cases} \left(\left((\tau/2)a+1\right)^2\right)^{1/\tau} & \tau \neq 0 \\ \exp(a) & \tau = 0 \end{cases}. \tag{E-2}$$

In this case the conditional mean function can be written as

$$E[y \mid x] = \exp(\ln((\tau/2)x\beta^* + 1)^{2/\tau}) + C(\sigma,\kappa)) = ((\tau/2)x\beta^* + 1)^{2/\tau})\exp(C(\sigma,\kappa)) \tag{E-3}$$

where

$$C = \left(\frac{\sigma}{\kappa}\right)\ln\left(\kappa^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\kappa^2}\right) + \left(\frac{\sigma}{\kappa}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\kappa^2}\right\}\right). \tag{E-4}$$

The marginal effect for a continuous policy variable, x_p, is defined as the derivative of the conditional mean of the outcome variable with respect to x_p, and defined as

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right]. \tag{E-5}$$

When the conditional mean is defined using an inverse Box-Cox transformation as in equation (E-3) we have three cases according to the value of parameter τ. These are summarized below:

Case 1: If parameter $\tau = 0$, we can rewrite parameter μ as $\mu = \ln(\exp(x\beta^*)) = x\beta^*$, and following this the conditional mean of the outcome variable y becomes

$$E[y \mid x] = \exp(x\beta^* + C) \tag{E-6}$$

which is the case for GG. Using (E-6) the marginal effect for the policy variable x_p can be derived as

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right] = E\left[\beta_p^* \exp\left(x\beta^* + C\right)\right] \tag{E-7}$$

where C is as given in equation (E-4).

The marginal effects as given in (E-7) can be consistently estimated using

$$ME = \hat{\beta}_p^* \left(\frac{\sum_{i=1}^{n} \exp\left(x_i \hat{\beta}^* + \hat{C}\right)}{n} \right) \tag{E-8}$$

where

$$\hat{C} = \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right) \ln\left(\hat{\kappa}^2\right) + \ln\left(\Gamma\left\{\left(\frac{1}{\hat{\kappa}^2}\right) + \left(\frac{\hat{\sigma}}{\hat{\kappa}}\right)\right\}\right) - \ln\left(\Gamma\left\{\frac{1}{\hat{\kappa}^2}\right\}\right). \tag{E-9}$$

Case 2: For $\tau = 2$, parameter μ becomes $\mu = \ln(((2/2)x\beta^* + 1)^{2/2}) = \ln(x\beta^* + 1)$ and conditional mean of y is now defined as

$$E[y \mid x] = \exp(\ln(x\beta^* + 1) + C) = (x\beta^* + 1)\exp(C) \tag{E-10}$$

where C is as given in equation (E-4). Above equation (E-10) is the case for linear conditional mean. For the linear case the marginal effect is simply defined as

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right] = E\left[\beta_p^* \exp(C)\right] \tag{E-11}$$

where C is as given in equation (E-4).

The consistent estimator for the marginal effect given in equation (E-11) is

$$ME = \hat{\beta}_p^* \exp\left(\hat{C}\right) \tag{E-12}$$

where $\hat{C}$ as is given in (E-9).

Case 3: In the last case the parameter $\tau \neq 0$ or $\tau \neq 2$, and the parameter μ can be defined as $\mu = \ln((((\tau/2)x\beta^* + 1)^2)^{1/\tau})$. Substituting this into the conditional mean equation gives

$$E[y \mid x] = \exp(\ln((\tau/2)x\beta^* + 1)^{2/\tau}) + C) = ((\tau/2)x\beta^* + 1)^{2/\tau} \exp(C) \quad \text{(E-13)}$$

where C is as given in equation (E-4). The derivative inside the expected value can be written as

$$\frac{\partial E[y \mid x]}{\partial x_p} = \frac{2}{\tau}((\tau/2)x\beta^* + 1)^{(2/\tau)-1}\frac{\tau}{2}\beta_p^* \exp(C) = ((\tau/2)x\beta^* + 1)^{(2/\tau)-1}\beta_p^* \exp(C) \quad \text{(E-14)}$$

where $C(\sigma,\kappa)$ is as given in equation (E-4). And after taking the expected value of the expression in equation (E-14) we can write the marginal effect of the policy variable as

$$ME = E\left[\frac{\partial E[y \mid x]}{\partial x_p}\right] = E\left[((\tau/2)x\beta^* + 1)^{(2/\tau)-1}\beta_p^* \exp(C)\right] \quad \text{(E-15)}$$

where C is as given in equation (E-4). The consistent estimator for the marginal effect given in equation (E-15) is

$$ME = \hat{\beta}_p^* \left(\frac{\sum_{i=1}^{n}((\hat{\tau}/2)x_i\hat{\beta}^* + 1)^{(2/\hat{\tau})-1} \exp(\hat{C})}{n} \right) \quad \text{(E-16)}$$

where $\hat{C}$ as is given in (E-9).

REFERENCES

Baser, O., Bradley, C.J., Gardiner, J.C., Given, C., 2003. Testing and correcting for non-random selection bias due to censoring: an application to medical costs. Health Services and Outcomes Research Methodology 4, 93-107.

Basu, A., Manning, W.G., 2006. A test for proportional hazards assumption within the class of exponential conditional mean framework. Health Services and Outcomes Research Methodology 6, 81-100.

Basu, A., Rathouz, P., 2005. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics 6, 93-109.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. Journal of the Royal Statistical Society Series B 26, 211-252.

Bradford, W.D., Zoller, J., Silvestri, G.A., 2010. Estimating the effect of individual time preferences on the on the use of disease screening. Southern Economic Journal 76 (4), 1005-1031.

Carpio, C.E., Wohlgenant, M.K., Boonsaeng, T., 2008. The demand for agritourism in the United States. Journal of Agricultural and Resource Economics 33 (2), 254-269.

DeSimone, J., 2002. Illegal drug use and employment. Journal of Labor Economics 20, 952-977.

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association 78, 605-610.

Etile, F., Jones, A.M., 2009. Smoking and education in France. Working Paper. http://www.paris.inra.fr/aliss/content/download/3255/31214/version/1/file/ALISSWP2009-04Etile.pdf.

Gavin, N.I., Adams, K., Manning, W.G., Raskind-Hood, C., Urato, M., 2007. The impact of welfare reform on insurance coverage before pregnancy and the timing of prenatal care initiation. Health Services Research 42, 1564-1588.

Gibson, T.B., Mark, T.L., Axelsen, K., Baser, O., Rublee, D.A., McGuigan, K.A., 2006. Impact of statin copayments on adherence and medical care utilization and expenditures. American Journal of Managed Care 12, SP11-SP19.

Gould, W., Pitblado, J., Sribney, W., 2003. Maximum Likelihood Estimation with Stata®. Stata® Press, Texas.

Heffler, S., Smith, S., Keehan, S., Borger, C., Clemens, M., Truffler, C., 2005. Trends: U.S. health spending projections for 2004-2014. Health Affairs 24 (suppl.), W5-W74.

Hill, S.C., Miller, G.E., 2009. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. Health Economics 19, 608-627.

Kenkel, D., Terza, J.V., 2001. The effect of physician advice on alcohol consumption: count regression with an endogenous treatment effect. Journal of Applied Econometrics 16, 165-184.

Lindrooth, R.C., Weisbrod, B.A., 2007. Do religious nonprofit and for-profit organizations respond differently to financial incentives? The hospice industry. Journal of Health Economics 26, 342-357.

Manning, W.G., Basu, A., Mullahy, J., 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. Journal of Health Economics 24, 245-488.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London.

Mullahy, J., 1997. Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior. The Review of Economics and Statistics 79 (4), 586-593.

Mullahy, J., 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. Journal of Health Economics 17, 247-281.

Murphy, K.M., Topel, R.H., 1985. Estimation and inference in two-step econometric models. Journal of Business and Economic Statistics 3 (4), 370-379.

Newey, W.K., McFadden, D.L., 1994. Large Sample Estimation and Hypothesis Testing. In: Engle and McFadden (Eds.), Handbook of Econometrics, Elsevier Science B.V., Amsterdam Ch. 36.

Norton, E.C., Van Houtven, C.H., 2006. Inter-vivos transfers and exchange. Southern Economic Journal 73, 157-172.

Pope, G.C., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Iezzoni, L.I., Ingber, M.J., Levy, J.M., Robst, J., 2004. Risk adjustment of medicare capitation payments using the CMS-HCC model. Health Care Financing Review 25, 119-141.

Richardson, L., Loomis, J., Champ, P.A., 2010. A comparison of methodologies for valuing decreased health effects from wildfire smoke. Working Paper. http://ageconsearch.umn.edu/bitstream/61252/2/AAEA.pdf.

Rubinstein, R.Y., 1981. Simulation and the Monte Carlo Method. Wiley, New York.

Shea, D.G., Terza, J.V., Stuart, B.C., Briesacher, B., 2007. Estimating the effects of prescription drug coverage for medicare beneficiaries. Health Services Research 43, 933-949.

Shin, J., Moon, S., 2007. Do HMO plans reduce health care expenditure in the private sector? Economic Inquiry 45, 82-99.

Stuart, B.C., Doshi, J., Terza, J.V., 2009. Assessing the impact of drug use on hospital costs. Health Services Research 44, 128-144.

Tadikamalla, P.R., 1979. Random sampling from the generalized gamma distribution. Computing 23, 199-203.

Terza, J.V., 2010. Health policy analysis via nonlinear regression methods: estimation and inference in the presence of endogeneity. Working Paper.

Terza, J.V., Basu, A., Rathouz, P.J., 2008a. Two stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of Health Economics 27, 531-543.

Terza, J.V., Bradford, W.D., Dismuke, C.E., 2008b. The use of linear instrumental variables methods in health services research and health economics: a cautionary note. Health Services Research 43, 1102-1120.

Vandegrift, D., Yavas, A., 2009. Men, women, and competition: an experimental test of behavior. Journal of Economic Behavior and Organization 72, 554-570.

Wooldridge, J.M., 1992. Some alternatives to the Box-Cox regression model. International Economic Review 33, 935-955.

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.

White, H., 1994. Estimation, Inference and Specification Analysis. Cambridge University Press, Cambridge.

Zhang, Y., 2008. Cost-saving effects of Olanzapine as long-term treatment for bipolar disorder. The Journal of Mental Health Policy and Economics 11, 135-146.

BIOGRAPHICAL SKETCH

Mujde Z. Erten completed her Bachelor of Arts degree in economics from Bogazici University, Istanbul, Turkey in 2001. She has been a Ph.D. student in the Department of Economics at the University of Florida, Gainesville, FL, USA since 2003. She earned her Master of Arts degree in economics from University of Florida, Gainesville, FL, USA in 2007. During her graduate studies she has been a Research and Teaching Assistant in the Department of Economics and Institute for Child Health Policy at the University of Florida, Gainesville, FL, USA. Her research interests include Applied Econometrics, Health Economics, Industrial Organization and Regulation. In her first year of study at the University of Florida she received Rafael Lusky Prize – Best First Year Graduate Student.