

The Sensitivity of Econometric Model Fit under Different Distributional Shapes

Manasigan Kanchanachitra*

University of North Carolina at Chapel Hill

October 18, 2009

*I would like to thank my advisor, Donna Gilleskie, as well as David Guilkey, Helen Tauchen, Sang Soo Park, Sally Stearns, and participants of the UNC Applied Microeconomics workshop for their helpful comments and suggestions.

1 Introduction

Many empirical questions in economics often involve the relationship between a set of explanatory variables and an outcome of interest. For example, a common question in health economics might be to understand how an individual's health affects his medical care expenditure. While it may be enough to quantify the effects of particular covariates on the expected or mean outcome, significant insight may be gained if the effects of covariates on the entire distribution of an outcome can be ascertained.

To illustrate the importance of knowing the distribution of an outcome in addition to its expected value, consider an economic decision making problem that involves uncertainty. The health insurance selection model, for example, states that an individual chooses a health insurance plan to maximize her expected utility. Health insurance is chosen without knowledge of one's medical care expenditures which depend on uncertain future health (and price and utilization).¹ The uncertainty of health requires that, at the time of the insurance decision, an individual forecasts the distribution of medical care expenditures. Health insurance is purchased to protect against the low probability of (a disastrous health event that requires) high medical care expenditures. Therefore, correct modeling of the tail of the medical care expenditure distribution (in addition to knowing the expected medical care expenditure) is important for accurately understanding observed health insurance decisions.

¹While future prices of medical care and future shocks to preferences are also unknown, we focus here on unobserved future health.

Another example in health economics is the vaccination decision. Whether or not an individual gets a vaccination for a particular disease depends on the probability of contracting the disease. An individual may not choose to get vaccinated if she underestimates her probability of contracting the illness, as she believes that the costs of getting vaccinated exceed the benefits, or the value of the expected lifetime utility associated with not vaccinating exceeds that of vaccinating. Again, the implications from ignoring the tails of a distribution may lead to socially suboptimal vaccination decisions.

Even when the entire distribution is not of specific interest to an economist, calculation of an outcome's expected value may depend on assumptions about the entire distribution. Policy makers may care about predicting average expenditures conditional on covariates, and estimation of the marginal effects often requires some assumptions about the underlying distribution of the outcome.

Approximation of the entire distribution of an outcome variable is often complicated by particular characteristics of the distribution shape. Many variables of interest to health economists are characterized by nonnegative outcomes, a nontrivial fraction of zero outcomes, and a positively skewed distribution with a long heavy right tail. Economists have many ways to deal with these empirical challenges. In the health economics literature, skewness is frequently addressed by three main approaches when estimating the marginal effects of expected outcome conditional on the covariates. The most widely used method is to take a logarithmic transformation of the dependent variable and apply ordinary least squares (OLS). The estimated results are then transformed back to the original scale to achieve interpretable results. The second technique is to use generalized linear models (GLM) that

assume a family distribution for the dependent variable. The final approach is a semi-parametric survival model such as the cox proportional hazard model.

While taking skewness into consideration, these three widely used approaches are limited to estimating the effect of variation in covariates on variation in the shape and location parameters of an assumed distribution.² These methods do not let the data fully dictate the shape of the distribution (or capture the tail per se), but rather allow the data to fit an imposed distribution.

In order to allow the data to influence the outcome's distributional shape, one may consider less parametric methods that impose minimal distributional assumptions. These methods include conditional density estimation and kernel regression. The use of these approaches in empirical applications is quite limited mainly due to their unavailability in commonly used statistical software packages.

This paper explores different econometric methods in their ability to take the nontrivial fraction of zeros, skewed and heavy tailed positive outcome distributions into account. The goals of this paper are threefold. First, I explore how well each method performs in estimating the first moments. In particular, I evaluate the sensitivity of these econometric models to outliers. Second, I evaluate how well each model approximates the true distribution of

²OLS estimates include μ and σ . GLM estimates are based on an assumed exponential family density such as gamma. The mean and variance that are specified in GLMs dictate the family of distribution to which the predicted values follow. Cox proportional hazard models are less parametric than OLS and GLM as it does not specify the functional form for y , but fully specifies the functional form for $x\beta$. However, the estimated β 's from a Cox proportional hazard model are not directly comparable to those of OLS or GLM since they reflect the hazard rate given the covariates.

outcomes, specifically focusing on the tails of the distribution. Third, I explore the implications of incorrectly capturing the tails. The scope of the alternative methods I consider include both parametric and nonparametric models that extend beyond the frequently used tools. These estimation procedures are evaluated under different data generating processes using Monte Carlo simulation experiments. The findings of this paper inform researchers on the optimal choice of estimation tool when faced with different types of data distributions and enable a more thorough interpretation of estimation results.

The rest of this paper proceeds as follows: Section 2 presents the background that includes a theoretical model, literature review, and econometric model descriptions. Section 3 describes the data generating mechanisms and evaluation criteria in the Monte Carlo experiment. Section 4 reports and discusses the preliminary results, and finally, Section 5 concludes and discusses future work.

2 Background

In this section, I describe an individual's optimization problem that illustrates how the specification of a distribution is an integral part of economic analysis. I then provide examples from the empirical economic literature that demonstrate attempts to capture distributions that affect decision making. Lastly, I describe the econometric tools that I evaluate in this paper in detail.

2.1 Theoretical Model

Consider a model of individual decision-making under uncertainty. Let y be the individual's income from which she directly derives utility, given by $U(y)$. Suppose there is a probability that she may lose $\$ \alpha$. I assume that the distribution of the monetary loss α is known. She does not know, however, her exact draw. She can formulate her expected utility by integrating over the entire distribution of the outcome. Therefore, her expected utility can be expressed as

$$E(U) = \int U(y - \alpha) f(\alpha) d\alpha \tag{1}$$

where $f(\alpha)$ is the probability density function of the potential loss of money.

Suppose now the agent can purchase insurance to compensate the loss in income. Let p be the price of the insurance contract and $v(\alpha)$ be the payment received from the insurance plan in the case where a $\$ \alpha$ of loss occurs. If there is no loss, then there is no compensation ($v(\alpha) = 0$). Her expected utility from purchasing the insurance is therefore

$$E(U) = \int U(y - \alpha - p + v(\alpha)) f(\alpha) d\alpha. \tag{2}$$

The agent buys the insurance if the expected utility from doing so exceeds the expected utility with no insurance. In this case, the price of insurance, p , and the payment made if a loss occurs, $v(\alpha)$, are important determinants of the expected utility with insurance. It is also important, however, how much the individual expects the loss to be, which is directly derived from the underlying distribution of α .

In the example given above, the insurance decision does not change the distribution of the monetary loss associated with illness. The individual's decision to purchase health insurance does not affect the probability that she gets sick. In other cases, however, the decision can directly affect the income loss distribution (even in the absence of moral hazard). Suppose again that the individual faces the expected utility in (1) if no decision is made. Suppose now that the purchase of insurance eliminates (or greatly reduces) the probability of a monetary loss. The expected utility of purchasing this insurance is now without the uncertainty of monetary loss,

$$U = U(y - p). \tag{3}$$

Again, the individual decides on insurance if (3) is greater than (1) and against it if (3) is less than (1). An example of this case is the vaccination decision. There is a probability of the individual contracting a particular disease without the vaccination, which is associated with potential monetary loss, α . With vaccination, however, the distribution of the probability of a monetary loss changes as the probability of contracting the disease is eliminated (from from $f(\alpha)$ in (1) to zero probability in (3)). The decision depends on what the individual expects her chances of getting sick with the disease to be, which requires a reliable approximation of the loss distribution.

In both cases, if the individual underestimates the probability of incurring a high loss (underestimates the upper tail of the distribution), she will under consume insurance. Correctly identifying the underlying distribution leads to more reliable research results, which are essential for policy implications.

2.2 Literature Review

In this section, I consider models that require the econometrician to approximate the entire distribution of a variable. I look at how studies attempt to capture the distribution of a variable of interest, particularly when these distributions are positively skewed. Studies that directly evaluate the performance of econometric methods are incorporated in the Model Description section.

I begin with studies that look at health insurance coverage and employment decisions. Rust and Phelan (1997) study whether Social Security and Medicare affect an individual's decision to retire, when the loans, annuities, and health insurance markets are incomplete. Individuals realize that there is a positive probability of incurring high medical care expenditures, and there may be some security value to remain employed until they are eligible for Medicare. In order to capture the distribution of medical care expenditures, the authors treat the expenditures as a mixture of a mass point of \$500 or less (to represent the high probability of incurring some health expenses), and a continuous long tail distribution of the medical care expenditure over \$500. They find that the continuous part is well described by the Pareto distribution, which has one parameter that characterizes the size of the tail. Their estimates yield a small parameter value for the Pareto distribution, implying a rather

large and long tail with a rather high probability of having a catastrophic medical care expenditure. Using a dynamic programming model, the paper finds that employed individuals entitled to Social Security benefits are significantly less likely to continue working as compared to those not entitled to the benefits while taking into consideration the uncertainty of medical care expenses.

Another study on health insurance and employment decisions examines the employment behavior of married couples who face uncertainty about future medical expenses (Gilleskie and Blau (2006)). Again, having health insurance helps couples smooth out their future utility of consumption across all possible scenarios. The medical care expenditure for each spouse is assumed to be a random draw from a known distribution. The authors model this underlying continuous distribution by using a discrete approximation. Specifically, they discretize the medical care expenditure into three categories (in 1992 dollars): \$0-1999, \$2000-14,999, and \$15,000 and above, and use multinomial logit models to estimate the probability of being in each category separately by sex and health status, with an intercept and a linear age term. From the linear regressions, the authors are able to assign the predicted mean expenditure by age for each sex/health category. The findings suggest a relatively modest impact of health insurance on employment behavior, given the uncertainty of medical care expenses.

Infant mortality is also a source of uncertainty that may affect a woman's fertility decisions. Mira (2007) studies the links between infant mortality and fertility decisions when there exists unobserved heterogeneity in infant mortality risks across mothers. In this dynamic stochastic model of life-cycle marital fertility behavior, the author focuses on mothers' learning about their own infant mortality probability after experiencing child deaths. The

paper assumes infant deaths conditional on birth are independent Bernoulli trials with a time-varying, mother-specific probability of death. From this framework, the findings suggest that women who experienced higher infant mortality have a higher probability of having additional births as an attempt to replace the children lost.

These are just a few examples where assumptions about a distribution are necessary for solution to the optimization problem. These assumptions, I conjecture, impact optimal decisions. In order to evaluate the importance of distributional assumptions, I consider alternative methods of capturing the distribution of an outcome, as well as econometric models that applied economists use to estimate the expected value of the outcome. I describe these models below.

2.3 Model Description

Throughout this section, consider an example where the outcome of interest (y) is medical care expenditures. Possible right hand side variables include sex, age, health status, marital status, education level, income, etc. I first describe models that deal with the non-trivial proportion of observations with zero medical care expenditures. In the econometrics literature, the zero observations are typically handled using one of these methods: the two-part model (2PM), the sample selection model (SSM). Then I discuss the ongoing debate on choice between 2PM and SSM. Lastly, I describe models that are aimed to deal with positively skewed distributions of medical care expenditures.

2.3.1 Two-Part Model

A two-part (or multi-part) model separates the observed positive outcome into two observed parts: $y > 0$ and $y|y > 0$. There are two separate equations to directly model for these two parts. The first part estimates the probability of observing positive medical care expenditures, $y > 0$, on the entire sample. The first equation is typically estimated using a standard probit model

$$I = x_1\beta_1 + e_1, \quad \text{where } e_1 \sim N(0, 1) \quad (4)$$

where $y > 0$ if $I > 0$ and $y = 0$ otherwise.

The second part of the model is estimated on those with any medical care expenditures, $y|y > 0$. Typically, the positive outcomes are estimated using OLS on log transformed variable or GLM methods. These techniques are discussed separately in subsequent sections.

2.3.2 Selection Models

Following Leung and Yu (1996), I focus on van de Ven and van Praag's (1981) version of the adjusted Tobit model. Again, there is a binary variable indicating positive medical care expenditure commonly modeled using a standard probit. However, the adjusted Tobit model takes into account the correlation between the probability of any medical care expenditure and the level of expenditure. Specifically,

$$I = x_1\beta_1 + e_3 \quad (5)$$

$$m = x_2\beta_2 + e_4 \quad (6)$$

where $\ln(y) = m$ if $I > 0$ and $-\infty$ otherwise, and

$$(e_3, e_4) \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix} \right).$$

The estimation methods that are most widely used are Heckman's two-step estimator and maximum likelihood. The first method augments the OLS regression with an omitted regressor $\hat{\lambda} = \lambda(x_1 \hat{\beta} a_1)$. Therefore, using all observations with $I > 0$, the OLS equation becomes $\ln(y) = x_2 \beta_2 + \rho \sigma \hat{\lambda} + \epsilon$, where $\hat{\lambda} = \phi(x_1 \hat{\beta} a_1) / \Phi(x_1 \hat{\beta} a_1)$ is the estimated inverse Mills' ratio, $\phi(\cdot)$ and $\Phi(\cdot)$ are the p.d.f. and c.d.f. of the standard normal distribution. Note that in the 2PM, the decision to have any medical care expenditure (e.g. whether to seek a physician) is independent from the level of expenditure incurred (e.g. how often to seek a physician). Therefore, the estimated level of expenditure is conditional on having any expenditures. On the contrary, the estimate from the level equation is an unconditional one. This fundamental difference between these two models lead to different interpretations of the estimated coefficients β s.

There has been an ongoing debate on the choice between a two-part model (2PM) and sample selection models (SSM). A fundamental question involves how one should treat the nontrivial fraction of zero outcomes. Specifically, are the zeros a result of a random process or a sample selection process? If the zeros appear random, then the use of a 2PM may be appropriate. On the other hand, if the observed zero outcomes are likely to be based on individual choice, then the use of SSM may be more appropriate in correcting for the selection bias.

The debate started when Hay and Olsen (1984) criticized Duan *et al.* (1983)'s paper that compared alternative models for the demand for medical care. In the paper, Duan

et al. (1983) evaluate the performance of a simple one part model, a two-part model, and a four-part model. Their results suggest that the multi-part models significantly outperform the simple one-part model in terms of mean squared forecast errors with the four-part model being the most preferred model. Hay and Olsen (1984) claim that the 2PM requires unusual assumptions on the model joint distribution and functional form. Moreover, the multipart model is nested in the more general selection model.

Duan *et al.* (1984) respond to the criticism claiming Hay and Olsen (1984)'s proof relies on an unmentioned restrictive assumption that cannot be satisfied. The authors demonstrated their point by offering a counterexample to prove that the 2PM is not nested within the SSM. Duan *et al.* (1984) further argue that the SSMs are *intrinsically flawed* since they rely on untestable assumptions.

Manning *et al.* (1987) attempt to settle the discussion by using Monte Carlo simulations to compare these two approaches. Using SSM as their theoretical benchmark, they find that 2PM outperforms SSM on statistical grounds. However, the Monte Carlo design of Manning *et al.* (1987) may have created collinearity problems that bias the results against SSM (Leung and Yu (1996)). The authors argue that if there is no collinearity, SSM would perform much better than 2PM. Leung and Yu (1996) conduct a different set of Monte Carlo experiments to compare the performance of sample selection and two-part models. They hypothesize that the reason that the 2PM performs well even when SSM is the true model is due to a subtle design problem in the simulation experiments. The authors believe that the range of a distribution which the regressors are drawn from in Manning *et al.* (1987) are far too narrow. SSM includes the inverse Mill's ratio, which is a function of the regressors. Therefore, when the range of the regressors is not wide enough, the SSM will be burdened with collinearity

problems. The findings from Leung and Yu (1996) suggest that no model outperforms the other under all conditions.

I now describe models that deal with outcomes that are positively skewed.

2.3.3 Ordinary least squares on log-transformed dependent variable

The ordinary least squares (OLS) on a log-transformed dependent variable is by far the most prevalent modeling approach used in labor and health economics. The method is simply to log transform the positively skewed medical care expenditures to $\ln(y)$ before applying ordinary least squares. If $\ln(y)$ has a normal distribution with mean $\mu = x\beta$ and variance σ^2 , the regression model is simply:

$$\ln(y) = x\beta + \epsilon \tag{7}$$

where x is a matrix of observed covariates, β is a column vector of coefficients to be estimated, and ϵ is the error term (Norton (2007)). Note that the error term need not be i.i.d.

The expected value of the logged expenditure is

$$\begin{aligned} E(\ln(y)) &= E(x\beta + \epsilon) \\ &= x\hat{\beta}. \end{aligned} \tag{8}$$

The log medical care expenditures need to be retransformed back to its original scale to get interpretable results. If homoskedasticity and normal error terms are assumed, the predicted retransformed medical care expenditure given the explanatory variables is

$$E(y) = \exp(x\hat{\beta} + .5\hat{\sigma}^2). \tag{9}$$

However, the assumption of normality is quite strong and may lead to biased estimates of y . Duan (1983) developed a way to estimate the retransformed dependent variable without

imposing the normality assumption by using a smearing factor, which is the average of the exponentiated estimated error terms. The predicted medical care expenditure without the normality assumption is

$$E(y) = \exp(x\hat{\beta})\left(\frac{1}{N} \sum_{i=1}^n \exp(\hat{\epsilon}_i)\right). \quad (10)$$

To obtain the marginal effect of x on the raw-scale medical care expenditure y , the calculation depends on the characteristics of the error terms. The estimated marginal effect of x is

$$\frac{\partial E(y)}{\partial x} = \hat{\beta}E(y) \quad (11)$$

where $E(y)$ is from equation (9) or (10), depending on whether or not the error term is assumed to be normal.

OLS with a log dependent variable will be resilient to many data problems (Manning and Mullahy (2001)). However, if the log-scale error term is heteroskedastic, then the estimates can be appreciably biased. To solve for this potential bias, a heteroskedastic retransformation can be employed (Duan (1983)). However, this heteroskedastic retransformation is currently beyond the scope of this paper.

2.3.4 Generalized linear model

In generalized linear models (GLM), one directly specifies the mean and variance functions for the observed medical care expenditure conditional on the covariates. The mean function for medical care expenditure is represented as

$$E(y|x) = \mu(x'\beta) \quad (12)$$

where μ is the inverse link between the expectation of the medical care expenditure and the linear predictor $x'\beta$. The log-link relationship is often chosen in health economics for the mean function such that

$$\begin{aligned} \ln(E(y)) &= x'\beta \\ E(y|x) &= \exp(x'\beta). \end{aligned} \tag{13}$$

One also specifies the variance function in GLM. The general form of the variance function is specified as

$$\nu(x) = \kappa(\mu(x'\beta))^\lambda \tag{14}$$

where λ must be nonnegative. The variance is constant when $\lambda = 0$, proportional to the mean (Poisson-like) if $\lambda = 1$, and proportional to the mean squared (gamma-like) if $\lambda = 2$.

GLM's main advantage is that one can directly specify how the expectation of medical care expenditure in its original scale is related to the covariates. If the link function is correctly specified, then the choice of the variance function will only affect the efficiency. However, if the link function is misspecified, which may likely be the case, then the model may not fit the data well across the entire range of the distribution. In this case, the specification of the variance function will determine the goodness of fit in the different parts of the distribution.

Manning and Mullahy (2001) compare log normal models and gamma models under different data generating specifications based on a Monte Carlo simulation. OLS with a homoskedastic retransformation was more resilient than GLM alternatives to heavy-tail data and large log-scale error term variance. However, the estimates from OLS with a homoskedastic retransformation can be substantially biased if the log-scale error term is heteroskedastic.

GLM methods also perform better than OLS in terms of precision when the distribution of the dependent variable is not bell-shaped or skewed bell-shaped.

2.3.5 Quantile regression

The quantile regression (QR) allows estimated coefficients to vary by quantiles, rather than being constant. Quantiles require to first order the observed medical care expenditure and divide it into segments. An observation is at the τ^{th} quantile when the proportion τ of the sample is below that observation and $(1-\tau)$ is above.

Quantile regressions can be considered an extension of a linear regression model. In the linear regression model, the coefficient estimates describe how the conditional mean of the dependent variable varies, given a change in the independent variable. The model does not take into account the full conditional distributional properties of the outcome variable. In a quantile regression however, the coefficient estimates capture the change in the conditional quantile. Therefore, one can analyze the full distributional properties of the dependent variable.

Moreover, OLS aims to describe the behavior of the location of the conditional distribution, using the mean to represent the central tendency. However, the mean may not be the best measure of central tendency for medical care expenditure as the distribution is skewed. The median, in this case, might be better at capturing the location shift of a distribution. Conditional median regression is the simplest form of quantile regression ($\tau = .5$). Other quantiles help in capturing the shape shift of the distribution. Therefore, a quantile regression models both location shifts and shape shifts as it estimates the potential different effect that a covariate has on different quantiles in the distribution.

With the homoskedasticity assumption in linear regression models, the variance is assumed to be constant for all values of x 's. The constant variance implies that OLS does not address the possibility of scale change, which is an important form of distributional shape change, with different values of the covariates.

The quantile regression minimizes the sum of absolute weighted residuals. In a median quantile regression, the minimized sum of weighted absolute residuals is symmetric in the sense that there are the same number of positive and negative residuals. For quantiles other than the median, the sum of asymmetrically weighted absolute residuals is minimized (Koenker and Hallock (2001)).

$$\min \sum \rho_{\tau}(y_i - \xi_i, \beta) \quad (15)$$

The function $\rho_{\tau}(\cdot)$ is the tilted absolute value function. The regression is solved by linear programming methods. Quantile regression is especially useful when one expects heterogeneity across different groups of individuals. OLS, for instance, would be appropriate only for the case where the interest lies in the average response of the population, or if responsiveness is similar across different groups of individuals. A quantile regression allows different groups of individuals to have different responses. For instance, Manning *et al.* (1995) find that the responsiveness to alcohol prices differ depending on whether the individual is a light, moderate, or heavy drinker.

2.3.6 The four parameter generalized beta of the second kind (GB2)

The GB2 is a family of distributions that is used to describe outcomes that are positive. The model is first introduced by McDonald (1984). It has one scale parameter, two shape

parameters, and one location parameter. More specifically,

$$f(y; \mu, \sigma, \alpha_1, \alpha_2) = \frac{[\exp(z)]^{\alpha_1}}{y|\sigma|B(\alpha_1, \alpha_2)[1 + \exp(z)]^{\alpha_1 + \alpha_2}}. \quad (16)$$

where $z = (\ln y - \mu)/\sigma$ and $B(\alpha_1, \alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1 + \alpha_2)$. The parameters are estimated via maximum likelihood and can be specified to vary with the covariates.³ With four parameters, the GB2 has the flexibility to fit data with thick or thin tails. It can also model positively skewed densities if $\alpha_1 > \alpha_2$, and negative if $\alpha_1 < \alpha_2$. The covariates are usually included in the location parameter, μ , as $x'\beta$. The covariates can also be included in the scale parameter to account for heterogeneity.

The next two models, the generalized gamma and the Singh-Maddala, are special cases of the GB2. The generalized gamma is the case when $\alpha_2 \rightarrow \infty$, and the Singh-Maddala is when $\alpha_1 = 1$. In general, the more parameters a model has, the better it will perform within a certain criterion (e.g. sum of squared error). However, more parameters also imply higher computational costs.

2.3.7 The three parameter generalized gamma

The three parameter generalized gamma (GG) has one scale parameter and two shape parameters. Following Manning et al. (2005), the probability density function for the generalized gamma is

$$f(y; \kappa, \mu, \sigma) = \frac{\gamma^\gamma}{\sigma y \sqrt{\gamma} \Gamma(\gamma)} \exp[z\sqrt{\gamma} - u] \quad y \geq 0 \quad (17)$$

where $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)[\ln(y) - \mu]/\sigma$, and $u = \gamma \exp(|\kappa|z)$. The parameter μ is set to equal $x'\beta$. Estimation is done using maximum likelihood methods. The expected value of y

³The GB2 distribution from Stata is $f(y; a, b, p, q) = \frac{ax^{ap-1}}{[(b^{a*p}) * B(p, q) * [1 + (x/b)^a]^{p+q}]}$

conditional on x for the generalized gamma is

$$E(y|x) = \exp[x\hat{\beta} + (\hat{\sigma}/\hat{\kappa}\ln(\hat{\kappa}^2) + \ln(\Gamma[(1/\hat{\kappa}^2) + (\hat{\sigma}/\hat{\kappa})]) - \ln(\Gamma[1/\hat{\kappa}^2]))] \quad (18)$$

where $\hat{\sigma} = (1/n) \sum \exp[\alpha_0 + \alpha_1 \ln(f(x_i))]$ if $\ln(\sigma)$ is parameterized as $\alpha_0 + \alpha_1 \ln(f(x))$.

Manning *et al.* (2005) extend the Manning and Mullahy (2001) paper by including the generalized gamma. The Monte Carlo experiment in the paper suggests that the generalized gamma yields estimates that are close to the true value on the log scale, except for when the error term is heteroskedastic. The precision loss associated with the generalized gamma is also smaller compared to GLM of gamma with a log link, which is a special case of the generalized gamma, when estimating a distribution with heavy tails. However, generalized gamma tends to under estimate the overall mean on the raw scale.

2.3.8 Singh Maddala

The Singh Maddala (SM) distribution is also designed to handle heavy tail data distribution. The density function for the Singh Maddala distribution is

$$f(y; \tau, \theta, \lambda) = \frac{\lambda \theta^\lambda \tau y^{\tau-1}}{(\theta + y^\tau)^{\lambda+1}}. \quad (19)$$

The SM is a special case of GB2 when $\sigma = 1/\tau$, $\mu = \ln\theta/\tau$, $\alpha_1 = 1$ and $\alpha_2 = \lambda$. One can allow either the shape parameter and/or location parameter to vary with covariates.⁴

McDonald (1984) finds that the Singh-Maddala distribution provides a better fit for the income distribution than the gamma, while the gamma outperforms the lognormal. The paper considers two generalized beta distributions. In particular, the generalized gamma and generalized beta of the first and second kinds and their special cases were fit to U.S.

⁴The SM density function from Stata is $f(x; a, b, q) = (aq/b)z^{-(q+1)}[(x/b)^{(a-1)}]$

family nominal income for 1970-1980. The author finds that the generalized beta of the second kind provides a better fit than the generalized gamma in terms of sum of squares or absolute errors, chi-square and log-likelihood criteria. The second best distribution function is Singh-Maddala, which performs better than the generalized beta of the first kind with four parameters.

2.3.9 Conditional density estimation

The conditional density estimation (CDE) approach is based on Gilleskie and Mroz (2004). Suppose we have some unknown distribution function for medical care expenditures y conditional on a set of covariates x . Let the density of this distribution be $f(y|X)$. We can break the range of the observed medical care expenditures into K intervals, where the k^{th} interval is defined by $[y_{k-1}, y_k)$, for $y_{k-1} \leq y_k$, $y_0 = -\infty$ and $y_K = \infty$.

The conditional probability that an observed value of medical care expenditure falls in the k^{th} interval given that it did not fall in one of the first $(k-1)$ intervals is

$$\lambda(k, x) = p[y_{k-1} \leq y < y_k] = \frac{\int_{y_{k-1}}^{y_k} f(y|x)dy}{1 - \int_{y_0}^{y_{k-1}} f(y|x)dy}. \quad (20)$$

Therefore, the probability that the random variable y falls in the k^{th} interval is given by

$$p[y_{k-1} \leq y < y_k|x] = \lambda(k, X) \prod_{j=1}^{k-1} [1 - \lambda(j, x)]. \quad (21)$$

The true expectation of any function $h(\cdot)$ of the random variable y given x , is

$$E(h(y)|x) = \int_{-\infty}^{\infty} h(y)f(y|x)dy \quad (22)$$

where $h(\cdot)$ is any smooth and continuous function of y .

The discrete approximation of the expected function using a partition of the support of y with K intervals is given by,

$$\tilde{E}(h(y)|x) = \sum_{k=1}^K h^*(k|K) \lambda(k, x) \prod_{j=1}^{k-1} [1 - \lambda(j, x)], \quad (23)$$

where each $h^*(k|K)$ is an approximation to $h(y)$ in the k^{th} interval. Gilleskie and Mroz are interested in the mean of y and therefore let $h^*(k|K)$ be the mean or the arithmetic average function, which does not vary with the x 's. This implies that, conditional on being in that interval, the x 's do not explain the mean value within their interval. That is, conditional on being in a particular interval, the value of y is random. The derivative of the conditional expected value in this case is

$$\frac{\partial \tilde{E}(h(y)|x)}{\partial x} = \sum_{k=1}^K h^*(k|K) \frac{\partial [\lambda(k, x) \prod_{j=1}^{k-1} (1 - \lambda(j, x))]}{\partial x} \quad (24)$$

One could also allow the mean to depend on x 's. If $h^*(k|K)$ varied with x , the derivative of the conditional expected value would have an additional term that reflects the fact that $h^*(\cdot)$ is also a function of x . Their paper suggests however, that the proposed conditional density estimator performs quite well under various data generating processes. More importantly, this estimation method allows effects of covariates to be different at different points of support in the distribution of the outcome.

2.3.10 Kernel regression

Kernel regression is a nonparametric technique that is less dependent on functional form assumptions than parametric models (Yatchew (1998)). Consider a model $y = f(x) + \epsilon$, where ϵ is i.i.d. with mean 0 and variance σ_ϵ^2 given x , which x is assumed to be a scalar

for the time being. The function f is unknown. A general formulation of local averaging estimator can be defined as

$$\hat{f}(x_0) = \sum w_t(x_0)y_t. \quad (25)$$

Several local averaging estimators can be used including kernel and nearest neighbor. Higher weights are assigned to observations closer to x_0 .

I focus on the kernel estimators. The weight function is specified as

$$w_t(x_0) = \frac{\frac{1}{\lambda T} K\left(\frac{x_t - x_0}{\lambda}\right)}{\frac{1}{\lambda T} \sum K\left(\frac{x_t - x_0}{\lambda}\right)} \quad (26)$$

where K is assumed to be a bounded function that sums to one and is symmetric around zero. It determines the shape of the weights and the magnitude is determined by the bandwidth, λ . The larger the bandwidth, λ , the more weight is being put on observations that are far from x_0 . The kernel regression function estimator is therefore

$$\hat{f}(x_0) = \frac{\frac{1}{\lambda T} \sum K\left(\frac{x_t - x_0}{\lambda}\right) y_t}{\frac{1}{\lambda T} \sum K\left(\frac{x_t - x_0}{\lambda}\right)} \quad (27)$$

There are several types of kernels. The simplest form is the uniform kernel which gives the weight of 1 on $[-1/2, 1/2]$ and 0 otherwise. The kernel that I use in this paper is the Epanechnikov function which takes the value of $\frac{3}{2}(1 - 4u^2)$ where $u \in [-1/2, 1/2]$. The choice of kernel is not as important as the choice of bandwidth size. The mean squared error can be minimized by increasing the bandwidth of the neighborhood until the increase in bias squared is offset but the decrease in variance.

Nonparametric regressions are not as widely used as one might expect. The main reasons are that the nonparametric techniques are more complex than procedures available in

standard statistical software packages, they are computationally intensive, and they require large datasets.⁵

3 Monte Carlo Experiment

To evaluate the performance of alternative econometrics model, I implement a Monte Carlo simulation experiment. I generate datasets such that there exists a nontrivial proportion of the sample with zero observed outcomes and for those with positive outcomes, the data are right skewed. Since the heart of the debate on the choice between the 2PM and SSM lies on whether the observed zeros can be treated as random or not, I consider two separate cases. In the first case, the data generating process is specified such that the two-part model is the true model. Specifically, the error terms are distributed $e_1 \sim N(0, 1)$ and $e_2|I > 0 \sim N(0, 1)$, and the two error terms are not correlated. In the second case, the sample selection model is the true model where the error terms are drawn from a bivariate normal distribution with variance 1 for each error term with a correlation of 0.5.

In both cases, $x \sim U[0, 10]$ ⁶ and the probability of having a positive outcome ($I > 0$) is set to be approximately 0.75 by setting the intercept $\beta_1 = -2.5$. This is consistent with actual

⁵The more regressors there are in the model, the larger number of observations is needed as Kernel regression relies on local weighted averaging. Observations are more sparsely distributed with higher dimensions. The convergence rate decreases with the number of regressors. This is also known as the curse of dimensionality.

⁶In my previous work, I had $x \sim U[0, 1]$. Leung and Yu (1996) find that when the range of the regressors is too narrow, the inverse Mills' ratio would be highly correlated with the regressor, leading to poor performance of the SSM even when the SSM is the true model.

medical care expenditure where 20-30% of the population do not incur any expenditure (for instance, 23% of individuals between 25-64 years of age in the MEPS 2005 data have zero health expenditures that year). The summary statistics for each case in provided in tables 1 and 2.

Table 1: Summary Statistics with 2PM as True Model

	2PM	SSM
Mean	8.84	6.58
Std. Dev.	3.26	4.77
Variance	10.6	22.77
Skewness	2.94	0.6
Kurtosis	36.89	8.71
<i>Number obs</i>	<i>74,480</i>	<i>10,000</i>

Table 2: Summary Statistics with SSM as True Model

	<i>y</i> > 0	all y
Mean	8.95	6.66
Std. Dev.	3.12	4.74
Variance	9.73	22.46
Skewness	2.41	0.35
Kurtosis	24.12	5.84
<i>Number obs</i>	<i>74,480</i>	<i>10,000</i>

To begin the Monte Carlo simulation experiment, I generate 10,000 observations for each case.⁷ I then implement alternative estimating models to attain the predicted y given each value of x . The process is simulated 200 times to obtain the average of the predicted y for each value of x across all replications.

The performance of each estimator is evaluated using the following criteria:

- The mean squared error (MSE) of the model, which is given by $MSE = \frac{1}{N} \sum (y_i - \hat{y}_i)^2$, to see how well the predicted $y|x$ fit true $y|x$.
- The mean squared error of the model by decile to examine the fit of the predicted $y|x$ in specific parts of the distribution.
- The quantile-quantile plot of the observed y and $E(y|x)$. The qq-plot is used to help evaluate how well the predicted values fit the actual data at each quantile. The qq-plot is useful particularly in examining the model fit in the extreme right tail. The predicted and actual values are first sorted from the lowest to the highest, then are plotted against one another. The qq-plot allows us to gauge each specification's performance across different values of y .
- Kolmogorov-Smirnov test (KS test) to evaluate whether the distribution of the predicted outcome fit the distribution of the observed data.

I am in search of other criteria that specifically evaluate how well estimates fit the entire observed distribution.

⁷Alternatively, I also consider a smaller number of observations of 500. Currently, I do not find that any estimating model performs significantly worse in terms of predicting the mean outcome when faced with a smaller number of observations. However, further investigation is needed.

4 Results

The goal of this paper is to evaluate how alternative models perform in estimating the effects of covariates on the entire distribution of an outcome. Examining the performance of each model in predicting y given x is the first step in achieving this goal. In table 3, I report the predicted y given x for positive outcomes when the true model is (1) 2PM and (2) SSM. The true mean is reported in the first row. The first and third columns report the predicted probabilities from a standard probit model for having any expenditures for 2PM, SSM, GLM, and Generalized Gamma. The zero observations are specified as an interval in CDE, and therefore are already accounted for in the model. 2PM, SSM and GLM appear to be performing well in estimating $E(y|x, y > 0)$ under both data generating processes.

The results in table 4 report the predicted y given x including zero observations, where $E(y|x) = \text{prob}(y > 0)E(y|x, y > 0)$. Again, 2PM, SSM and GLM seem to perform well under both cases.

Tables 5 and 6 report the MSE by decile under 2PM and SSM respectively. The MSE is broken into deciles to evaluate model prediction in each part of the data. The results suggest that the models considered do not fit the data well in the tail of the distribution (in the 10th decile).

Table 3: Predicted y given x for $y > 0$

	Probit	Mean	S.E.	Probit	Mean	S.E.
	(1)	(1)	(1)	(2)	(2)	(2)
<i>Truth</i>	0.74	8.84	(3.26)	0.74	8.95	(3.12)
Two-Part Model	0.74	8.84	(2.47)	0.74	8.97	(3.06)
Sample Selection	0.74	8.85	(2.35)	0.74	8.94	(2.21)
GLM (Gamma with a log link)	0.74	8.85	(2.36)	0.74	8.95	(2.08)
Generalized gamma	0.74	8.53	(2.71)	0.74	8.56	(2.59)
CDE	-	7.87	(3.10)	-	7.97	(2.94)
<i>Observations</i>	7,448			7,448		

Table 4: Predicted y given x for all y

	Mean	S.E.	MSE	Mean	S.E.	MSE
	(1)	(1)	(1)	(2)	(2)	(2)
<i>Truth</i>	6.58	4.77	-	6.66	4.74	-
Two-Part Model	6.59	(4.22)	5.73	6.68	(4.58)	6.72
Sample Selection	6.59	(4.19)	5.64	6.66	(4.13)	6.04
GLM (Gamma with a log link)	6.59	(4.17)	5.69	6.66	(4.06)	6.03
Generalized gamma	8.18	(3.25)	7.00	8.19	(3.15)	7.40
CDE	6.18	(3.97)	22.91	6.32	(3.83)	6.46
<i>Observations</i>	10,000			10,000		

Table 5: MSE by decile under 2PM

	2PM	SSM	GLM	GG	CDE
0-10	-	-	-	-	1.55
11-20	2.70	3.10	2.61	3.26	3.75
21-30	1.64	2.06	1.65	1.78	1.43
31-40	1.61	1.95	1.60	1.79	0.52
41-50	1.32	1.48	1.29	1.47	0.08
51-60	2.47	2.40	2.35	2.82	0.09
61-70	2.44	2.14	2.27	2.80	0.06
71-80	2.73	2.08	2.45	3.22	0.06
81-90	3.67	2.75	3.27	4.23	0.30
91-100	39.72	39.60	39.85	41.47	37.45

Table 6: MSE by decile under SSM

	2PM	SSM	GLM	GG	CDE
0-10	-	-	-	-	2.32
11-20	3.68	2.93	2.80	3.50	2.74
21-30	2.53	2.42	2.42	2.41	1.04
31-40	2.48	2.28	2.24	2.31	0.34
41-50	2.55	2.03	1.95	2.30	0.15
51-60	2.96	2.07	1.94	2.49	0.07
61-70	3.73	2.14	1.97	2.62	0.06
71-80	6.41	3.29	3.01	4.17	0.13
81-90	7.91	3.45	3.15	4.69	0.57
91-100	39.87	40.06	40.30	42.16	35.99

5 Future work

In order to evaluate the models in greater detail, I intend to experiment with different data generating processes such as:

- Different degrees of skewness, kurtosis
- Heteroskedasticity in the error terms
- Different sample sizes

In order to evaluate the goodness-of-fit of the entire distribution, I also intend to implement the Kolmogorov-Smirnov test. As mentioned, I also intend to find other goodness-of-fit tests.

Future work also includes implementing these econometric methods using a real dataset, the 2005 Medical Expenditure Panel Survey (MEPS). This dataset is a national survey on the financing and utilization of medical care in the United States. The data are from the Household Component, which provides information on a sample of family and individuals drawn from a nationally representative subsample of households that participated in the prior year's National Health Interview Survey (NHIS). The dataset includes data on demographic characteristics, health status, medical service usage, charges and sources of payments, health insurance coverage, income, and employment, amongst others.

The dependent variable y is annual individual medical care expenditures or the dollar value of medical care consumed (i.e., not just out-of-pocket expenses). From the 26,609 individuals in the full sample (after dropping those with no medical expenditures), the medical care expenditures are highly right-skewed with a mean of \$7,049.18, which is much

higher than the median of \$929. The skewness is also reflected in the kurtosis of 570.807, which is much higher than the normal distribution value of 3. Once we take the log of medical care expenditures, the distribution becomes closer to normal. The mean and median are much closer together at 7.02 and 6.83 respectively. The kurtosis after the log transformation is also closer to 3 at 2.843. The summary statistics for medical care expenditures from the 2005 MEPS are shown in table 7.

Table 7: Descriptive Statistics for medical care expenditures

	Dollars	Log Dollars
Mean	7,049.18	7.02
Median	929	6.83
Min	3	1.099
Max	1,546,859	14.252
Std. Dev.	29,553.76	1.82
Variance	8.73e+08	3.312
Skewness	18.00255	0.393
Kurtosis	570.807	2.843
<i>Number of observations</i>	<i>26,609</i>	<i>26,609</i>

References

- DUAN, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, **78** (383), 605–610.
- , MANNING, J., WILLARD G., MORRIS, C. N. and NEWHOUSE, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, **1** (2), 115–126.
- , —, — and — (1984). Choosing between the sample-selection model and the multi-part model. *Journal of Business and Economic Statistics*, **2** (3), 283–289.
- GILLESKIE, D. B. and BLAU, D. M. (2006). Health insurance and retirement of married couples. *Journal of Applied Econometrics*, **21** (7), 935–953.
- and MROZ, T. A. (2004). A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics*, **23** (2), 391 – 418.
- HAY, J. W. and OLSEN, R. J. (1984). Let them eat cake: A note on comparing alternative models of the demand for medical care. *Journal of Business and Economic Statistics*, **2** (3), 279–282.
- KOENKER, R. and HALLOCK, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, **15** (4), 143 – 156.
- LEUNG, S. F. and YU, S. (1996). On the choice between sample selection and two-part models. *Journal of Econometrics*, **72** (1-2), 197 – 229.

- MANNING, W. G., BASU, A. and MULLAHY, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, **24** (3), 465 – 488.
- , BLUMBERG, L. and MOULTON, L. H. (1995). The demand for alcohol: The differential response to price. *Journal of Health Economics*, **14** (2), 123 – 148.
- , DUAN, N. and ROGERS, W. H. (1987). Monte carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, **35** (1), 59 – 82.
- and MULLAHY, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, **20** (4), 461 – 494.
- MCDONALD, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, **52** (3), 647–663.
- MIRA, P. (2007). Uncertain Infant Mortality, Learning, and Life-Cycle Fertility. *International Economic Review*, **48** (3), 809–846.
- NORTON, E. C. (2007). HPA 883 Class notes: Analysis of categorical data.
- RUST, J. and PHELAN, C. (1997). How Social Security and Medicare Affect Retirement Behavior In a World of Incomplete Markets. *Econometrica*, **65** (4), 781–831.
- YATCHEW, A. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature*, **36** (2), 669–721.