TECHNICAL WORKING PAPER SERIES

GENERALIZED MODELING APPROACHES TO RISK ADJUSTMENT OF SKEWED OUTCOMES DATA

Willard G. Manning
Anirban Basu
John Mullahy

Technical Working Paper 293
http://www.nber.org/papers/T0293

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2003

We would like to thank Mindy Drum, Alberto Holly, Joe Hilbe, Dan Polsky, Paul Rathouz, and Frank Windmeijer for their help and comments. The opinions expressed are those of the authors, and not those of the University of Chicago, or the University of Wisconsin. This work was supported in part by the National Institute of Alcohol Abuse and Alcoholism (NIAAA) grant 1RO1 AA12664-01 A2. The views expressed in this paper are those of the authors and not necessarily those of the National Bureau of Economic Research.

Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data
Willard G. Manning, Anirban Basu, and John Mullahy
NBER Technical Working Paper No. 293
October 2003
JEL No. I1

ABSTRACT

There are two broad classes of models used to address the econometric problems caused by skewness in data commonly encountered in health care applications: (1) transformation to deal with skewness (e.g., OLS on ln(y)); and (2) alternative weighting approaches based on exponential conditional models (ECM) and generalized linear model (GLM) approaches. In this paper, we encompass these two classes of models using the three parameter generalized gamma (GGM) distribution, which includes several of the standard alternatives as special cases – OLS with a normal error, OLS for the log normal, the standard gamma and exponential with a log link, and the Weibull. Using simulation methods, we find the tests of identifying distributions to be robust. The GGM also provides a potentially more robust alternative estimator to the standard alternatives. An example using inpatient expenditures is also analyzed.

Willard G. Manning
Harris School of Public Policy Studies
The University of Chicago
1155 East 60th Street, Room 176
Chicago, IL 60637
w-manning@uchicago.edu

Anirban Basu
Harris School of Public Policy Studies
The University of Chicago
1155 East 60th Street
Chicago, IL 60637
abasu@midway.uchicago.edu

John Mullahy
Department of Population Health Sciences
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53705
and NBER
jmullahy@wisc.edu

1. INTRODUCTION

Many past studies of health care costs and their responses to health insurance, treatment modalities or patient characteristics indicate that estimates of mean responses may be quite sensitive to how estimators treat the skewness in the outcome (y) and other statistical problems that are common in such data. Some of the solutions that have been used in the literature rely on transformation to deal with skewness (most commonly, OLS on ln(y)), alternative weighting approaches based on exponential conditional models (ECM) and generalized linear model (GLM) approaches, the decomposition of the response into a series of estimation models that deal with specific parts of the distribution (e.g., multi-part models), or various combinations of these. The default alternative has been to ignore the data characteristics and to apply OLS without further modification.

In two recent papers, we have explored the performance of some of the alternatives found in the literature. In Manning and Mullahy (2001), we compared models for estimating the exponential conditional mean – how the log of the expected value of y varied with observed covariates x. That analysis compared OLS on log transformed dependent variables and a range of GLM alternatives with log links under a variety of data conditions that researchers often encounter in health care cost data. In Basu, Manning, and Mullahy (2003), we compared log OLS, the gamma with a log link, and an alternative from the survival model literature, the Cox proportional hazard regression. In both papers, we proposed a set of tests that can be employed to select among the competing estimators, because we found no single estimator dominates the other alternatives or is a close second best.

In this paper, we again compare exponential conditional mean models.[1] Our primary interest is in the marginal effect of a covariate x_1 on E(y|x), where x_1 could be a treatment or behavioral variable of interest.[2] If E(y|x) is exponential conditional mean, then the marginal effect is

$$m_1(x) = \frac{\partial E(y \mid x)}{\partial x_1} = \beta_1 \, e^{\beta_0 + \beta_1 x}$$

which is nonlinear in x. But if we log both sides, then we can summarize the marginal effect by

$$\frac{\partial \ln(E(y \mid x))}{\partial x_1} = \frac{\partial \ln(m_1(x))}{\partial x_1} = \beta_1$$

In what follows, we focus on this as a summary of the response of y to x.

This time, we examine regression modeling using the generalized gamma distribution. The generalized gamma is appealing because it includes several of the standard alternatives as

[1] This focus rules out situations where the analyst is interested in some latent variable construct.
[2] In practice, the vector of covariates x may include other explanatory variables.

special cases – OLS with a normal error, OLS for the log normal, the standard gamma and exponential with a log link, and the Weibull. We see two potential advantages to using this distribution. First, it provides nested comparisons for some alternative estimators, and hence a formal alternative to the somewhat cumbersome and incomplete testing procedure in Manning and Mullahy (2001). Second, if none of the standard approaches is appropriate for the data, then the generalized gamma provides an alternative estimator that will be more robust to violations of distributional assumptions.

The plan for the paper is as follows. In the next section, we describe the generalized gamma in greater detail, showing its connection to more commonly used estimators. Section 3 describes the general modeling approaches that we consider, and our simulation framework. Section 4 summarizes the results of the simulations and examines an application: (1) a study of inpatient expenditures that we have used in previous papers. The final section contains our discussion and conclusions.

2. GENERALIZED GAMMA MODELLING FRAMEWORK

We confine our discussion here to the case with strictly positive values of y to streamline the analysis. We do not address issues related to truncation, censoring, or the "zeros" aspects of data (or "part one of a two-part model"). The focus is on the exponential conditional mean (log link) model because of its widespread use in health economics and health services research. However, the estimation approaches examined here can be extended to include Box-Cox models and alternative power links for GLM and generalized gamma models.

Our modeling framework compares the generalized gamma estimator to several alternative estimators that are most commonly used to model health care costs. We give a list of these alternative estimators below. But before that, we describe the generalized gamma distribution in detail. The generalized gamma distribution has one scale parameter and two shape parameters. This form is also referred to as the family of generalized gamma distributions because the standard gamma, Weibull, exponential and the log normal are all special cases of this distribution. Hence, it provides a convenient form to identify the data generating mechanism of the dependent variable and in turn helps to select the best estimator by applying maximum likelihood methods to estimate a regression model based on the generalized gamma distribution.

2.1. The Standard Version.

The probability density function for the generalized gamma is parameterized as a function of κ, μ, and σ:

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

where $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)\{\ln(y) - \mu\}/\sigma$, and $u = \gamma \exp(|\kappa|z)$.[3] The parameter μ is replaced by $x'\beta = \beta_0 + \beta_1 x_1$, where x is the matrix of the covariates including an intercept, and the β's are coefficients to be estimated. As an extension, we can allow σ to also depend on x.

For the generalized gamma distribution, the expected value of y conditional on x is given by:

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

The other moments of this distribution are:

$$r^{th} \text{ moment} = E(y^r) = \{\exp(\mu)\cdot\kappa^{\frac{2\sigma}{\kappa}}\}^r \cdot \left\{\Gamma\{(1/\hat{\kappa}^2) + (r\hat{\sigma}/\hat{\kappa})\}/\Gamma(1/\hat{\kappa}^2)\right\}$$

$$\begin{aligned}\text{Variance} &= E(y^2) - E(y)^2 \\ &= \{\exp(\hat{\mu})\cdot\hat{\kappa}^{\frac{2\hat{\sigma}}{\hat{\kappa}}}\}^2 \left\{\left[\Gamma((1/\hat{\kappa}^2) + (2\hat{\sigma}/\hat{\kappa}))/\Gamma(1/\hat{\kappa}^2)\right] - \left[\Gamma((1/\hat{\kappa}^2) + (\hat{\sigma}/\hat{\kappa}))/\Gamma(1/\hat{\kappa}^2)\right]^2\right\}\end{aligned} \quad (3)$$

We can also extend the GGM to allow for heteroscedasticity (GGM-het) by parameterizing $\ln(\sigma)$ as $\alpha_0 + \alpha_1 \ln(f(x))$ so that σ is estimated as $\hat{\sigma} = (1/n)\sum_i \exp\{\hat{\alpha}_0 + \hat{\alpha}_1 \ln(f(x_i))\}$. The marginal effect of a covariate (x_k) on the expected value of y is then given by:

$$\frac{\partial \ln(E(y|x))}{\partial x_k} = \hat{\beta}_k + \frac{\hat{\kappa}}{\ln(\hat{\kappa}^2)}\cdot\frac{\partial\hat{\sigma}}{\partial x_k} + \frac{\Gamma'(z)}{\Gamma(z)}\cdot\frac{\partial\hat{\sigma}}{\partial x_k} \quad (4)$$

where $z = [(1/\hat{\kappa}^2) + (\hat{\sigma}/\hat{\kappa})]$, $\partial\hat{\sigma}/\partial x = \hat{\sigma}[\hat{\alpha}_1 f'(x)/f(x)]$, and $\Gamma'(z)/\Gamma(z)$ is the digamma function. When σ is not modeled as a function of x, then $\partial\ln(E(y|x))/\partial x = \hat{\beta}$.

2.2. Special cases.

The specific values for the shape parameters of the generalized gamma distribution yield several possible distributions as special cases. Table I lists the special cases. Using the maximum likelihood estimates of parameter σ and κ and the likelihood function, we can perform hypothesis tests of the appropriateness of each special case.

---

[3] This formulation is consistent with the formulation used by Stata Corp Inc., Version 7. An alternative formulation of the three-parameter generalized gamma distribution was proposed by Stacy (1962) and the form commonly used in practice was suggested by Stacy and Minhram (1965). Appendix A contains a crosswalk between the form in (1) and the Stacy and Minhram parameterization.

This formulation can also be modified to deal with a series of issues.

2.2.1. Heteroscedastic log normal distribution. The error terms in models for ln(y) are often heteroscedastic in at least one of the covariates. In such situations, heteroscedastic retransformation of log-scale prediction from OLS based model is necessary to obtain unbiased estimators of E(y|x) (Manning 1998). If ln(y) ~ Normal ($\mu=x\delta$, $\sigma^2=f(x)$), then $E(y|x) = \exp(x\delta + 0.5f(x))$. The generalized gamma regression provides an opportunity to simultaneously model both the full response of E(y) to covariates x. Thus, a direct test of the presence of heteroscedasticity can be performed with the parameter estimates of the model. For example, in the generalized gamma regression if $\ln(\sigma)$ is parameterized as $\alpha_0 + \alpha_1 \ln(f(x))$, the test of $\alpha_1 = 0$ is a test for heteroscedasticity on the log-scale, as long as α_1 can be identified with respect to the specification used in the main model. Moreover, E(y|x) can be obtained directly using parameter estimates of the model without any retransformation.

2.2.2. Mixture models. Some studies deal with dependent measures and error terms that are heavier tailed (on the log-scale) than even the log normal. In these scenarios, a mixture of log-normals may better approximate the appropriate distribution. However, GLM models tend to be inefficient in the presence of heavier tails. Log OLS models seem to provide a more precise fit to these data (Manning and Mullahy 2001), baring other problems. We expect that the generalized gamma regression to have results equivalent to the log OLS.

However, if we can identify the process behind the generation of the mixture, then we can incorporate these into the specification of the generalized gamma regression. For example, let the error (ε) on the log scale be a mixture of two normal distributions, $N(0, v_1)$ and $N(0, v_2)$. Let δ_i be an indicator for the first of the two distributions. Then, $\varepsilon_i \mid \delta_i \sim N(0, \delta_i v_1 + (1-\delta_i)v_2)$. δ_i could be stochastic (e.g. Bernoulli) or deterministic (a dummy variable). If this δ_i is observable, then one can model $\ln(\sigma) = \alpha_0 + \alpha_1\delta$, indicating that $\exp(2\alpha_0) = v_2$ and $\exp(2\alpha_0 + 2\alpha_1) = v_1$. Significant efficiency gains may accrue from such a formulation if the error terms are homoscedastic in x, but heavy-tailed on the log scale.

To achieve these gains, the δ_i must be observable, rather than latent. However, both in the case of heteroscedasticity and mixture model, the true variance function are seldom known *a priori.* Nevertheless, modeling $\ln(\sigma)$ as a linear function of observable covariates may overcome the biases in many applications and may also result in efficiency gains compared to GLM models.

3. **METHODS**

To evaluate the performance of the generalized gamma estimator, we rely on Monte Carlo simulation of how this estimator behaves over a range of data circumstances and compare it with the behavior of alternative estimators from the literature, including one that is optimal in terms of bias and efficiency for the given data generating mechanism. We consider a broad range of data circumstances that are common in health economics and health services research. They are: (1) skewness in the raw-scale dependent variable y; (2) a log-scale error that is heteroscedastic; (3) a pdf that is monotonically declining rather than bell shaped; and (4) heavy tailed distributions (even after the use of log transformations to reduce skewness on the raw scale). This set of generating mechanisms includes many of the alternatives from Manning and Mullahy (2001).[4] The following describes the data generating processes that exhibit these four properties, and each of the estimation methods that we use to estimate the mean response E(y|x)).

3.1. Data generating processes.

In this work, we consider several different data generating processes that yield strictly positive outcomes that are skewed to the right and exhibit the exponential conditional mean property. They differ in their degree of skewness, kurtosis and also in their dependence on a linear combination of covariate x. We evenly spaced the single covariate x over the [0, 1] interval.[5] The first data generating processes is $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$, ε is N(0, v) with variance v = 1.0 and 2.0. The greater the error variance, the more skewed y becomes on the raw scale. $E(x'\varepsilon) = 0$ and β_1 equals 1.0, which is also the slope of the log of the marginal effect. The value of the intercept (β_0) is selected so that the unconditional mean of y is one.

Heteroscedasticity in the log-scale error term of a linear specification for E(ln(y|x)) is a common feature in health economics data. Estimates based on OLS on the log-scale can provide a biased assessment of the impact of the covariate x on E(y|x) (Manning 1998). In this case, the constant variance v from above is replaced by some log-scale variance function v(x). The expectation of y on the raw-scale becomes: $E(y|x) = \exp(\beta_0 + \beta_1 x + 0.5v(x))$. To construct the heteroscedastic log normal data, we generate the error term ε as the product of a N(0, 1) variable and either (1+x) or its square root. The latter has error variance that is linear is x (v = 1+x),

---

[4] We do not deal with either truncation or censoring. Nor do we consider models based on survival methods, such as those with the proportional hazards property; see Basu, Manning, and Mullahy (2003) for a comparison of survival based estimators with exponential conditional mean estimators.

[5] For each sample, there are 10 subsamples of 1000 with values for x, with x evenly spaced at the 0.001 times the observation number, less 0.0005

while the former is quadratic in x ($v = 1 + 2x + x^2$). Again, β_1 equals 1.0 and β_0 is selected so that $E(y) = 1$.

The third data generating process is based on the gamma distribution. The gamma has a pdf that can be either monotonically declining throughout the range of support or bell shaped, but skewed to the right. The pdf for the standard gamma variate is given in Table I. The scale parameter μ equals $\beta_0 + \beta_1 x$, where β_1 equals 1.0 and β_0 is selected so that $E(y|x) = 1$. The shape parameter $1/|\kappa|^2 = 0.5$, 1.0, or 2.0. The first and second values of the shape parameter yield monotonically declining pdfs conditional on x, while the last is the bell-shaped but skewed right. If the shape parameter equals 1.0, then we have the exponential distribution.

We also consider data generating process based on the Weibull distribution, which (like the exponential distribution) exhibits both exponential conditional mean and proportional hazard properties. The Weibull variate has two parameters. The scale parameter μ equals $\beta_0 + \beta_1 x$, where β_1 equals 1.0 and β_0 is selected so that $E(y) = 1$. We set the shape parameter σ to 0.5, which yields a linearly increasing hazard function with y.

As noted earlier, some studies deal with dependent measures and error terms that are heavier tailed (on the log-scale) than even the log normal. We consider two alternative data generating mechanisms with ε being heavy tailed (kurtosis > 3). In the first, ε is drawn from a mixture of normals, each with mean zero. The ($p \times 100\%$) of the population have a log-scale variance of 1, and $(1-p) \times 100\%$ have a higher variance. In the first case, the higher variance is 3.3 and $p = 0.9$, yielding a log-scale error term with a coefficient of kurtosis of 4.0. In the second case, the higher variance is 4.6 and $p = 0.9$, giving a log-scale error term with a coefficient of kurtosis of 5.0.

Table II summarizes the data generating mechanisms that we consider.

3.2. Estimators.

We employ several different estimators for each type of data generated. The first estimator is a regression model based on the generalized gamma distribution. We employ three versions of the generalized gamma regression. The first version is the regular regression where σ is estimated as a constant; we refer to this model as "GGM." In the second version, we model a 'working' version of the variance function that may not represent the true underlying variance function; we refer to this model as "GGM-het1." Specifically, for GGM-het1, we model $\ln(\sigma)$ as a linear function of x, i.e. $\ln(\sigma) = \alpha_0 + \alpha_1 x$. Both the GGM and GGM-het1 models are run on all data generating processes, whether heteroscedasticity is present or not. Finally, we also employ a third model only for heteroscedastic and heavy tailed data where the true underlying model of σ is used to illustrate the best case scenario; we refer to this model as GGM_het2. Several

methods exist for the estimation of the parameters of this distribution (Hagen and Bain 1970, Lawless 1980, Wingo 1987, Cohen and Whitten 1988, Stacy and Mihram 1965, Balakrishnan and Chan 1994). We employ full-information maximum likelihood method to estimate the parameters of the model. Full information maximum likelihood is implemented in Stata 7's -streg- option, to obtain MLE estimates for β, σ and κ.

The other estimators that we employ include: ordinary least square (OLS) regression of ln(y) on x and an intercept with a homoscedastic smearing factor for the retransformation (Duan 1983); the gamma generalized linear model (GLM) for y with a log link function (McCullagh and Nelder, 1989); and a maximum-likelihood estimator of Weibull model for y.

3.2.1. Least Squares on ln(y). By far the most prevalent estimation approach used in health economics and health services research is to use ordinary least squares or a least-squares variant with ln(y) as the dependent variable. One rationale for this transformation is that the resulting error term is often approximately normal. If that were the case, the regression model would be $\ln(y) = x\beta + \varepsilon$, where x is a matrix of observations on covariates, β is a column vector of coefficients to be estimated, and ε is the column vector of error terms. We assume that E(ε) = 0 and E(x′ε) = 0, but the error term ε need not be i.i.d. If the error term is normally distributed N(0, σ_ε^2), then $E(y|x) = \exp(x\delta + 0.5\sigma_\varepsilon^2)$. If ε is not normally distributed, but is i.i.d., or if exp(ε) has constant mean and variance, then $E(y|x) = s \times \exp(x\beta)$, where s = E(exp($\varepsilon$)).[6] In either case, the expectation of y is proportional to the exponential of the log-scale prediction from the LS-based estimator.

However, if the error term is heteroscedastic in x – i.e. E(exp(ε|x)) is some function f(x) -- then $E(y|x) = f(x) \times \exp(x\beta)$, or, equivalently,

$$\ln(E(y \mid x)) = x\beta + \ln(f(x)) \tag{5}$$

and in the log normal case,

$$\ln(E(y \mid x)) = x\beta + 0.5\sigma_\varepsilon^{\,2}(x) \tag{6}$$

where the last term in Equation 6 is the error variance as a function of x on the log scale (Manning, 1998).

3.2.2. Gamma Models. In GLM modeling, one specifies a mean and variance function for the observed raw scale variable y, conditional on x (McCullagh and Nelder, 1989). Because of the work by Blough, Madden, and Hornbook (1999), we will focus on the gamma regression model with a log link. Like the log normal, the gamma distribution has a variance function that is

[6] Duan (1983) shows that one can use the average of the exponentiated residuals to provide a consistent estimate of the smearing factor.

proportional to the square of the mean function, a property approximately characteristic of many health data sets. The exponential distribution is a limiting case of the standard gamma when $\kappa = 1$.

3.2.3. Weibull Models. The last estimator that we consider is the Weibull, which is frequently used as a parametric alternative for dealing with survival or failure time data. Here, the Weibull is implemented as a GLM model where $E(y|x) = \mu = \exp(x\beta)$ and $v(y|x) = \xi (\mu(x))^2$, where $\xi = [\Gamma(1 + 2\sigma) - \Gamma(1 + \sigma)]$. The Weibull is the only distribution in the generalized gamma family of distributions that has this property.

3.2.4. Estimators Not Considered. In principle, we could have adapted the OLS model for ln(y) to allow for heteroscedastic retransformation (Manning, 1998; Manning and Mullahy, 2001). In the case of a simple form of heteroscedasticity based on a categorical variable or if the underlying log-scale error term was actually normally distributed, then a heteroscedastic retransformation would be a viable alternative. However, if the log-scale error is heteroscedastic in continuous variables or is not normally distributed, then this alternative is cumbersome and difficult to implement. Having explored this alternative earlier, we forego it here.

Similarly, we could have adapted an MLE estimator for ln(y) to allow for a heteroscedastic, normally distributed error. If the log-scale error is heteroscedastic, but not normally distributed, then the estimates from such a model will provide a biased estimates of E(y|x), because that expectation depends on the expected value of the exponentiated log-scale error term, and will not necessarily equal $\exp(0.5\sigma^2(x))$ as it does in the log normal case.

3.3. Design and Evaluation.

Each of the estimators is evaluated on 500 random samples from each of the data generating processes, with each sample having a sample size of 10,000. All models are evaluated in each replicate of a data generating mechanism. This allows us to reduce the Monte Carlo simulation variance by holding the specific draws of the underlying random numbers constant when comparing alternative estimators. The primary estimates of interest are:

(1) The mean, standard error and 95% interval of the simulation estimates of the slope β_1 of ln(E(y)) with respect to x. The mean provides evidence on the consistency of the estimator, while the standard error indicates the precision of the estimate.

(2) The mean residual, to see if there is any overall bias in the prediction of y. The mean provides evidence on the consistency of the overall level of the response.

(3) The bootstrap estimate of the variance of the slope of the (log of the) expected value of y with respect to x provides estimates of the precision of the estimator that is not sensitive to the over-fitting problems in a specific sample.

(4) In evaluating the predictive validity of the alternative estimators, we compare the variance in estimating μ(x) at different values of x across alternative estimators. A plot of standard deviations of μ(x) against x is plotted for each estimator under each data generating mechanism. The pattern in the standard deviations indicates the estimated prediction variance on the raw scale and thus presents a sense of comparative efficiency across estimators.

Finally we also employ all the tests for identifying distributions based on the generalized gamma regression discussed in Section 2. We performed four Wald tests on the parameter and variance estimates of the ancillary parameter. The tests are: a test for the standard gamma ($\exp(\ln(\hat{\sigma}))=\hat{\kappa}$); a test for the log normal ($\hat{\kappa}=0$); a test for the Weibull ($\hat{\kappa}=1$); and, a test for the exponential ($\ln(\hat{\sigma})=0, \hat{\kappa}=1$). We report the proportion of the simulations where the chi-square statistic from each of these tests is significant at the 5% level.

We used Stata 7.0 for all of the estimation. For the generalized gamma, we employed the -- streg -- command in Stata.[7]

4. RESULTS: SIMULATIONS AND AN EMPIRICAL EXAMPLE

4.1. Simulation Results.

Table II provides some of the sample statistics for the dependent measure y on the raw scale across the various data generating mechanisms. As indicated earlier, the intercepts have been set so that the E(y) is 1. For each case, the dependent variable y is skewed to the right and heavy tailed. Table III provides the results on the consistency and precision in the estimate of β_1, the slope of ln(E(y|x)) with respect to x, for each of the alternative estimators for different data generating processes. Appendix B provides tests of the goodness of fit measures for the alternative estimators, including the mean of the raw scale residuals for each estimator for each data generating process. Table IV reports the tests for identifying distributions tests performed after the generalized gamma regressions on each data type. Finally, Figures 1, 2 and 3 shows the relative precision of alternative estimators in predicting y in the raw scale at different values of x.

[7] We have written three Stata ado files that can be used to estimate these models and do the associated tests. They are available from the corresponding author by e-mail request.

4.1.1. Homoscedastic Log Normal Data. All the estimators seem to produce consistent estimates of the slope β_1 for the homoscedastic log normal data (Table III). Log-OLS seems to provide the most precise estimate when compared to the Gamma and Weibull estimators. However, the standard form of the three parameter generalized gamma (GGM) provides equally consistent and precise results as the log OLS. The GGM-het1 model is also consistent and is more precise than the standard Gamma model. This was expected since log normal distribution is a special case of the generalized gamma. On average, the alternative estimators make unbiased predictions, as seen in Appendix B, Table I. Again, the GGM fairs as well as the OLS estimates in terms of bias and goodness of fit measures, with the exception of the very heavy tailed alternatives. The Weibull estimates show a downward bias (under prediction) with higher error variance. For data such as these, the results for OLS estimate based on a logged dependent variable is BLUE. The results for the standard Gamma estimator are consistent (Manning and Mullahy, 2001) but less precise than the OLS estimate based on ln(y). This is especially true at extreme values of x as evident in Figure 1.

The test for log normal ($\kappa = 0$) after the GGM regression was rejected only 7 percent of the times for log-scale error variance of 1 and 6 percent for log-scale error variance of 2 at the 5 percent significant level (Table IV). The tests for Gamma, Weibull and Exponential were rejected for all samples of data such as these.

4.1.2. Heteroscedastic Log Normal Data. As expected, OLS with homoscedastic retransformation yields a biased estimate of the slope ln(E(y|x)) with respect to x (Table III). The standard gamma provides a consistent estimate of the slope, though the consistency comes at some expense of precision. However, the Weibull model seems to provide biased estimates with larger bias for the quadratic variance.

The regular GGM estimate performs exactly like the OLS estimate for ln(y) with homoscedastic smearing and thus provides a biased estimate of the slope. This may come as a surprise when a special case of GGM, the standard gamma GLM is an unbiased estimator. We conjecture that that this anomaly is due to a special feature of generalized gamma distribution and the implementation of GGM in Stata 7. Using a separate simulation framework we found that as the coefficient of skewness of the error on the log scale approaches zero, the MLE for κ also approaches 0. When $\hat{\kappa}$ is close to 0, Stata 7 maximizes the log-normal distribution instead of generalized gamma. Consequently, since heteroscedastic log normal data is symmetric on the log scale, the GGM model gives identical results to a log OLS model. We, therefore suggest cautious interpretation of results from a GGM model when $\hat{\kappa}$ is close to 0 and no hetero correction is applied. However, when we model the random part with the appropriate variance function, the heteroscedastic generalized gamma model (GGM-het2) gives a consistent estimate

of the slope with reasonable precision. Thus, it provides an alternative to some heteroscedastic generalizations of Duan's (1983) smearing estimate. When we use an 'working' variance function (GGM-het1) and not the true one, the heteroscedastic GGM model still gives consistent estimate of the slope and is more efficient than the standard Gamma model. However, the efficiency of the heteroscedastic GGM model will depend on the distribution of x. At higher values of x, the GGM-het1 is more inefficient than the standard Gamma (Figure 1) while the opposite is true at lower values of x.

Even for the heteroscedastic log normal, the test of log normality after the GGM regression seem to fail only 5 percent of the time, whereas other distributions were rejected for all the replicates at the 5 percent significance level (Table IV).

4.1.3. Heavy-tailed Data. The presence of a heavy-tailed error distribution on the log-scale does not cause consistency problems for any of the estimators, but it does generate much more imprecise estimates for the Gamma and Weibull models. The standard errors are about 2 and 4 times larger for the Gamma and Weibull models respectively than the OLS estimate if the kurtosis is 4. These estimates rise to 4 and 10 respectively if kurtosis is 5. The regular GGM produces both an unbiased and precise estimate of the slope. The GGM-het2 (where σ is models as a function of the mixing process) also provides unbiased estimate and only modest precision gain over the regular GGM. The GGM-het1 (where a 'working' variance functions is used) also provides unbiased estimate of the slope with modest precision loss over the regular GGM. The standard Gamma model is highly inefficient for this data generating mechanism especially at the tails of the distribution of x (Figure 2).

Regular GGM predictions tend to be upward biased by about 8 percent if kurtosis is 4 and by 20 percent if kurtosis is 5 (Appendix B Table 1). However, GGM-het2 overcomes this problem and produces consistent predictions. This may be indicative of the difficulty in modeling a mixture distribution. The test of log normality after the GGM regression seems to fail only 5 percent of the time, whereas other distributions were rejected for all the replicates at the 5 percent significance level (Table IV).

4.1.4. Data from the Gamma and Weibull Families. Each of the estimators provides a consistent estimate of the slope for the data generating mechanism of gamma with shapes 0.5 (monotonically declining pdf), 1.0 (exponential distribution) and 2.0 (bell-shaped pdf skewed to the right) and of Weibull with shape 0.5 (linearly increasing hazard). The OLS estimator experienced some precision loss mainly for the gamma with shape 0.5 (Table III and Figure 3). In terms of prediction, all estimators provide unbiased predictions expect that Weibull model tends to over predict at all the deciles of x for gamma with shape 0.5. The GGM does not

provide any evidence for lack of fit as it is the MLE as well as BLUE for these data generating mechanisms.

The tests for identifying distributions correctly identify the gamma or the Weibull data while rejecting all other distributions (Table IV). For the exponential data (gamma with shape 1.0), the tests correctly identifies it as gamma, Weibull and exponential since the exponential distribution is a special case of both gamma and Weibull.

4.2. Choosing an Estimator.

In Manning and Mullahy (2001), we suggested an algorithm for selecting among the exponential conditional mean models that we had examined. The set of checks involved looking at two sets of residuals: (1) the log-scale residuals[8] from a least squares model for ln(y); and (2) the raw-scale residuals from a generalized linear model with a log link. If the log-scale residuals showed evidence that the error was appreciably and significantly heteroscedastic, especially if it was heteroscedastic across a number of variables, then the appropriate choice was one of the GLM models. Although the heteroscedastic retransformation used on the Health Insurance Experiment, and discussed in Manning (1998), was a potential solution, it was often too cumbersome to employ. If the residuals were not heteroscedastic, then the choice would depend on whether the log-scale residuals were heavy tailed or the raw-scale residuals exhibited a monotonically declining pdf. If the log-scale residuals were heavy-tailed, but roughly symmetric, then OLS on ln(y) is the more precise estimator. If the raw-scale residuals were monotonically declining, then one of the GLM alternatives, possibly the gamma, was appropriate. Finally, one could use the squared raw-scale residual in a modified Park test to determine the appropriate family (distribution) function among the GLM alternatives.

This algorithm did not deal with certain situations. If the log-scale residuals are symmetric, heavy-tailed, and heteroscedastic, then OLS without suitable heteroscedastic retransformation will be biased. But a suitable retransformation is often difficult to execute. The GLM alternatives will be unbiased, but suffer substantial losses in precision.

One of the motivations for the current analysis was to examine the generalized gamma as a formal alternative to this earlier algorithm. We in fact set up a program to execute the algorithm above, modified so that heteroscedasticity always leads to the choice of a GLM model, monotonically pdfs (otherwise) lead to GLM, and heavy tails (but homoscedastic on the log-scale) lead to OLS on ln(y). The results indicate that the generalized gamma alternative did

[8] We would suggest using the standardized or studentized residuals rather than the conventional residuals $e = \ln(y) - xb$, where b is the OLS estimate of β. The OLS residual is heteroscedastic, by construction, even when the true error ε is not. The variance-covariance matrix for the least squares residual is $\sigma^2 (I-X(X'X)^{-1}X')$.

better over a range of data generating functions that were characterized by log-scale homoscedasticity, but asymmetric log-scale residuals. In particular, the earlier algorithm would often choose OLS on ln(y) over the gamma regression alternative when the true data generating function was a gamma with a log link and a shape parameter greater than one. The generalized gamma model, which includes both the log normal and the gamma with log link as special cases, never made this mistake.

As a result, we would suggest that anyone using the earlier algorithm and its rule about heavy tails require that the log-scale residuals be roughly symmetric before choosing OLS on ln(y). Alternatively, we suggest using the generalized gamma and employing the tests used in this paper, including those in Appendix B.

4.3. Empirical Example – The University of Chicago Hospitalist Study.

We use data from a study of hospitalists that is currently being conducted at the University of Chicago by Meltzer, Manning, et al. [2002]. Hospitalists are attending physicians who spend three months a year attending on the inpatient words, rather than the one month a year typical of most academic medical centers. The policy issue is whether hospitalists provide less expensive care or better quality of care than the traditional arrangement for attending physicians. The evidence to date suggests that costs and length of stay are lower. The behavioral issue in Meltzer, Manning et al. is whether these differences are due to increased experience in attending on the wards – as experience (number of cases treated) increases, do expenditures fall? Does the introduction of a covariate for total experience and one for experience with the disease specific to that patient eliminate the explanatory power of the indicator for the hospitalists?

The data cover all admissions over a twenty-four month period. All patients are adults drawn from medical wards at the University of Chicago. Patients were assigned in a quasi-random manner based on date of admission. The hospitalist and non-hospitalist attending teams rotated days in fixed order through the calendar, ensuring a balance of days of the week and months across the two sets of attending physicians. There is no evidence of significant or appreciable differences in the two groups of patients in terms of demographics, diagnoses or other baseline characteristics. The sample size is 6511 cases for length of stay analyses and 6500 for inpatient costs. We deleted eleven cases because of missing values for the inpatient expenditure variable.

The hospitalist study shows that there were no differences in cost per stay between the two groups of attending physicians at the beginning of the study. This indicates that there were no significant or appreciable differences in baseline skills or experience between the hospitalist and traditional attending teams. Instead, it appears that the differences evolve over time and are directly related to experience to the date of admission of the observation. To illustrate the

alternative estimators, we re-estimate the models from the earlier study using inpatient (facility) expenditures as dependent variables, and the following estimators: Ordinary least squares on ln(y), Gamma regression with a log link, Weibull regression with a log link, and the Generalized Gamma estimator. Table VI provides the estimates of the coefficients for the indicator for the hospitalist variable, the overall measure of experience-to-date, and the disease specific measure of experience-to-date. We have suppressed the estimates of the coefficients of the other variables. The standard errors reported are robust estimates using the appropriate analog of the Huber/White correction to the variance/covariance matrix.

The results indicate that the coefficient on the hospitalist variable is not significantly different from zero once we correct for the inherent differences in experience between hospitalists and conventional atttendings. There are two interpretations of the insignificant hospitalist coefficient. First, hospitalists have no further effects on costs, except through their experience variable. Second, at the beginning (no experience), they were not different from non-hospitalists in their costs. Further, it is disease specific experience, not total experience that matters. These conclusions are unaffected by choice of estimator.

The different estimators estimate different estimates of the magnitudes of the experience response. As a result, the results in Table VI are not directly comparable. First, the OLS on ln(y) estimates are really about the geometric mean. Because the error term is heteroscedastic, these estimates are inconsistent in terms of the natural log of E(y|x). Second, the Gamma and Weibull models do provide consistent estimates of the natural log of E(y|x). Finally the generalized gamma regression models the deterministic part and the random part separately and hence provides a consistent estimate of log E(y|x) when estimates from both these part are taken into account.

To make the results directly comparable, we predicted what each estimator would predict would be the results on the raw-scale of y – inpatient dollars. In Table VI, we also provide the sample means of inpatient expenditures based on the deciles of disease-specific experience. For the OLS on ln(y), we used a homoscedastic smearing factor because we could determine no simple fix for the complex heteroscedasticity in the OLS residuals on the log scale. In Appendix B, we also provide some tests of model fit. Also, for the generalized gamma model, we report whether any particular distribution is identified by testing the ancillary parameters.

The regular GGM produces results identical to log OLS model in terms of slope and goodness of fit test. However, the average residual from prediction is about 15 times lower than that of log OLS. The test of log normality fails to reject the log normal distribution. A heteroscedastic version of GGM is fitted by modeling $\ln(\sigma) = \alpha_0 + \alpha_1 LNCNT2 + \alpha_2 LND3CNT2$. This indicates that we assume that the heteroscedasticity is of the form: $\sigma^2 = K_1 (CNT2)^{K2} (D3CNT2)^{K3}$, where CNT2 is cumulative disease specific experience to date and D3CNT2 is

specific experience-to-date. Though model fit with GGM-het was not much different than in regular GGM, the slopes of LNCNT2 and LND3CNT2 were comparable to the gamma regression with log link.

5. CONCLUSIONS

In Manning and Mullahy (2001), we explored the performance of alternative least squares and generalized linear model estimators for the response of the expected value of y to a set of covariates x under a range of data generating processes. No single estimator was dominant or nearly dominant under all circumstances. But two patterns were clear. First, least squares could provide biased estimates of the mean response of the (untransformed) outcome variable if there was heteroscedasticity in the log scale error. Second, the GLM models would be unbiased but could be quite imprecise if the log-scale error was symmetric but heavy-tailed or if the log scale error variance is large (>1). We proposed a set of tests that would allow analysts to choose among the competing exponential conditional mean (ECM) models.[9]

This paper takes a different approach. It has considered the estimation of a regression model using maximum likelihood for a specific distribution – the generalized gamma -- that includes some of the ECM estimators, notably the gamma and the log normal, as special cases. Using similar simulation comparisons to our two earlier papers, we find that the GGM performs well against the special cases. It handily rejects alternatives that do not apply to a specific data generating mechanism – for example, the log normal when the data are generated from a gamma with shape less than or equal to one. It rarely rejects the correct distribution that applies. The estimates provided by the GGM are consistent for the log-scale slope and almost as precise as the appropriate model for that data generating process. The one exception to the consistency for the ECM data generating processes is the case where $\ln(y) = x\beta + \varepsilon$, where ε is heteroscedastic in x. This appears to be the result of the GGM selecting a κ close to zero because of the symmetry in the log scale error term. Under these circumstances, the GGM estimates the log normal model, ignoring the heteroscedasticity. This anomaly can be remedied by allowing the GGM to have a heteroscedastic error.

---

[9] The results here confirm the earlier results and indicate how well the GG model works when the data generating process satisfies the proportional hazard assumption. . In Basu, Manning, and Mullahy (2002), we also considered a set of alternatives derived from the literature on survival models with proportional hazard (PH) assumptions. We provided a set of tests to choose between the two quite different exponential approaches that would allow a test of ECM vs PH alternatives.

Unlike the gamma with log link, the GGM can estimate the heavy-tailed alternatives without a noticeable loss in precision. In those heavy-tailed cases, GGM is consistent for the slope of ln(E(y|x) with respect to x, but tends to under predict the overall mean on the raw scale because of bias in the estimate of the intercept.

This approach to choosing among the competing models is more appealing that the algorithm that we proposed in Manning and Mullahy (2001). In practice, that algorithm often chooses the log normal when the true model is gamma with log link. The new GGM approach thus deals well with the range of data generating conditions and problems that was problematic before. Thus, it picks the right special case with high probability. And it does so with little loss of precision in the log-scale slope.

Another advantage of the GGM is that it can be a robust or more general alternative to the two parameter log normal and exponential conditional models when the data do not fit one of the two or one parameter alternatives. Thus, the generalized gamma provides an appealing encompassing model for several of the estimators that have been proposed.

But the generalized gamma model is not without some limitations of its own:

- The standard formulation of the generalized gamma is not consistent when the data generating mechanism is the heteroscedastic model for ln(y), such as the examples in Manning (1998), Mullahy (1998), and Manning and Mullahy (2001). This is in contrast to the results for the conventional gamma model, which is consistent (but inefficient) under those circumstances. When such heteroscasticity is present on the log scale, the GGM must be adapted to handle heteroscedasticity, as we have done here to produce consistent and reasonably precise estimates.
- Although it can be adapted to deal with heteroscedasticity and mixture models, the GGM does not have the robust, less parametric alternative like Duan's smearing estimator smearing estimator for the least squares model. But if the link function and the specification of the covariates in x are appropriate, then the choice of the wrong distribution/family does not lead to bias in the parameter estimates.
- The GGM is not a full substitute for a careful examination of the model to see if the data exhibit the pattern we would expect of this class of models. Nor is it a substitute for a careful examination of linearity, functional form, and the link function. In a related paper, Basu and Rathouz (2003) extend the formulation of GLM models to select the power for the link and variance functions (distribution) simultaneously. They show the nature in the bias from selecting the wrong link function.
- Our concern here has been with modeling the mean response of the outcome variable to changes in the covariates, or some function of the mean, such as the marginal or

incremental effect of some covariate, controlling for other variables. Many applications will require more attention to the distribution functions because the analyst is interested in a different task – such as the probability that the outcome will exceed some critical or policy important threshold. For such analyses, the distributions studied here may not provide a close enough approximation.

Nevertheless, we hope that the alternatives considered here are useful in helping analysts to deal with data on outcomes that are inherently skewed, but may not necessarily fall into certain simple situations. Those applications include the analysis of expenditures on health and other commodities and services, earnings, and many other economic outcomes which are often very skewed to the right. Given the potential for bias and inefficiency in standard approaches, the GGM and its heteroscedastic adaptation provide a fresh, more flexible alternative.

References

Balakrashnan N, and Chan PS. 1994. Maximum likelihood estimation for the three parameter log-gamma distribution. In *Recent Advances in Life Testing and Reliability*. Boca Raton, FL.

Basu, A., W.G. Manning, and J. Mullahy. 2003. Comparing alternative model: Log vs Cox proportional hazard? Draft, University of Chicago.

Basu, A. and P. Rathouz. 2003. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models, Draft, University of Chicago.

Blough, D.K., C.W. Madden, and M.C. Hornbrook. 1999. Modeling risk using generalized linear models. *Journal of Health Economics* 18: 153-171.

Cohen AC, and Whitten BJ. 1986. Modified moment estimation for the three-parameter gamma distribution. *Journal of Quality Technology* 17:147-154.

Cox D.R. 1972. Regression Models and life-tables. *Journal of the Royal Statistical Society* B 34: 187-200.

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

Hager HW, and Bain LJ. 1970. Inferential procedures for the generalized gamma distribution. *Journal of the American Statistical Association* 65:1601-1609.

Hosmer, D.W., and S. Lemeshow. 1995. *Applied Logistic Regression,* 2nd Edition. New York, John Wiley & Sons.

Lawless JF. 1980. Inference in the generalized gamma and the log gamma distribution. *Technometrics* 22: 409-419.

Manning, W.G. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 17: 283-295.

Manning, W.G., and J. Mullahy. 2001. Estimating log models: To transform or not to transform? *Journal of Health Economics* 20(4): 461-494.

McCullagh, P. and J.A. Nelder. 1989. *Generalized linear models*, 2nd Edition. London: Chapman and Hall.

Meltzer, D.O., W.G. Manning, J. Morrison, T. Guth, A. Hernanadez, A. Dhar, L. Jin, and W. Levinson. 2002. Effects of hospitalist physicians on an academic general medicine service: results of a randomized trial. *Archives of Internal Medicine,* 137(11): 866-874..

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

Pregibon, D. 1980. Goodness of link tests for generalized linear models. *Applied Statistics*, 29: 15-24.

Pregibon, D. 1981 Logistic regression diagnostics. *Annals of Statistics* 9: 705-724.

Stacy, E.W. 1962. A generalization of gamma distribution. *Annals of Mathematical Statistics* 33: 1187-1192.

Stacy, E.W., and Mihram, G.A., 1965. Parameter estimation for a generalized gamma distribution. *Technometrics* 7: 349-358.

Wingo DR. 1987. Computing maximum-likelihood parameter estimates of the generalized gamma distribution by numerical root isolation. *IEEE Transactions on Reliability* R-36:586-590.

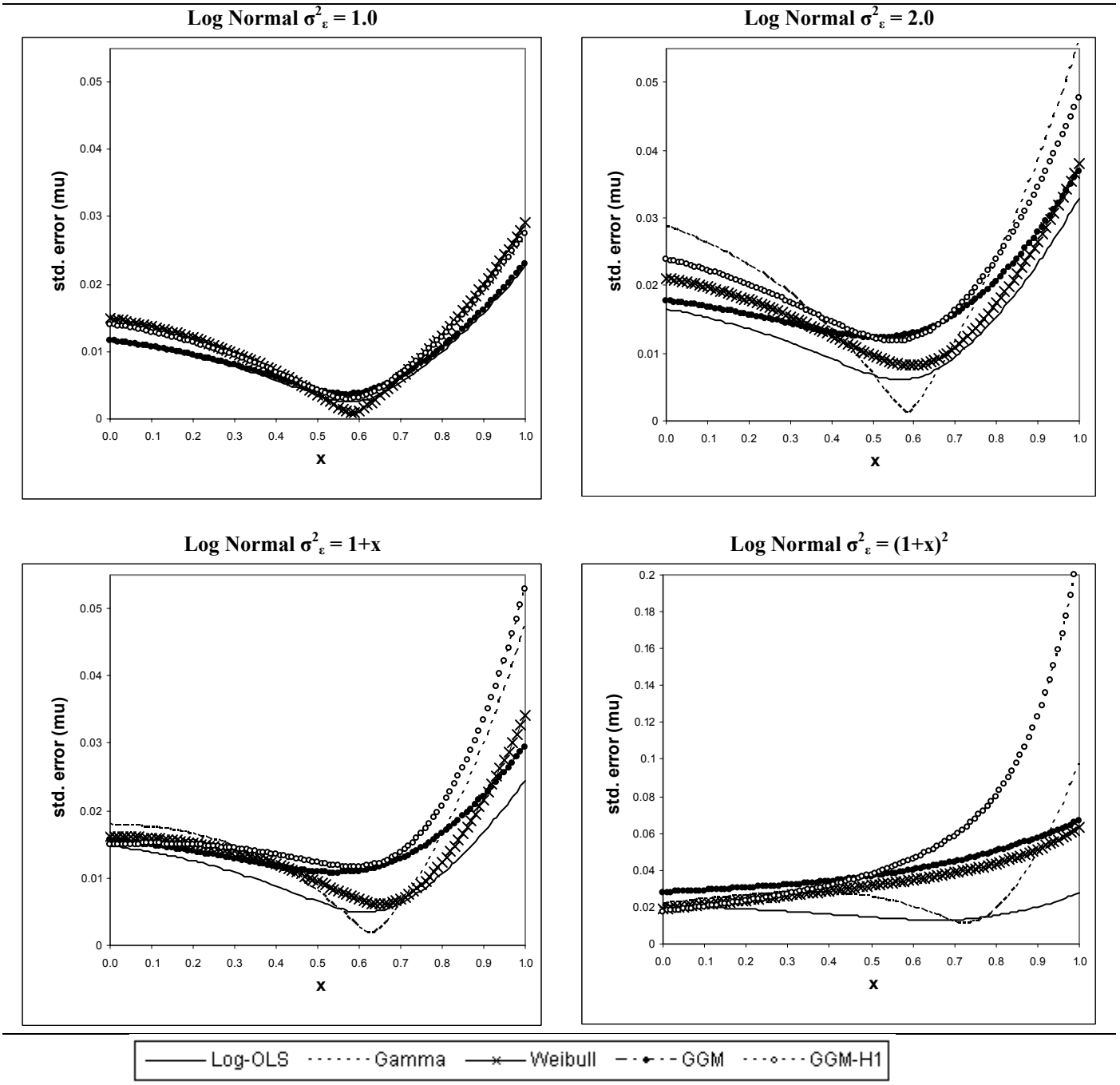

Figure 1: Standard error of predictions from different estimators across different values of 'X' for Log Normal data with and without heteroscedasticity.

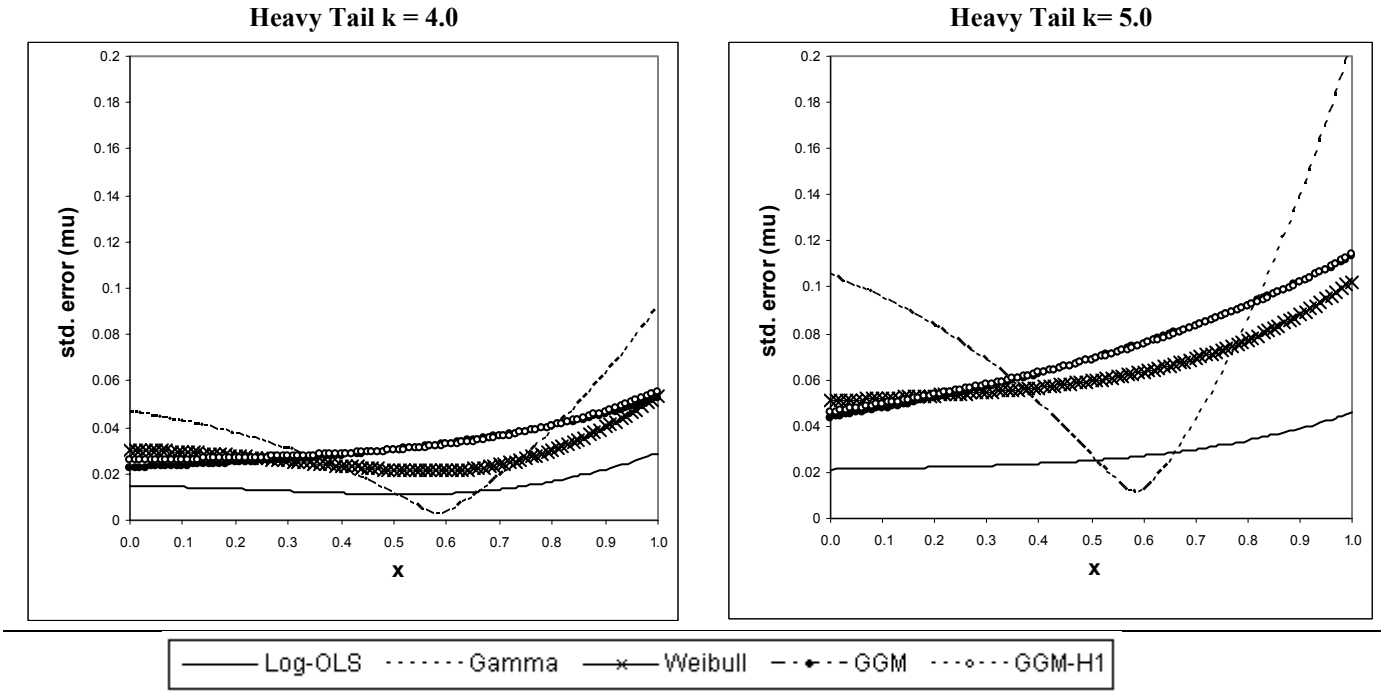

Figure 2: Standard error of predictions from different estimators across different values of 'X' for data with heavy tails.

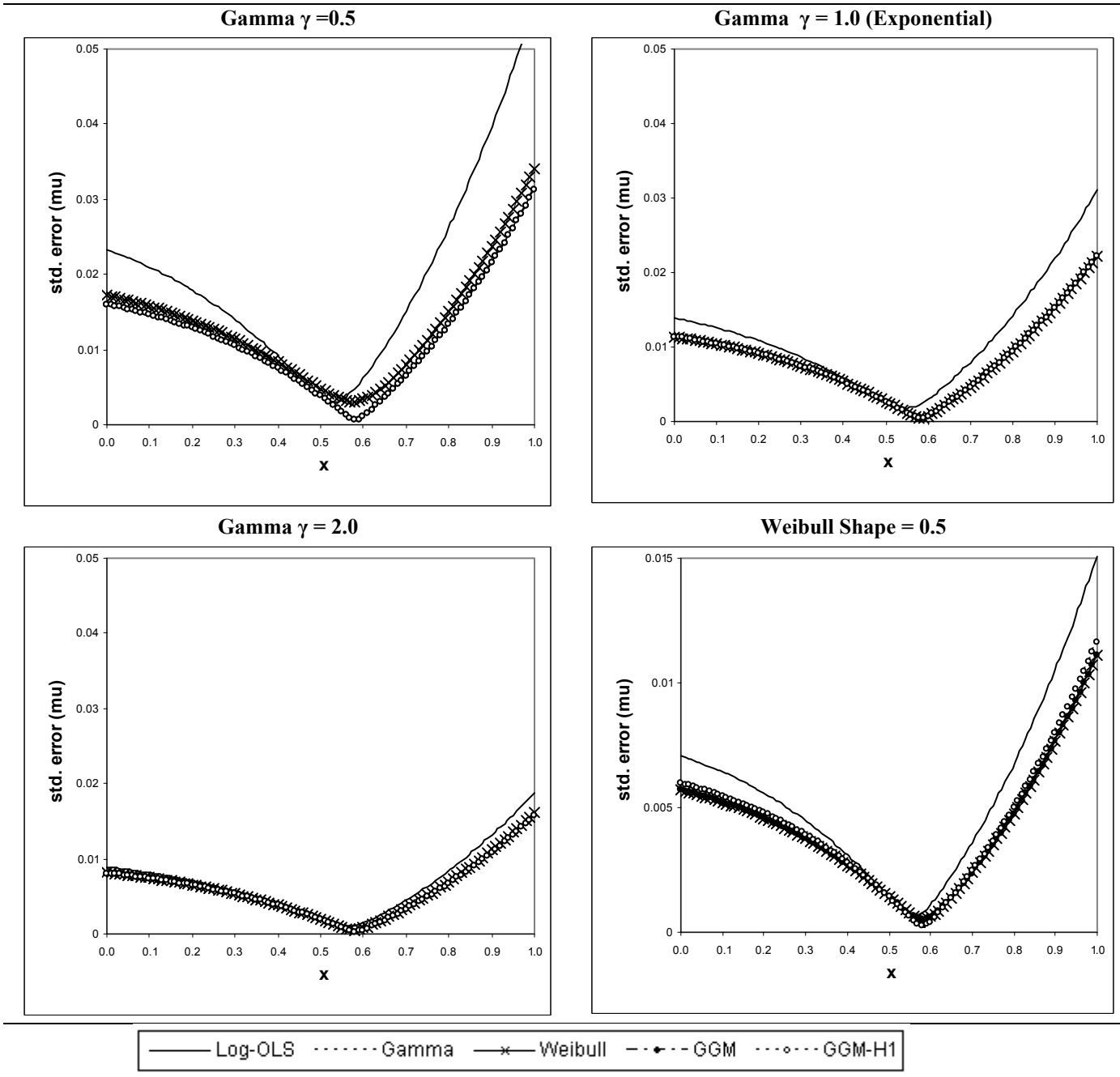

Figure 3: Standard error of predictions from different estimators across different values of 'X' for Gamma, Exponential and Weibull data.

Table I: Special Cases of Generalized Gamma Distributions

Ancillary Parameters		Distribution	Probability Density Function
σ	**κ**		
>0	*	Generalized Gamma	$\frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\Gamma(\gamma)}\exp\left[z\sqrt{\gamma}-u\right]$
= κ	= σ, > 0	Standard Gamma	$\frac{\gamma^{\gamma}}{y\Gamma(\gamma)}\exp\left[z\sqrt{\gamma}-\gamma\exp(\sigma z)\right]$
*	= 1	Weibull	$\frac{1}{\sigma y}\exp\left[z-\exp(z)\right]$
= 1	= 1	Exponential	$\exp\left[\frac{y}{\exp(\mu)}-\mu\right]$
*	→ 0	Log Normal	$\frac{1}{\sigma y\sqrt{2\pi}}\exp\left[-z^2/2\right]$

where, $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)\{\ln(y) - \mu\}/\sigma$, and $u= \gamma\exp(|\kappa|z)$; *= can take can value on $\Re$.

Table II: Monte Carlo Simulation Design and Sample Statistics for Distributions.

Alternative Data Generating Processes		Descriptive statistics on y			
Data (y)	**Brief Description**	**Mean***	**S.D.**	**Skewness**	**Kurtosis**
Log Normal $\sigma^2_\varepsilon = 1$	ECM only	1.0	3.95	6.34	99.36
Log Normal $\sigma^2_\varepsilon = 2$	ECM only	1.0	12.18	15.17	516.7
Heteroscedastic $\sigma^2_\varepsilon = 1+x$	ECM only	1.0	9.06	15.14	529.7
Heteroscedastic $\sigma^2_\varepsilon = (1+x)^2$	ECM only	1.0	39.63	33.55	1902.5
Gamma, shape = 0.5	Monotonically declining pdf, ECM only	1.0	2.58	3.28	20.2
Gamma, shape = 1.0	Exponential Model, ECM & PH	1.0	1.86	2.41	12.3
Gamma, shape = 2.0	Bell-shaped pdf, ECM only	1.0	1.36	1.81	8.3
Weibull, shape = 0.5	Linearly increasing hazard, ECM & PH	1.0	0.94	1.15	4.69
Heavy tailed, k=4.0	Mixture of Normals with v=1.0 & 3.3, p=0.9	1.0	13.42	35.71	2124.2
Heavy tailed, k=5.0	Mixture of Normals with v=1.0 & 4.6, p=0.9	1.0	37.41	47.72	3187.1

* Standardized so that mean(y) =1. ECM = Exponential conditional mean property. PH = proportional hazards property.
k = log-scale coefficient of kurtosis; σ_ε^2 = log-scale variance.

Table III: Effect of alternative estimator on log-scale coefficient on slope of ln(E(y|x)).

Data	Estimator	Mean Slope	S.E.	95% simulation interval	
				Lower	Upper
Log Normal	OLS for ln(y)	1.0006	0.0338	0.9407	1.0710
$\sigma_\varepsilon^2 = 1.0$	Gamma	1.0003	0.0440	0.9178	1.0856
(true = 1.0)	Weibull	1.0003	0.0440	0.9176	1.0860
	GGM	1.0006	0.0338	0.9410	1.0703
	GGM – het(1)	1.0005	0.0412	0.9169	1.0835
Log Normal	OLS for ln(y)	1.0009	0.0478	0.9162	1.1004
$\sigma_\varepsilon^2 = 2.0$	Gamma	0.9999	0.0847	0.8445	1.1711
(true = 1.0)	Weibull	1.0004	0.0622	0.8835	1.1215
	GGM	1.0009	0.0478	0.9165	1.0994
	GGM – het(1)	1.0005	0.0672	0.8643	1.1317
Log Normal	OLS for ln(y)[1]	1.0008	0.0415	0.9375	1.0886
$\sigma_\varepsilon^2 = 1+x$	Gamma	1.5000	0.0662	1.3778	1.6423
(true = 1.5)	Weibull	1.4084	0.0547	1.3100	1.5134
	GGM	1.0007	0.0423	0.9359	1.0888
	GGM – het(1)	1.5129	0.0559	1.3989	1.6269
	GGM – het(2)	1.5021	0.0573	1.3897	1.6162
Log Normal	OLS for ln(y)[1]	1.0011	0.0526	0.9070	1.1069
$\sigma_\varepsilon^2 = (1+x)^2$	Gamma	2.4949	0.1354	2.2722	2.7866
(true = 2.5)	Weibull	1.9815	0.0711	1.8474	2.1184
	GGM	1.0011	0.0598	0.8949	1.1220
	GGM – het(1)	2.5983	0.0893	2.4233	2.7899
	GGM – het(2)	2.5024	0.0805	2.3417	2.6721
Heavy-tailed	OLS for ln(y)	1.0006	0.0378	0.9312	1.0743
k = 4.0	Gamma	0.9973	0.1383	0.7268	1.2966
(true = 1.0)	Weibull	0.9992	0.0808	0.8544	1.1543
	GGM	1.0007	0.0377	0.9314	1.0731
	GGM – het(1)	1.0016	0.0523	0.9029	1.1084
	GGM – het(2)	1.0007	0.0348	0.9380	1.0720
Heavy-tailed	OLS for ln(y)	1.0006	0.0398	0.9249	1.0754
k = 5.0	Gamma	0.9923	0.3110	0.3257	1.7007
(true = 1.0)	Weibull	0.9984	0.1102	0.7912	1.2017
	GGM	1.0007	0.0397	0.9250	1.0756
	GGM – het(1)	1.0022	0.0611	0.8899	1.1228
	GGM – het(2)	1.0008	0.0349	0.9376	1.0728

NOTE: Based on 500 simulations, each with n=10,000. Numbers normalized so that E(y)=1.0. Slope parameter GGM – het models obtained via numerical methods.
[1] With heteroscedastic smearing. GGM = Generalized Gamma;
het(1) = hetero. model with $\ln(\sigma) = \alpha_0 + \alpha_1 x$; het(2) = true underlying hetero. model.
k = coefficient of kurtosis, σ^2 =log-scale variance.

Table III (cont'd): Effect of alternative estimator on log-scale coefficient on slope of ln(E(y|x)).

Data	Estimator	Mean Slope	S.E.	95% simulation interval	
				Lower	Upper
Gamma	OLS for ln(y)	1.0015	0.0750	0.8654	1.1504
γ = 0.5	Gamma	1.0006	0.0471	0.9150	1.0950
(true = 1.0)	Weibull	1.0007	0.0491	0.9054	1.1025
	Cox	0.9777	0.0472	0.8868	1.0679
	GGM	1.0005	0.0472	0.9157	1.0950
	GGM – het(1)	1.0006	0.0472	0.9156	1.0944
Gamma	OLS for ln(y)	1.0009	0.0433	0.9236	1.0881
γ = 1.0	Gamma	1.0005	0.0335	0.9364	1.0659
(Exponential)	Weibull	1.0005	0.0335	0.9364	1.0659
(true = 1.0)	GGM	1.0005	0.0335	0.9369	1.0661
	GGM – het(1)	1.0004	0.0335	0.9366	1.0659
Gamma	OLS for ln(y)	1.0005	0.0271	0.9522	1.0548
γ = 2.0	Gamma	1.0004	0.0238	0.9543	1.0496
(true = 1.0)	Weibull	1.0003	0.0241	0.9557	1.0483
	GGM	1.0003	0.0238	0.9540	1.0495
	GGM – het(1)	1.0004	0.0238	0.9544	1.0497
Weibull	OLS for ln(y)	1.0004	0.0217	0.9618	1.0441
σ = 0.5	Gamma	1.0003	0.0177	0.9679	1.0376
(true = 1.0)	Weibull	1.0002	0.0167	0.9682	1.0329
	GGM	1.0002	0.0168	0.9684	1.0331
	GGM – het(1)	1.0003	0.0167	0.9684	1.0370

NOTE: Based on 500 simulations, each with n=10,000. Numbers normalized so that E(y)=1.0. Slope parameter GGM – het models obtained via numerical methods.
k = coefficient of kurtosis, σ^2 =log-scale variance. GGM = Generalized Gamma;
het(1) = hetero. model with $\ln(\sigma) = \alpha_0 + \alpha_1 x$;

Table IV: Identifying Distribution Tests from the Generalized Gamma Regression.

Data Generating Mechanism	Proportion significant at 5% level			
	Gamma $\exp(\ln \hat{\sigma}) = \hat{\kappa}$	**Log Normal** $\hat{\kappa} = 0$	**Weibull** $\hat{\kappa} = 1$	**Exponential** $\ln \hat{\sigma} = 0,\ \hat{\kappa} = 1$
Log normal $\sigma^2 = 1$	1.0000	0.0720	1.0000	1.0000
Log normal $\sigma^2 = 2$	1.0000	0.056	1.0000	1.0000
Log normal $\sigma^2 = 1+x$	1.0000	0.056	1.0000	1.0000
Log normal $\sigma^2 = (1+x)^2$	1.0000	0.058	1.0000	1.0000
Heavy tailed k = 4	1.0000	0.0520	1.0000	1.0000
Heavy tailed k = 5	1.0000	0.0620	1.0000	1.0000
Gamma shape = 0.5	0.0480	1.0000	1.0000	1.0000
Gamma shape = 1.0	0.0460	1.0000	0.0440	0.0300
Gamma shape = 2	0.0480	1.0000	1.0000	1.0000
Weibull shape = 0.5	1.0000	1.0000	0.0440	1.0000

Note: k = coefficient of kurtosis

Table VI: Alternative Regression Estimates from Hospitalist Study.

Estimator		Coefficient	Robust Std. Err.	z-statistic	P > | t |
Outcome: Inpatient Expenditure					
OLS on ln(y)	hsplist	-.00319	.02920	-0.11	0.913
	lncnt2	-.00723	.01015	-0.71	0.476
	lnd3cnt2	-.04328	.01569	-2.76	0.006
Gamma Regression	hsplist	.00703	.03599	0.20	0.845
with log link	lncnt2	-.00070	.01235	-0.06	0.955
	lnd3cnt2	-.05468	.01905	-2.87	0.004
Weibull Regression	hsplist	.01315	.04145	0.32	0.751
with log link	lncnt2	.00423	.01381	0.31	0.759
	lnd3cnt2	-.06601	.02166	-3.05	0.002
Generalized Gamma	hsplist	-.00267	.02909	-0.09	0.927
Regression	lncnt2	-.00723	.01094	-0.69	0.489
	lnd3cnt2	-.04288	.01597	-2.69	0.007
GGM-het					
lny:	hsplist	-.00314	.02902	-0.11	0.914
	lncnt2	-.00753	.01044	-0.72	0.471
	lnd3cnt2	-.04474	.01587	-2.82	0.005
ln(sigma):	lncnt2	.03180	.00900	3.53	<.001
	lnd3cnt2	-.04539	.01287	-3.52	<.001
lnE(y):	lncnt2	.00916	.01133	0.81	0.419
	lnd3cnt2	-.06854	.01734	-3.95	<.001
		Mean	Std. Error	t	p
Deciles					
1		8717.93	430.80	20.24	0.000
2		10193.44	575.85	17.70	0.000
3		9884.57	686.14	14.41	0.000
4		9066.21	467.62	19.39	0.000
5		8946.39	530.70	16.86	0.000
6		9082.10	473.85	19.17	0.000
7		8098.04	439.66	18.42	0.000
8		7528.42	392.02	19.20	0.000
9		7375.74	413.86	17.82	0.000
10		6384.19	397.41	16.06	0.000

hsplist: Indicator for patient of hospitalist
lncnt2: log cumulative experience–to–date
lnd3cnt2: log disease specific experience–to–date

Table VII: Goodness of Fit on the Raw Scale of Inpatient Expenditures from Hospitalist Study.

Estimator	**Average Residual[1]**	**H.L. F Test (p)**	**Pregibon Test (p)**	**Pearson Correl. (p)**
OLS for ln(y)	-73.63	1.09 (0.36)	-1.62 (0.106)	-0.1387 (<0.0001)
Gamma Regression	-101.37	1.25 (0.26)	-2.04 (.041)	-.1313 (<0.0001)
Weibull Regression	-993.49	7.72 (<0.0001)	-0.07 (0.94)	-0.2212 (<0.0001)
Generalized Gamma	-5.29	0.94 (0.50)	-1.65 (0.10)	-.1304 (<0.0001)
GGM-het	-27.29	0.92 (0.52)	-1.65 (0.10)	-0.1489 (<0.0001)
Tests for identifying Distributions		Chi Sq Statistic	df	p-value
Std. Gamma	$\kappa = \sigma$	303.30	1	<0.0001
Log Normal	$\kappa = 0$	2.45	1	0.1178
Weibull	$\kappa = 1$	587.05	1	<0.0001
Exponential	$\kappa = \sigma = 1$	1273.80	2	<0.0001

Appendix A

We start with (1) which is the formulation of the generalized gamma density function that Stata uses:

$$P(y;\kappa,\mu,\sigma) = \frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\Gamma(\gamma)}\exp\left[z\sqrt{\gamma}-u\right]$$

where, $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)\{\ln(y) - \mu\}/\sigma$, and $u= \gamma\exp(|\kappa|z)$. Rearranging we have,

$$P(y;\kappa,\mu,\sigma) = \frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\,\Gamma(\gamma)}\exp\left[\frac{sign(\kappa)(\ln y-\mu)}{\sigma/\sqrt{\gamma}}-\gamma\exp\left\{\frac{sign(\kappa)(\ln y-\mu)}{\sigma\sqrt{\gamma}}\right\}\right]$$

$$= \frac{\gamma^{\gamma}}{\sigma y\sqrt{\gamma}\Gamma(\gamma)}\left[\frac{y}{\exp(\mu)}\right]^{\frac{sign(\kappa)\gamma}{\sigma\sqrt{\gamma}}}\exp\left[-\gamma\left\{\frac{y}{\exp(\mu)}\right\}^{\frac{sign(\kappa)1}{\sigma\sqrt{\gamma}}}\right]$$

$$= \frac{\left(\frac{1}{\sigma\sqrt{\gamma}}\right)[y]^{\left(\frac{sign(\kappa)1}{\sigma\sqrt{\gamma}}\gamma-1\right)}\exp\left[-\left\{\frac{y}{\exp(\mu)\Big/\gamma^{\left(sign(\kappa)\sigma\sqrt{\gamma}\right)}}\right\}^{\frac{sign(\kappa)1}{\sigma\sqrt{\gamma}}}\right]}{\Gamma(\gamma)\left[\frac{\exp(\mu)}{\gamma^{\left(sign(\kappa)\sigma\sqrt{\gamma}\right)}}\right]^{\frac{sign(\kappa)}{\sigma\sqrt{\gamma}}\gamma}}$$

$$= \frac{|c|(y)^{ca-1}\exp\left[-\left(y/b\right)^{c}\right]}{\Gamma(a)b^{ca}}$$ (Stacy and Minhram, 1965)

where $a = \gamma = 1/|\kappa|^2$,

$b = \exp(\mu)/(1/|\kappa|^2)^{\text{sign}(\kappa)\sigma/|\kappa|} = \exp(\mu)/(1/|\kappa|^2)^{\sigma/\kappa}$, and

$c = \text{sign}(\kappa)|\kappa|/\sigma = \kappa/\sigma$.

Appendix B

We were also interested in the ability of standard tests to pick up model misspecification when the data on the dependent variables were as skewed as the cases considered here. We considered three tests. The first examines the mean of the raw-scale residuals across deciles of x. By looking at the pattern in the residuals as a function of x, we can determine whether there is a systematic pattern of bias in the forecasts. A formal version of this test is provided by a variant of test of goodness of fit proposed by Hosmer and Lemeshow [1995], using an F test that the means of the raw scale residual across all 10 of the deciles are not significantly different from zero. If the residual pattern is u-shaped, then there is evidence for a different nonlinear response than was assumed. We report the proportion of the simulations where the F was significant at the five percent level.

The second is a more parsimonious test for nonlinearity known as Pregibon's Link Test [1981, 1982]. Based on the initial estimate of the regression coefficients, we create a prediction of ($x\beta$) on the scale of estimation. This variable and its square are included as the only covariates in a second version of the model. If the model is linear, then the coefficient on the square of the prediction should be insignificantly different from zero. We report the proportion of the simulations where the t test for the second term is significant at the 5 percent level.

The third test uses the Pearson correlation between the raw-scale (y-scale) residual and x. If this statistic is significantly different from zero, then the model is providing a biased prediction of $E(y|x)$. Unlike Pregibon's Link Test, this test examines a propensity of the estimated impact of x on y (the slope) to be either too high or too low. We report the proportion of the simulations where the correlation is significantly different from zero.

B.1. Simulation Results.

Appendix B, Tables 1 and 2 report the results for the same set of data generating mechanisms and estimators examined in earlier part of the paper. Figures 1 and 2 show the pattern of the means across deciles that correspond to the Hosmer-Lemeshow tests on the raw scale.

In terms of predictions, log OLS, regular GGM, and the Weibull make biased predictions across all the deciles of X (Appendix B Figure 1). The biases are larger for the quadratic variance. These estimators also fail the goodness of fit. In contrast, the regular Gamma and the GGM-het make unbiased predictions and seem to provide a good fit to the data.

The results for the diagnostic tests were mixed. In principle, we should find no evidence of model misfit when the estimates should be consistent. Except for the OLS on log(y) (with homoscedastic retransformation) and GGM model when the errors are heteroscedastic, all of the estimators should be consistent with zero average residuals on the raw scale. None of the results were statistically significant from zero at usual levels. However, the Weibull and GGM

were off by appreciable amounts for the heavy tailed distribution – 14 and 20 percent of mean values. The consistent estimators should also have failed Pregibon's Linktest and the modified Hosmer-Lemeshow only about 5 percent of the time. Pregibon's test was well behaved, with model rejections in the 5 – 6 percent range for the consistent models, with only the Weibull having too many failures. The modified Hosmer-Lemeshow test indicated that there was a problem too often, especially for the heavy tailed and highly heteroscedastic data generating processes. This is also evident in the figures in this appendix, which indicate the degree of misfit across quantiles of the single covariate x. Pearson's correlation of the raw-scale prediction and residual often failed to have adequate coverage.

The inconsistent estimators had small average bias, but failed to pass the modified Hosmer-Lemeshow and Pearson correlation tests in a high proportion of the cases. Pregibon's Linktest failed to pick up the misspecification of the model in the heteroscedastic cases where there should have been a problem (OLS on ln(y) and GGM.) Thus, it appears that Pregibon's Linktest is an under powered test for assessing the type of bias that one would get in this case.

Thus the evidence indicates that some of the diagnostics are weak (Pregibon's Linktest) or may indicate failure too often (the modified Hosmer – Lemeshow in the face of heavy tails). Thus these diagnostics should be used with some caution.

B.2. Hospitalist Example

We also examined the fit for the hospitalist data. The test results for inpatient expenditures for the OLS on ln(y) model and for the gamma model with log link are mixed. There is no evidence of significant nonlinearity by the Hosmer-Lemeshow test or for Pregibon's Link Test for OLS. The gamma model fails Pregibon's Link Test ($p = 0.04$). Both gamma models fail the Pearson correlation test ($p < 0.0001$). It appears that the problem is the lack of fit for the casemix measures (the DRG relative weight and the Charlson Index).

The Weibull regression model fails all of the tests of fit (Table VII) and tends to over-predict the grand mean and the mean by deciles of disease specific experience (Appendix B, Figure 3).

The regular GGM produces results identical to log OLS model in terms of slope and goodness of fit test. However, the average residual from prediction is about 15 times lower than that of log OLS. The test of log normality fails to reject the log normal distribution. A heteroscedastic version of GGM is fitted by modeling $\ln(\sigma) = \alpha_0 + \alpha_1 LNCNT2 + \alpha_2 LND3CNT2$. This indicates that we assume that the heteroscedasticity is of the form: $\sigma^2 = K_1 (CNT2)^{K2} (D3CNT2)^{K3}$, where CNT2 is cumulative disease specific experience to date and D3CNT2 is specific experience-to-date. Though, model fit with GGM-het was not much different than in regular GGM, the slope of LNCNT2 and LND3CNT2 were comparable to the gamma regression with log link.

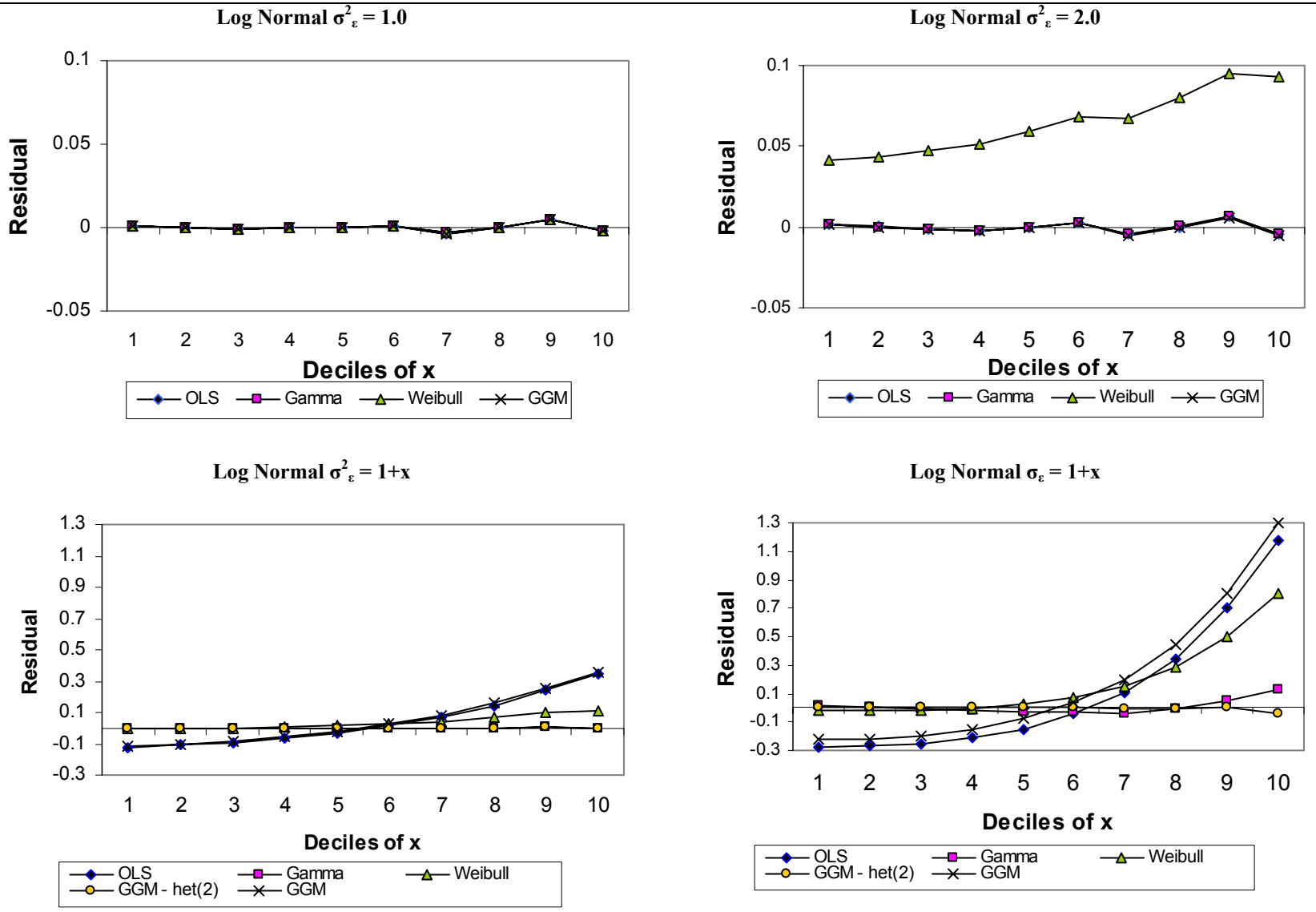

Figure 1: Mean Residual from different estimators across deciles of 'X' for Log Normal data with and without heteroscedasticity.

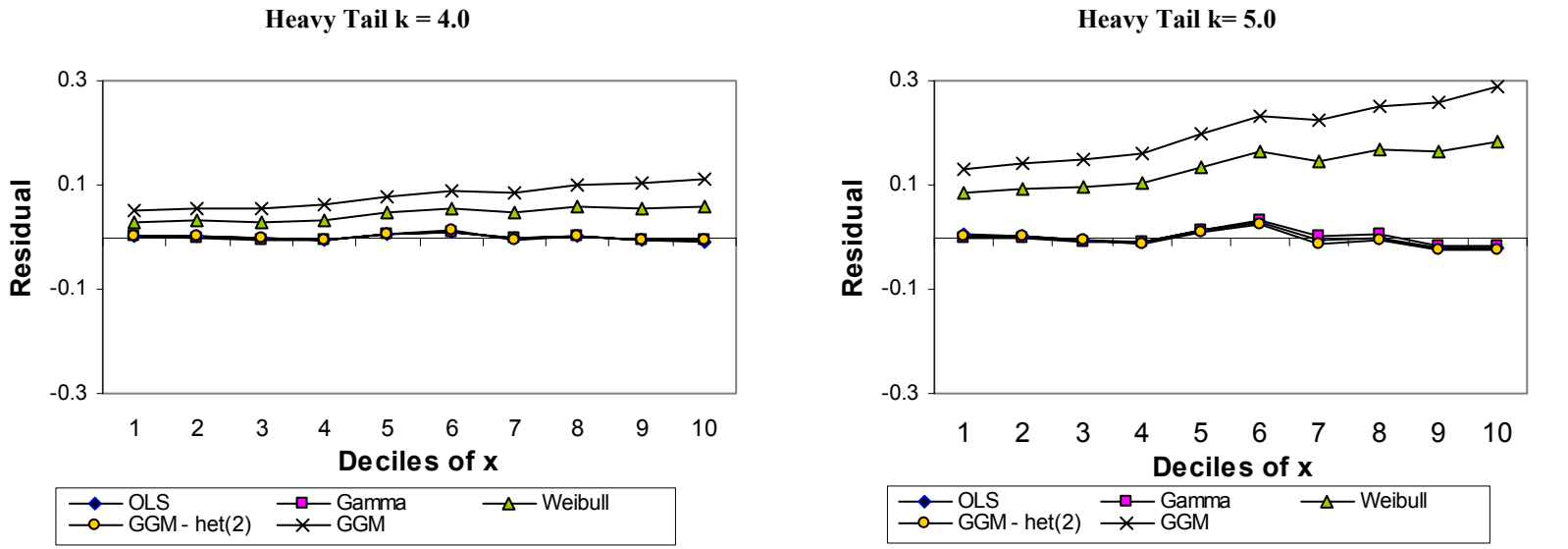

Figure 2: Mean Residual from different estimators across deciles of 'X' for data with heavy tails.

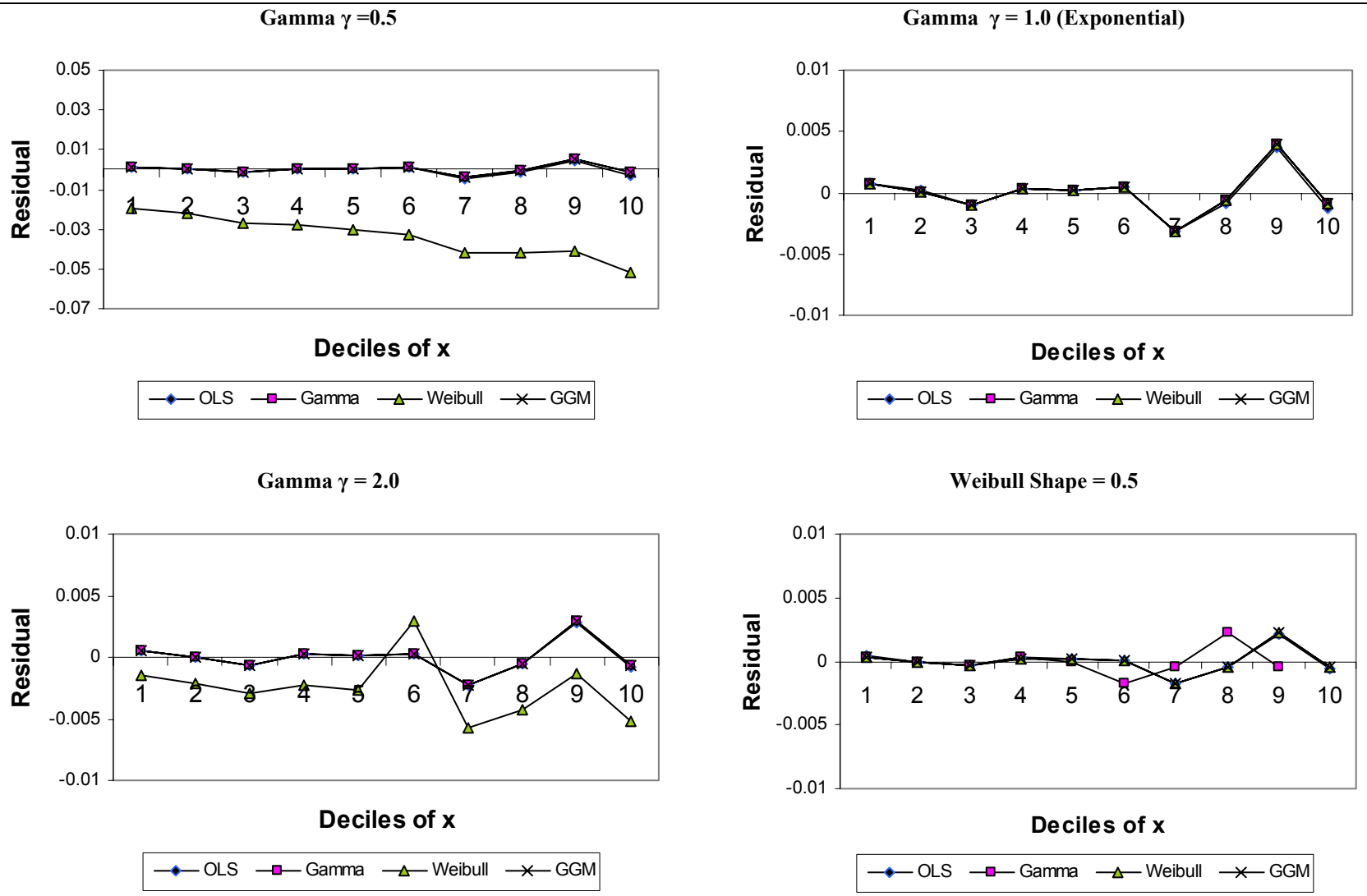

Figure 3: Mean Residual from different estimators across deciles of 'X' for Gamma, Exponential and Weibull data.

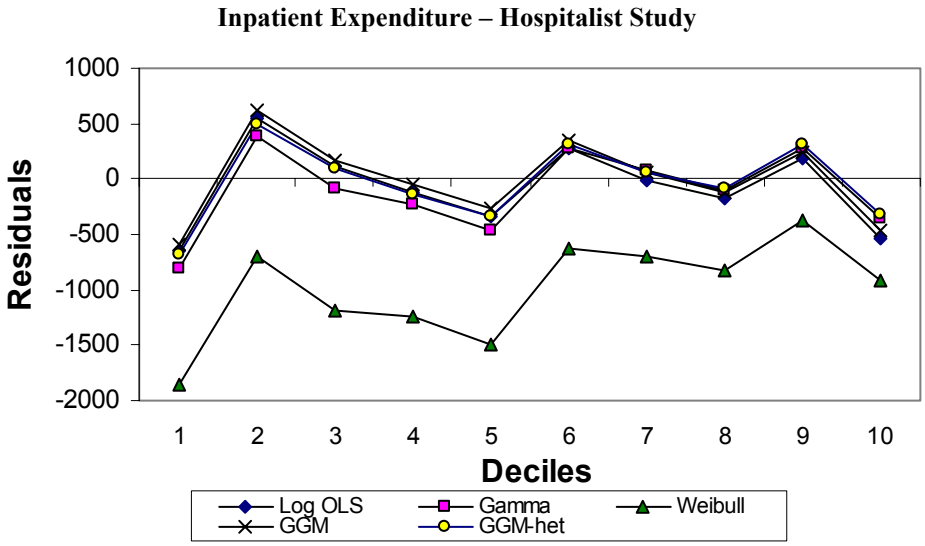

Figure 4: Mean Residual from different estimators across deciles of disease-specific experience for inpatient expenditures from the Hospitalist Study.

Appendix B Table 1: Alternative Estimator Results for Log normal, heteroscedastic and heavy tailed distributions.

Data	Estimator	Average Residual[1]	Prob. H.L.[2] Signif.[3]	Prob. Pregibon Signif.[3]	Prob. Pearson Signif.[3]
Log Normal	OLS for ln(y)	- 0.00002	0.0460	0.0580	0.0040
$\sigma_\varepsilon^2 = 1.0$	Gamma	0.00001	0.0400	0.0560	0.0000
	Weibull	0.0001	0.0400	0.0580	0.0000
	GGM	-0.00004	0.0460	0.0620	0.0040
	GGM – het(1)	-0.00007	0.0400	0.0600	0.0000
Log Normal	OLS for ln(y)	-0.0001	0.1180	0.0580	0.0120
$\sigma_\varepsilon^2 = 2.0$	Gamma	0.00001	0.1060	0.0600	0.0000
	Weibull	0.0647	0.0520	0.0580	0.0020
	GGM	-0.0004	0.1180	0.0560	0.0260
	GGM – het(1)	-0.0004	0.1100	0.0620	0.0000
Log Normal	OLS for ln(y)[4]	0.0395	1.0000	0.0640	1.0000
$\sigma_\varepsilon^2 = 1+x$	Gamma	0.00004	0.0700	0.0620	0.0000
	Weibull	0.0372	0.0460	0.0620	0.2100
	GGM	0.0494	1.0000	0.0620	1.0000
	GGM – het(1)	-0.0027	0.0760	0.0560	0.0240
	GGM – het(2)	-0.00034	0.0660	0.0560	0.0120
Log Normal	OLS for ln(y)[4]	0.1119	1.0000	0.0640	1.0000
$\sigma_\varepsilon^2 = (1+x)^2$	Gamma	0.0075	0.3000	0.1600	0.0000
	Weibull	0.1767	0.9320	0.2000	0.9980
	GGM	0.1908	1.0000	0.0620	1.0000
	GGM – het(1)	-0.0284	0.2860	0.0580	0.3420
	GGM – het(2)	-0.0032	0.1721	0.0656	0.1148
Heavy-tailed	OLS for ln(y)	-0.0004	0.4800	0.0620	0.0380
k = 4.0	Gamma	-0.00007	0.4320	0.0460	0.0000
	Weibull	0.0441	0.1480	0.0600	0.0000
	GGM	0.0787	0.0680	0.0620	0.0500
	GGM – het(1)	0.0786	0.0660	0.0620	0.0240
	GGM – het(2)	0.0002	0.5040	0.0580	0.0480
Heavy-tailed	OLS for ln(y)	-0.0015	0.8060	0.0620	0.0540
k = 5.0	Gamma	-0.0003	0.7560	0.0620	0.0000
	Weibull	0.1336	0.2080	0.0660	0.0020
	GGM	0.2031	0.1180	0.0620	0.0680
	GGM – het(1)	0.2030	0.1020	0.0620	0.0400
	GGM – het(2)	-0.0049	0.9020	0.0600	0.0800

NOTE: Based on 500 simulations, each with n=10,000. Numbers normalized so that E(y)=1.0.
[1]Residual on raw scale for y. [2]Hosmer-Lemeshow test. [3]At the 5 percent level.
[4] With homoscedastic smearing. GGM = Generalized Gamma; het(1) = hetero. model with $\ln(\sigma) = \alpha_0 + \alpha_1 x$; het(2) = true underlying hetero. model.

Appendix B Table II: Alternative Estimator Results for Gamma, Weibull and Gompertz distributions.

Data	Estimator	Average Residual[1]	Prob. H.L.[2] Signif.[3]	Prob. Pregibon Signif.[3]	Prob. Pearson Signif.[3]
Gamma	OLS on ln(y)	-0.0003	0.0440	0.0640	0.0960
$\gamma = 0.5$	Gamma	0.0000	0.0200	0.0540	0.0000
	Weibull	-0.0337	0.2260	0.0540	0.0000
	GGM	0.00002	0.0200	0.0520	0.0000
	GGM – het(1)	0.00002	0.0200	0.0520	0.0000
Gamma	OLS on ln(y)	-0.00008	0.0320	0.0640	0.0200
$\gamma = 1.0$	Gamma	0.00001	0.0180	0.0600	0.0000
(Exponential)	Weibull	0.00001	00180	0.0600	0.0000
	GGM	0.00001	0.0180	0.0300	0.0440
	GGM – het(1)	0.00001	0.0180	0.0600	0.0000
Gamma	OLS on ln(y)	-0.00002	0.0180	0.0620	0.0000
$\gamma = 2.0$	Gamma	0.00001	0.0180	0.0520	0.0000
	Weibull	-0.0031	0.0200	0.0520	0.0000
	GGM	0.0000	0.0180	0.0500	0.0000
	GGM – het(1)	0.0000	0.0180	0.0500	0.0000
Weibull	OLS on ln(y)	-0.00001	0.0160	0.0640	0.0000
$\sigma = 0.5$	Gamma	0.0000	0.0140	0.0520	0.0000
	Weibull	0.00001	0.0160	0.0600	0.0000
	GGM	0.00001	0.0140	0.0600	0.0000
	GGM – het(1)	0.0000	0.0180	0.0500	0.0000

NOTE: Based on 500 simulations, each with n =10,000. Numbers normalized so that E(y) = 1.0.
[1]Residual on raw scale for y. [2]Hosmer-Lemeshow test. [3]At the 5 percent level.
[4] With homoscedastic smearing. GGM = Generalized Gamma;
het(1) = hetero. model with $\ln(\sigma) = \alpha_0 + \alpha_1 x$;

Appendix B, Table III: Goodness of Fit on the Raw Scale of Inpatient Expenditures from Hospitalist Study.

Estimator	**Average Residual[1]**	**H.L. F Test (p)**	**Pregibon Test (p)**	**Pearson Correl. (p)**
OLS for ln(y)	-73.63	1.09 (0.36)	-1.62 (0.106)	-0.1387 (<0.0001)
Gamma Regression	-101.37	1.25 (0.26)	-2.04 (.041)	-.1313 (<0.0001)
Weibull Regression	-993.49	7.72 (<0.0001)	-0.07 (0.94)	-0.2212 (<0.0001)
Cox Regression	-138.17	1.38 (0.18)	2.90 (0.004)	.1702 (<0.0001)
Generalized Gamma	-5.29	0.94 (0.50)	-1.65 (0.10)	-.1304 (<0.0001)
GGM-het	-27.29	0.92 (0.52)	-1.65 (0.10)	-0.1489 (<0.0001)

Tests for identifying Distributions		Chi Sq Statistic	df	p-value
Std. Gamma	$\kappa = \sigma$	303.30	1	<0.0001
Log Normal	$\kappa = 0$	2.45	1	0.1178
Weibull	$\kappa = 1$	587.05	1	<0.0001
Exponential	$\kappa = \sigma = 1$	1273.80	2	<0.0001