TECHNICAL WORKING PAPER SERIES

ESTIMATING LOG MODELS: TO TRANSFORM OR NOT TO TRANSFORM?

Willard G. Manning
John Mullahy

Technical Working Paper 246
http://www.nber.org/papers/T0246

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 1999

This research was supported in part by a grant from the National Institute of Alcohol Abuse and Alcoholism (NIAAA) under Grants AA10392 and AA10393 and by Janssen Pharmaceutica, LP. The opinions expressed are those of the authors, and not those of NIAAA, the National Bureau of Economic Research, the University of Chicago, the University of Wisconsin, or Janssen Pharmaceutica, LP. We would like to thank Ashoke Bhattacharjya, Partha Deb, Tom DeLiere, Edward Norton, and Daniel Polsky for their comments on an earlier draft. An earlier version of this paper was presented at the Second World Conference of the International Health Economics Association, Rotterdam, the Netherlands, June 6-9, 1999. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Estimating Log Models: To Transform or Not to Transform?
Willard G. Manning and John Mullahy
NBER Technical Working Paper 246
November 1999
JEL No. C2, I1

ABSTRACT

Data on health care expenditures, length of stay, utilization of health services, consumption of unhealthy commodities, etc. are typically characterized by: (a) nonnegative outcomes; (b) nontrivial fractions of zero outcomes in the population (and sample); and (c) positively-skewed distributions of the nonzero realizations. Similar data structures are encountered in labor economics as well. This paper provides simulation-based evidence on the finite-sample behavior of two sets of estimators designed to look at the effect of a set of covariates x on the expected outcome, E(y|x), under a range of data problems encountered in every day practice: generalized linear models (GLM), a subset of which can simply be viewed as differentially weighted nonlinear least-squares estimators, and those derived from least-squares estimators for the ln(y). We consider the first- and second-order behavior of these candidate estimators under alternative assumptions on the data generating processes. Our results indicate that the choice of estimator for models of ln(E(x|y)) can have major implications for empirical results if the estimator is not designed to deal with the specific data generating mechanism. Garden-variety statistical problems - skewness, kurtosis, and heteroscedasticity - can lead to an appreciable bias for some estimators or appreciable losses in precision for others.

Willard G. Manning
Department of Health Studies
and Harris School
The University of Chicago
5841 So. Maryland Ave.
Chicago, IL 60637
w-manning@uchicago.edu

John Mullahy
Departments of Preventive Medicine
and Economics
University of Wisconsin - Madison
787 WARF, 610 Walnut St.
Madison, WI 53705
and NBER
jmullahy@facstaff.wisc.edu

I. Introduction

Health economists need little convincing that many of the outcomes with which they are concerned are awkward to analyze empirically; see Jones (1999) for an excellent overview. The circumstances that concern us in this analysis are those involving data like those typically encountered on health care expenditures, length of stay, utilization of health care services, consumption of unhealthy commodities, and others. Such data are typically characterized by: (a) nonnegative measurements of the outcomes; (b) a nontrivial fraction of zero outcomes in the population (and sample); and (c) a positively-skewed empirical distribution of the nonzero realizations. Econometric strategies for the analysis of such data have been discussed extensively (Duan, Manning, et al, 1983; Jones, 1999; Manning, 1998; Mullahy, 1998).

This paper provides some simulation-based evidence on the finite-sample behavior of two sets of estimators designed to look at the effect of a set of covariates x on the expected outcome, E(y), under a range of data problems encountered in every day practice. We work largely within the two classes of estimators: generalized linear models (GLM), and those derived from least-squares estimators for the ln(y), a subset of which can simply be viewed as differentially weighted nonlinear least-squares estimators. We consider the first- and second-order behavior of these candidate estimators under alternative assumptions on the data generating processes.

We are also investigating the performance of two variants of the traditional OLS model for the ln(y). Although technically, models for ln(y) get at the expectation of the log, rather than at the log of the expectation, they are interesting for two reasons. First, OLS for ln(y) is by far the most prevalently used (and most prevalently mis-used) model for looking at such data. Second, it is possible to go from the E(ln(y)) to the ln(E(y)) by retransformation (Duan, 1983; Manning, 1998). While these two classes of models – the GLM and OLS-based – overlap for some data structures, neither is a proper subset of the other.

The results indicate that there are important tradeoffs in terms of precision and bias. The OLS based methods can be biased in the face of heteroscedasticity if not appropriately retransformed. The GLM models can yield very imprecise estimates if the error term is heavy tailed on the log scale. Even if the estimators considered are consistent, there can be major gains in precision from selecting a more appropriate estimator. We develop a method for determining

which estimation method to choose for any application, using tests that are relatively easy to implement. We illustrate the approach with data on doctor visits from the National Health Interview Survey.

The plan for the paper is as follows. Section II describes the general modeling approaches that we consider. Section III presents our simulation framework. Section IV summarizes the results of the simulations as well as of an empirical example that focuses on the outcome of annual physician visits. Section V concludes.

II. Modeling Framework

In what follows, we adopt the perspective that the purpose of the analysis is to say something about how the expected outcome, E(y), responds to shifts in a set of covariates x.[1,2] Whether E(y) will always be the most interesting feature of the joint distribution $\phi(y,x)$ to analyze is, of course, a situation-specific issue, but the dominance of conditional-mean modeling in health econometrics renders what we suggest below of central practical importance. While many aspects of the following discussion apply for the more general case of nonnegative y, the discussion here is confined to the strictly-positive-y case to streamline the analysis. As a result, issues related to truncation/censoring or the "zeros" aspects of data (or "part one of a two-part model") are ignored here, but will be addressed in future work.

Our modeling framework includes two classes of estimators: Generalized Linear Models (GLM) with a logarithmic link function, and least squares for models with logged dependent variables. These specific GLM models estimate the ln(E(y|x)) directly, while the least squares estimates E(ln(y|x)), which can at least in principle be converted to E(y|x) by a suitable retransformation. As we have stressed elsewhere (Manning, 1998; and Mullahy, 1998), it is essential to distinguish these related but distinct models.

---

[1] We use the E(.) and var(.) notation as shorthand for E(.|x) and var(.|x) throughout. Essentially all moments considered here are conditional-on-x moments.

[2] This rules out situations where the analyst is interested in some latent variable construct.

A. GLM Modeling

In the version of the generalized linear model (GLM) framework (McCullagh and Nelder, 1989) used here, the central structure of the model is an exponential conditional mean (ECM) or log-link relationship:

$$\ln(E(y)) = x\beta \tag{1a}$$

or

$$E(y) = \exp(x\beta) = \mu(x;\beta). \tag{1b}$$

In GLM modeling, one specifies a mean and variance function for the observed raw scale variable y, conditional on x. Three stochastic families are studied here, the key attributes of which involve their respective conditional mean-variance relationships. These relationships can be described using the general structure

$$\text{Var}\,(y) = \sigma^2\, v(x). \tag{2}$$

Rather than write Var (y(x)), we will use v(x) directly. The first case is the homoscedastic or "classical" nonlinear regression model with v(x)=1; that is, the variance of y (conditional on x) is unrelated to x. The second case has a Poisson-like structure with $v(x)=\kappa_1 \mu(x)$, where $\kappa_1 > 0$; that is the variance is proportional to the mean, which is itself a function of x. The third has a gamma structure with $v(x)=\kappa_2 (\mu(x)^2)$, where $\kappa_2 > 0$; that is, the standard deviation is proportional to the mean. Within this class of power-proportional variance functions, it is useful to think more generally of the variance function v(x) being:

$$v(x) = \kappa(\mu(x\beta))^{\lambda} \tag{3}$$

where λ must be finite and non-negative. In the case $\lambda = 0$, we get the usual nonlinear least-squares estimator. In the case $\lambda = 1$, we get the Poisson-like class. In the case $\lambda = 2$, we get the gamma, the homoscedastic log normal, the Weibull, and the chi square, with the suitable

specification of a distribution.[3] In the case $\lambda = 3$, we get the inverse Gaussian (or Wald) distribution. Throughout this paper, we are assuming a log link for the expectation of y given x, $\mu = \exp(x\beta)$.

Estimation of the conditional mean parameters β given such structural assumptions proceeds using what economists think of as GMM estimation but what is more generally spoken of by statisticians as GLM modeling using quasi-likelihoods or generalized estimating equations (GEE). Regardless of how interpreted, the key features of such estimation approaches are the moment or quasi-score equations

$$0 = \sum_{i=1}^{N} \frac{\partial \mu(x_i;\beta)}{\partial \beta} \times v(x_i)^{-1} \times (y_i - \mu(x_i;\beta)), \tag{4}$$

whose solutions $\hat{\beta}$ are the estimators of interest. The v(x) are assumed to be functions of $\mu = \exp(x\beta)$, not of individual covariates in x more generally.

B. OLS-Based Models

By far the more prevalent modeling approach is to use ordinary least-squares or a variant with ln(y) as the dependent variable. In this case, the assumed regression model is

---

[3] Note that the "gamma-class" (λ=2) models are in some respects a natural "baseline" specification. That is, if the model is taken to be

$$y = \exp(x\beta) \times u$$

and if u is taken to be homoskedastic, then it is indeed natural to suggest that Var[y|x] is proportional to E[y|x]-squared. Thus, just as the homoskedastic linear model

$$y = x\beta + u$$

generates a "natural" constant-variance perspective in the linear context, the exponential mean model generates a "natural" "gamma-class-variance" perspective in the log-linear context.

$$\ln(y) = x\delta + \varepsilon \tag{5}$$

where it is assumed that $E(x'\varepsilon) = 0$. The error term need not be i.i.d. If the error term is normally distributed $N(0, \sigma^2)$, then $E(y) = \exp(x\delta + 0.5\sigma^2)$. If ε is not normally distributed, but it is homoscedastic, then $E(y) = s \times \exp(x\delta)$, where $s = E(\exp(\varepsilon))$.[4] In either case, the expectation of y is proportional to the exponential of the log scale prediction from the OLS or LS-based estimator.

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

$$\ln(E(y)) = x\delta + \ln(f(x)) \tag{6}$$

and in the log normal case,

$$\ln(E(y)) = x\delta + 0.5\sigma^2(x) \tag{7}$$

where the last variance term is the error variance on the log scale.

In general, the presence of heteroscedasticity on the log scale for an LS-based models implies that the exponentiated log scale prediction $[s(\exp(x\delta))]$ provides a biased estimate of the E(y|x), and is biased in a way that depends on x; the s here is the (homoscedastic) smearing

factor. This bias can be eliminated by including an estimate of the variance function, v(x), if the error is log normal, or more generally, of $E(\exp(\varepsilon)|x)$.

III. Methods

To evaluate the performance of the two alternative classes of estimators for log models, we rely on a Monte Carlo simulation of how each estimator behaves under a range of data circumstances that are common in health economics and health services research studies. There are five data situations that we consider: (1) skewness in the dependent variable; (2) heavy-tailed distributions (even after use of log transformations to reduce skewness); (3) pdf's that are monotonically declining rather than bell-shaped; (4) data with nonlinear responses but additive errors;and, (5) log error terms that are heteroscedastic. We do not deal with either truncation or censoring.

We also provide a set of tests for determining which estimator is appropriate for a given data set, using easy to implement tests. We illustrate the approach using data on doctor visits from the National Health Interview Survey.

A. Alternative Data Generating Structures

As we noted earlier, one of the major motivations for using a logarithmic transformation of the dependent variable is a concern over the severe skewness in health care utilization and expenditures. By transforming the dependent variable, the goal is to be able to use ordinary least squares estimators without having to worry about the sensitivity of the results to skewness.

Some applications have more skewed dependent variables than others. For example, the inpatient days are more skewed than the number of inpatient stays, among those with any hospitalizations. Inpatient expenditures tend to be more skewed (and kurtotic) than inpatient days.

---

[4] Duan (1983) shows that one can substitute the estimated residual for ε to get a consistent estimate of the smearing factor s.

To determine the effect of the level of skewness on the estimated outcome, we examine two classes of data generating mechanisms: (1) log normal distributions with increasing log scale error variances; and (2) gamma distributions with decreasing shape functions. In the case of the log normal, the raw scale mean, variance, skewness, and kurtosis are all increasing functions of the variance on the log scale. If the log scale error ε is normally distributed with mean 0 and variance v, then the raw scale skewness (S) for this data generating mechanism is:

$$S_{raw}=(w + 2)\,((w - 1)^{0.5}) \tag{8}$$

where w = exp (v). Using a N(0,v) deviate, we let the log scale variance range from 0.5 to 2.0 in steps of 0.5. Thus, the skewness of exp (ε) varied from 2.94 to 23.7, compared to zero for a normal deviate.

Specifically, we assume that the true model is:

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon \tag{9}$$

where x is uniform (0,1), ε is N(0,v) with variance v = 0.5, 1.0, 1.5, or 2.0, and E (x′ ε) = 0. β_1 equals 1.0. The value for the intercept β_0 is selected so that E (y) = 1.

Note that for this data generating mechanism, the expectation of y is:

$$E(y) = e^{(\beta_0 + \beta_1 x + 0.5v)} \tag{10}$$

The slope of E(y) with respect to x equals $\beta_1 \exp(\beta_0 + \beta_1 x + 0.5v)$.

Some studies deal with dependent measures and error terms that are heavier tailed (on the log scale) than even the log normal. For example, the residual for Edward Norton et al.'s study of (log) length of stay for Medicaid psychiatric inpatient care has a log scale kurtosis of 3.5, compared to a value of 3 for a normal (or in that case log normal). David Meltzer's hospitalist study has a kurtosis of 3 for log length of stay, but over 6 for log costs.

We generate 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. p percent

of the population have a log scale variance of 1, and (1-p) have a higher variance. In the first case, the higher variance is 3.3, yielding a log scale error term with a kurtosis of 4.0. In the second case, the higher variance is 4.6, giving a log scale error term with a kurtosis of 5.0. For the normal distribution, the kurtosis is 3.0.

We also consider models based on the gamma distribution. The gamma has a pdf that can be either monotonically declining throughout the range or bell-shaped, but skewed right. The pdf for the gamma variable y is:

$$f(y) = (y/b)^{(c-1)} \frac{[\exp(-y/b)]}{b\Gamma(c)} \qquad (11)$$

where b is the scale parameter and c is the shape parameter; some parameterizations use a = 1/b. The scale parameter b equals exp ($\beta_0 + \beta_1$ x), where $\beta_1 = 1$, and β_0 is selected so that the E(y) = 1. The shape parameter c is 0.5, 1.0, or 4.0. The first and second values of the shape parameter yield monotonically declining pdf's, conditional on x, while the last is bell shaped but skewed right. The first is a chi square with one degree of freedom if b equals 1. The second is an exponential variate. As the shape c increases to infinity, the distribution approaches a normal. Thus the skewness S on the raw scale is a declining function of c, $S = 2\ c^{-0.5}$ if we ignore the covariates.

The next class of data generating mechanisms is the one with an additive error term that corresponds to the nonlinear least-squares model:

$$y = e^{x\beta} + \varepsilon \qquad (12)$$

where ε is a normal deviate with mean zero and standard deviation 0.3. In principle, the NLS estimator should be ideal for this data generating mechanism.

Finally, it is not uncommon to encounter heteroscedasticity in the error term of a linear specification for ln(y). In this case, estimates based on OLS on the log scale can provide a biased assessment of the impact of the covariate x on E(y); see Manning (1998) for a discussion. In this case, the constant variance v in Equation 3, is replaced by some log scale variance function v(x). The expectation of y on the raw scale becomes:

$$E(y) = e^{(\beta_0 + \beta_1 x + 0.5v(x))} \tag{13}$$

if the underlying error term ε is N(0, v(x)). The slope of the expectation of y is now:

$$\frac{\partial E(y)}{\partial x} = y\ (\beta_1 + 0.5\frac{\partial v(x)}{\partial x}) \tag{14}$$

To construct the heteroscedastic log normal data, the error term ε is the product of a N(0,1) variable and either (1 + x) or its square root. The latter has error variance that is linear in x (v = (1+x)), while the former is quadratic in x ($v = 1 + 2x + x^2$). Again, $\beta_1 = 1$, and β_0 is selected so that E(y) = 1.

Table 1 summarizes the data generating mechanisms that we consider.

B. Alternative Estimators

We employ five different estimators for each of these data generating processes. The first two are from the least squares class. The first relies on ordinary least-squares (OLS) regression of ln(y) on x and an intercept, and uses a homoscedastic smearing factor to retransform the results to obtain E(y|x) . The second also relies on ordinary least-squares regression of ln(y) on x and an intercept, but uses a heteroscedastic retransformation; see below. The other three models are variants of generalized linear models (GLM) for y with a log-link function (McCullagh and Nelder, 1989). In the first GLM case, the error term is additive on the raw scale and has a variance that does not depend on E(y) or x. This is basically the nonlinear least-squares estimator proposed by Mullahy (1998). The second GLM estimator assumes that the raw scale variance is proportional to the E(y), which is a Poisson-like assumption without the discrete nature of the dependent measure. The third GLM approach assumes has a raw scale standard deviation is proportional to E(y), which is a gamma-like assumption similar to Blough et al. (1999). In all three GLM models,

$$E(y) = e^{(\beta_0 + \beta_1 x)} \tag{15}$$

Because the OLS estimates are for the E(ln(y)), we *retransform* the log scale estimates to obtain raw scale estimates of E(y). For all of the OLS-based estimators (except for the heteroscedastic retransformation cases), we use Duan's (1983) smearing estimator to obtain an estimate of E(y). The smearing estimator is the average of the exponentiated residuals from the ln(y) regression.[5] If the log scale errors are not heteroscedastic in some function of x or of E(y), then the smearing estimate provides a consistent estimate of E[exp(ε)]. If the error ε is truly normal, then the smearing estimate is less precise than using exp(0.5v), where v is a consistent estimate of the log scale residual variance.

We also generate predictions based on heteroscedastic retransformation

$$v = \mathrm{E}(\varepsilon^2) = \delta_0 + \delta_1 x + \delta_2 x^2 \qquad (16)$$

if the variance is (1+x), then we omitted the x squared term from a regression of squared residuals on x and x squared. For all of the GLM generated data, we assume that the variance function is linear in x.

All of the equations are estimated in STATA 5.0, using either the standard regression command ("reg") or the appropriate GLM command:

glm y x, family(xxx) link(log)

where xxx is either Gaussian, Poisson, or gamma.[6]

C. Design and Evaluation.

Each model is evaluated on 1000 random samples, with each having a sample size of 10,000. Except for the two heteroscedastic cases, all models are evaluated in each replicate of a

---

[5] We did not use the normal theory retransformation from equation 7 because it would be inconsistent for several of our data generating mechanisms. Except for the heteroscedastic log normal cases, the smearing estimate should provide a consistent retransformation.

[6] In practice, we recommend the use of Stata's "xtgee" or "rglm" command instead of "glm," because the first two accommodate robust covariance matrix estimation while the last does not.

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 percent confidence interval of the estimate 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 and 95 percent confidence interval indicate the precision of the estimate.

(2) The mean squared error (MSE) of the model on the original estimation sample. The MSE indicates how well the estimate minimized the original residual error on the raw scale.

(3) The absolute prediction error (APE) of the estimate of β_1, where the APE is the absolute value of the estimate of β_1 minus its true value.[7] A more precise estimator should be closer to the true value.

If a model has low MSE and high APE, then there is strong evidence that that estimator has overfitted the estimation sample. The 95 percent confidence intervals are based on the 0.025 and 0.975th percentiles of the estimates, rather than using the normal theory estimate. Not all of the estimated values of the β's are normally distributed, or whose distribution is well approximated by a normal. Estimators are compared on APE and MSE by comparing the number of times that estimator A had a lower APE (or lower MSE) than estimator B. With n replicates with random draws, the proportion p where A is lower than B should be 0.5 under the null that the two estimators are equally good, and the variance of p is p(1-p)/ (n).

D. Diagnostics for Variance Functions (Park Tests)

The results below will provide a compelling demonstration of the importance in terms of

---

[7] For this study, we know the true values of the parameters. But in most applications, the analyst does not know the true population parameters to use in constructing the APE. Another alternative is to use a split-sample, cross-validation approach. If overfitting occurs, the estimator will perform better on the estimation sample than on the validation sample. See Duan et al. (1983) for an example comparing alternative estimators for health expenditures.

precision of specifying a (conditional) variance function that captures the true conditional variance in the data. In this section, we propose a simple strategy for selecting such a specification, one that should be of considerable use in practice.

As above, we focus on the GLM class of variance functions where:

$$\text{var}(y) = \alpha[E(y)]^{\lambda}. \tag{17}$$

because this specification captures most of the alternative estimators that we are interested in. In a generalized method-of-moments environment, this variance function specification would imply a set of moment conditions proportional to

$$m(y_i,x_i;\beta,\alpha,\lambda) = [\ (y_i - \exp(x_i\beta))^2 - \alpha\exp(\lambda x_i\beta)\ ] \tag{18}$$

such that E[m(.)]=0 under the assumption of correct specification of the conditional mean and conditional variance (e.g. Wooldridge, 1991).

This moment structure (with a consistent initial estimate of β) is similar to one of the early tests for heteroscedasticity. In the Park test (Park, 1966), the log of the estimated residual squared (on the scale of the analysis) is regressed on some factor z thought to cause heteroscedasticity in the error on the scale of the analysis. Here, we propose to use the residuals and predictions on the raw (untransformed) scale for y to estimate and test a very specific form of heteroscedasticity – one where the raw scale variance is a power function of the raw scale mean function. The OLS version of equation 17 is:

$$\ln((y_i - \hat{y}_i)^2) = \lambda_o + \lambda_1 \ln(\hat{y}_i) + \nu_i \tag{19}$$

where $\hat{y}_i = \exp(x_i\beta)$ in the GLM specifications, and $\exp(x_i\beta + 0.5\sigma^2(x))$ in the log normal specifications. The coefficient λ_1 on the log of the raw scale prediction will tell us which GLM model to employ if the GLM option is chosen.

While the purpose of the Park's original approach was to *test* for heteroscedasticity for a specific variable, we choose instead to exploit and interpret this approach as a guide to

specifying the λ parameter for purposes of weighted NLS or GLM estimation. Specifically, to the extent that the Park test estimate of λ captures the true variance function, we can be build a downstream GLM regression strategy for the choice of particular GLM models (NLS, Poisson, Gamma, etc.) whose variance (inverse weighting) function is specified to be $[\exp(x_i\hat{\beta})]^{\hat{\lambda}}$.

One concern with this approach is that we are focusing on the raw scale behavior of conditional means and variances in applications where skewness in the dependent measure y often leads to log transformation to obtain more robust results. Under these circumstances, how informative are these particular Park tests? To assess the utility of such a strategy, we return to the simulation designs described above and estimate the λ parameter for a subset of the data structures where y is skewed to the right: log normal, with log scale variance = 1; gamma, with shape=1; the 90/10 mixture of log normals with the kurtosis of 5 for the log error term ε; and heteroscedastic log normal, with log scale standard deviation = 1+x. Note that in the first two data generating specifications, the conditional variance is proportional to the square of the conditional mean (λ=2). In the third specification (the heavy tailed distribution from a mixture of log normals), the proportionality assumption is valid but it operates across different variance structures in the data. In the last data specification (heteroscedastic log normal), the proportionality specification is no longer strictly appropriate.

IV. Results: Simulations and Empirical Example

Table 2 provides some 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.

A. Skewness.

Given that the severe skewness in health utilization is often a major rationale for using a log approach, we begin with skewness. The skewness in y on the raw scale increases in the variance v for the log normal models. Table 3 provides the results on the consistency and precision in the estimates β_1, the slope of ln(E(y)) with respect to x, for each of the alternative estimators for the log normal data generating processes. In the absence of heteroscedasticity in x in the error ε, the

OLS model with homoscedastic retransformation,[8] the NLS, Poisson, and Gamma models all produce consistent estimates of the slope β_1.

Thus, if consistency were the only concern, and if there is no evidence of heteroscedasticity, then each of the models considered here is admissible.

However, if there is also a concern about precision, then the most precise estimates can be obtained by OLS, with the gamma, Poisson, and NLS versions of the GLM model trailing in that order from lower to higher variance. The differences in precision among the estimators increase as the log scale variance increases. At a variance of 0.5 on the log scale, the gamma standard error is roughly 13 percent larger, and it would take a sample size 28 [$0.28 = (1.133^2 - 1)$] percent larger to give the same precision as OLS with homoscedastic retransformation. At a variance of 2.0 on the log scale, the gamma standard error is roughly 74 percent larger, and it would take a sample size three times as large to give the same precision as OLS with homoscedastic retransformation. The NLS would require a sample almost four times as large as the OLS sample to have the same level of precision.

Thus, the efficiency losses (relative to the OLS-based estimator) from using GLM methods can be substantial and increasing in the variance on the log scale.

B. Heavy Tailed Data.

The presence of a heavy-tailed error distribution on the log scale does not cause consistency problems for these estimators, but it does generate much more imprecise estimates for the three GLM models; see Table 4. In the absence of heavy tails, the standard errors for the gamma estimates of the slope are 13 percent larger than for the OLS estimate. For the mixture of normals case, the standard errors are about 3.5 times larger for the gamma model and 4.6 times larger for the NLS estimator if the kurtosis is 4. They are over seven times larger for the gamma and over 130 times larger for the NLS if the kurtosis is 5.3.[9]

---

[8] We used Duan's (1983) smearing estimator.

[9] The poor performance of the NLS in terms of the standard error of the estimate of β_1 is heavily influenced by the estimate from one random sample. However, if we were to use a more robust estimate of dispersion, the inter-quartile range, we would still find the NLS to be the least precise estimator. The difference among the estimators would be less dramatic, but qualitatively similar.

Thus, the efficiency losses of GLM models (relative to the OLS-based estimator) are substantial and increasing in kurtosis of the log scale error.

C. Alternative Shapes to PDFs

To test the sensitivity of the results to differences in the shape parameter of the pdf, we use alternative gamma models, with shapes of 0.5, 1.0, and 4.0. These correspond to two monotonically declining, and one (skewed) bell-shaped pdf. As Table 5 indicates, all of the estimators yield consistent estimates of β_1. Not surprisingly, the gamma regression models yield the most precise estimates and OLS on ln(y) yields the least precise estimates. The Poisson-like GLM and NLS estimators are in between, but closer to the precision available from the gamma regression model than to that from the OLS-based model. The size of the discrepancy in precision is greatest for c = 0.5, and the least for a shape c = 4.0; the former has a monotonically declining pdf (conditional on x), while the latter has a skewed bell shape. It would take a sample size 2.5 times as large for OLS to generate the same precision as the gamma model if the shape c = 0.5, but only 14 percent larger if the shape c = 4.0.

Thus, the efficiency losses (relative to the gamma-based estimator) from using OLS based estimators can be substantial, but decreasing in the shape parameter c. The losses are greater if the pdf is monotonically declining, than if it is a skewed bell shape.

D. NLS-like Data Generating Mechanisms

The GLM models provide consistent estimates of β_1 when the data generating mechanism has an additive error ε on the raw scale. The homoscedastic retransformation of Log OLS model provides a significantly biased estimate of the true value, but one that is not appreciably biased – the bias is only on the order of four percent. The NLS estimate is the most precise of the estimates of β_1, while the Log OLS estimates are the least precise. The gain from using the NLS estimator in this case is roughly equivalent to an increase of three-quarters in the sample size; see Table 7 and Appendix Table 1.

E. Heteroscedasticity

As the discussion indicated, heteroscedasticity that depends on x can lead to biased estimates of the impact of x on the E(y) if OLS is used on ln(y) without an appropriate heteroscedastic

retransformation. Table 6 indicates that GLM models capture consistently the effect of x on ln(E(y)) when the error variance is linear in x, with their estimated values of β_1 averaging 1.5, the true value. The OLS model with homoscedastic retransformation provides an estimate that is significantly less than the true value. In essence, it captures only the "deterministic" part β_1 on the log scale, not the full effect: $\beta_1 + 0.5\ \partial v(x)/\partial x$.

However, by estimating v(x) from the OLS residuals on the log scale, the heteroscedastic retransformation of the OLS ln(y) model does provide a consistent estimate of the full effect of x on ln(E(y)). Of the consistent estimators, the heteroscedastic retransformation version is the most precise, followed by the gamma, the Poisson, and NLS models, in that order. The gamma model would require a sample 47 percent larger to give the same precision as the heteroscedastic retransformation version of OLS, and the NLS would require a sample 250 percent larger.

When the error variance on the log scale is quadratic in x, the story is more complicated. Unless a quadratic model is estimated for the GLM alternatives or in the variance function for the heteroscedastic version of OLS, then the estimates of $\partial \ln(E(y))/\partial x$ will be biased. If the square of x is added to the list of regressors,[10] then the GLM and the heteroscedastic retransformation version of OLS are all consistent. However consistent the GLM methods are, they do not provide a very powerful indication of the nonlinearity caused by this form of heteroscedasticity. The 95 percent confidence interval for the quadratic term for the NLS is [-1.99, +3.58], for the Poisson [-0.83, +2.12], and for the gamma [-0.41, +1.44] when the true value is 0.5. Only the OLS with heteroscedastic retransformation is able to pick up a result that is significantly different from zero; the 95 percent confidence interval is [0.002, 0.97].[11] As in

---

[10] In the case of the OLS based model, the square of x is added as a regressor in the variance function in equation 16, not to equation 9.

[11] The absence of a significant quadratic effect in the GLM is not due to lack of precision for quadratic terms in general for GLM models, but lack of precision when they are not the true model. For example, we also examined a gamma model with ln(E(y)) a quadratic function in x, shape = 1, and the same coefficients for the linear and quadratic effects as implied by the heteroscedastic model above. All three of the GLM models' coefficients for the quadratic terms have p values < 0.01, and are notably more precise than a quadratic OLS model for ln(y). The gamma regression model is the most precise of the alternatives under these specific circumstances.

the other heteroscedastic case, the homoscedastic retransformation version is appreciably biased, because it omits the term $+0.5\ \partial v(x)/\partial x$.

Thus, if consistency is the concern, the usual OLS-based model for ln(y) is inconsistent unless transformed by an appropriate heteroscedastic factor. All of the other estimators considered are consistent.

To the extent that precision is a concern, the heteroscedastic retransformation of the OLS-based results is the most precise alternative considered here.

For each of the data generating mechanisms that we have examined, we have estimated both heteroscedastic and homoscedastic retransformation results for the OLS-based estimators. Except for the cases that were truly heteroscedastic, the heteroscedastic version is usually less precise than the homoscedastic version. Except for the cases that were truly heteroscedastic, the both versions are consistent.

As each of these alternatives has suggested, there are substantial gains from selecting the best estimator for a given data situation. Different data generating mechanisms lead to different choices of estimators. Table 7 and Appendix Table 1 show that the precision gains from selecting a more appropriate model can be quite substantial. Within the class of GLM models, the choice of an inappropriate variance function can lead to a substantial loss in precision.

F. Overfitting

One of the concerns that has motivated the use of log models instead of OLS on raw scale dependent variables has been the fear that OLS would overfit the extreme cases. That is, the estimate of the β 's would be pulled around by extreme cases and not reflect well the true values. GLM models, especially the NLS, could have a similar problem, because they do not deal necessarily deal well with the skewness in the data unless the variance function is appropriately specified.

To address this issue, we examine both the mean-squared error (MSE) on the raw scale for the estimation sample and how close the estimated slope is to the true value, as measured by the absolute prediction error (APE) for β_1.[12] If over-fitting occurs, we would expect the MSE to be low, while the APE for that estimator to be high. For each of the estimation models, we

[12] APE = absolute (estimated β_1 - true β_1) averaged over replications.

compare alternative estimators in terms of which model had lower MSE's or lower APE's; see Appendix Table 2. For the within-sample measure of MSE, NLS generally has smaller MSE's than any of the other estimators, followed by the Poisson, gamma, and retransformed OLS models, in that order. This pattern holds across a number of different kinds of data problems, except for the NLS-like data generating mechanism. In contrast, the APE results suggest that the retranformed OLS model is closer to true, followed by the gamma, Poisson, and NLS in that order. Given the biased estimate for the homoscedastic retransformation method for OLS, when the error is heteroscedastic, this model is the worst behaved of all if there is heteroscedasticity, but the best of all (on APE grounds) if there is no heteroscedasticity.

In any event, the within sample estimator of fit, the mean squared error, is quite sensitive to skewness and other data problems. It tends to pick estimators that have higher true variances for estimates of β_1 than the within-sample estimate of mean squared error would indicate. The NLS and Poisson models are especially prone to this kind of overfitting in the face of skewness in the raw-scale version of y or of kurtosis in the log-scale error

G. Diagnostics for Variance Functions (Park Tests)

These results provide a compelling demonstration of the importance in terms of precision of specifying a (conditional) variance function that captures the true conditional variance in the data. In this section, we use the simulation approach to evaluate a simple strategy for selecting such a specification, one that is likely to be of considerable use in practice. Using the estimates from each model, we can construct raw scale residuals for each estimator. Then one can either use a non-linear least squares estimator for this residual squared versus a power function of the predicted (raw scale) value for the dependent measure, or one can regress by OLS the log of the raw scale residuals squared on the log of the raw scale prediction.

Table 8 provides a summary of the Park test simulations. We focus on the performance of the Park test OLS slope estimator for the different baseline estimators (linear least-squares on the log scale, nonlinear least-squares with a log link, Poisson with a log link, and gamma with a log link). For the first two data generating mechanisms, the performance of the Park test estimate is quite good for all the estimators. Despite the skewness in the dependent variable y, the estimates of λ centered quite tightly around the true value λ=2. Further, for these two data

generating mechanisms, there is no appreciable difference in precision across the estimators. In the heavy-tailed distribution specification, the replicate means and medians of the OLS estimator center on λ=2, whereas the cross-replicate performance of the nonlinear GLM estimators (NLS, Poisson, and Gamma) shows significant divergence between the mean and median of the estimates of λ. Specifically, although the median of the point estimates centers on λ=2, the mean estimate is attenuated due to the mixing, presumably a "Jensen's inequality type" consequence of mixing nonlinear functions. In any event, the Park test is not as informative about which value of λ to choose for the GLM models as it was for the log normal and gamma data generating mechanisms. Finally, for the heteroscedastic log normal case, we find that the simple homoscedastic OLS strategy misses the fact that the data are no longer structured such that λ=2, whereas the other estimators are not so fooled. The heteroscedastic version of the log based model is as well-behaved and as precise as the GLM alternatives.

If the conditional variance structure is proportional to some integer power of the conditional mean, then there is likely to be a substantial payoff from a strategy: (1) using one of the GLM estimators to obtain a baseline estimate of β, (2) conducting the Park test to obtain an estimate of λ, and (3) utilizing this estimate of λ to formulate a second-stage GLM weighting function. However, analysts should be alert to the fact that the second-stage GLM estimates may be based on estimated conditional variance functions that may not necessarily converge in the limit to the true conditional variance functions. As a result, they should continue to use robust (Huber/White-type or bootstrapped) estimates of the corresponding variance-covariance matrix for the estimate of β.

H. An Empirical Example

To illustrate the empirical importance of the above issues in the context of real data, we estimate a set of GLM models using the 1992 National Health Interview Survey (NHIS) data that were the basis of the study by Mullahy, 1998. These data comprise 27,598 observations on adults who had at least one doctor visit during the twelve months prior to the survey, this being the outcome we consider here. The summary statistics for this outcome measure are mean=6.42, median=3, variance=204.7, skewness=9.79, and kurtosis=158.6.

The model specification used here is identical to that used in the earlier study, including as covariates age in years (AGE), gender (MALE), years of completed schooling (EDUC), race

(WHITE), marital status (MARRIED), and health status (XCELLENT, VERYGOOD, GOOD). The models are estimated using λ values of 0 (NLLS class), 1 (Poisson class), 2 (gamma class), and 3 (inverse Gaussian class), as well as the "optimal" value derived using the nonlinear "Park test" procedure described above (which turns out to be 1.917887 for these data).

The results are summarized in table 9. It is apparent in the table that the precision of the point estimates in this case is best in the "optimal" case. The traditional GLM cases that bracket this case (λ=1,2,3) are not terribly inferior, while the λ=0 case is generally much less precise than the others. No less importantly, note too that the magnitudes of the point estimates in some cases vary dramatically across the values of λ (with MALE and EDUC being perhaps most striking in this respect).

From this example, we are left to conclude that considerations of the variance structure may have considerable implications for analysts' inferences in applied research.

V. Conclusions

Our results indicate that the choice of estimator for examining the ln(E(y)) can have major implications for the empirical results if the estimator is not designed to deal with the specific data generating mechanism. Garden-variety statistical problems – skewness, kurtosis, and heteroscedasticity – can lead to an appreciable bias for some estimators (e.g., simple OLS for ln(y)) or appreciable losses in precision for others (e.g., GLM).

The standard use of ordinary least-squares with a logged dependent variable reminds us of Longfellow's nursery rhyme. "When she was good, she was very, very good. But when she was bad, she was horrid!" OLS with homoscedastic retransformation seems to be resilient to various data problems, except for one. It deals much better with heavy-tailed distributions (heavy-tailed on the log scale) than any of the GLM alternatives that we have considered. Unfortunately, when the log scale error term ε is heteroscedastic, the OLS (with homoscedastic retransformation) estimates can be significantly biased. Moreover, when the pdf is not bell-shaped or a skewed bell-shape, then the OLS-based models are notably less precise than some of the GLM alternatives, but remain consistent.

The bias in the homoscedastic version of OLS can be corrected if one estimates the variance function v(x) for the log scale error, and then uses that to obtain a retransformed

prediction of ln (E(y)); see Manning (1998). Although consistent, this approach is much more cumbersome because it requires more investment in the "finer" details of econometric modeling than many analysts have been willing to invest. The heteroscedastic retransformation can be done easily if there is heteroscedasticity across mutually exclusive groups (e.g., health insurance plans), but is difficult for heteroscedasticity across multiple factors or continuous measures.

The GLM models, such as the NLS, Poisson-like, and Gamma models, provide the alternative of directly estimating what most economists are really after, E(y) or of ln (E(y)), without having to go through the process of estimating the variance function v(x) that is required for retransforming Log OLS results. Unfortunately, if the true model is a heteroscedastic equation for ln(y), then the GLM methods are less precise for dealing with some problems – the quadratic variance case. The precision of the GLM approaches is also diminished more by higher variance and kurtosis on the log scale than are OLS-based methods. Nevertheless, when the GLM models are designed for the data generating process, they can be substantially more precise than OLS-based methods.

In our analysis, we concentrated our attention on data generating mechanisms based on the log normal and on the gamma.[13]. Both of these have the characteristic that the raw scale standard deviation is a constant multiple of the mean – a constant coefficient of variation. It has been our experience that many health care expenditure and use data have this attribute. However, not all do. Some have relationships of raw scale means and variances that are characteristic of either the nonlinear least-squares (NLS) model or the Poisson-like models. In these cases, these other two GLM estimators are more precise than either OLS-based models or gamma regression models; our one NLS-like example illustrates this point.

The sensitivity of the results to common data issues appears to leave us with a quandary in model selection. If our *only* concern were bias in assessing the effect of $\partial \ln(E(y))/\partial x$, then we would recommend the GLM models. These also would be easier to use than the heteroscedastic retransformation suggested by Manning (1998) and used in many of the papers from the Health Insurance Experiment (Manning, et al., 1987; Newhouse et al., 1993). However, we are often quite concerned about precision; the usual difficulty that most health economists face is lack of precision due to the high variance in utilization and expenditures. Depending on the data, some GLM methods will be more precise than others. Unfortunately, we cannot rely on

[13] With some attention to NLS-like mechanisms with log-link functions, but additive error.

within-sample diagnostics to make the choice, because some models are more likely to lead to overfitting if the dependent variable is appreciably skewed right. This is particularly a problem for the NLS alternative suggested by Mullahy (1998) when facing very skewed data. Blough et al.'s (1999) recommendation of the gamma, and Mullahy's (1998) NLS recommendation are particularly sensitive to kurtosis on the log scale of the kind often seen in studies of hospital length of stay or inpatient costs.

Our recommendation is for the analyst to begin with both the raw-scale and log-scale residuals from one of the consistent estimators. If the log-scale residuals are heavy-tailed (kurtosis > 3), then consider the OLS-based models with ln(y) as the dependent variable. If there is no kurtosis on the log scale to speak of (k about 3), use the Park test on the raw-scale residuals to select one of the GLM models. If the raw-scale variance does not depend on the raw-scale prediction ($\lambda = 0$, in the notation of eq. (17)), then consider the NLS. If the raw-scale variance is proportional to the raw-scale prediction ($\lambda = 1$), consider the Poisson-like model. If the raw-scale variance is quadratic in the raw-scale prediction ($\lambda = 2$), then consider either the gamma model or the homoscedastic Log OLS model.[14] If the raw-scale variance is cubic in the raw-scale prediction ($\lambda = 3$), then consider the inverse Gaussian (Wald) model. Alternatively, one could use the results of the Park tests to estimate an iteratively, reweighted nonlinear least-squares model.

For those who are wedded to OLS based models with a logged dependent measure as a starting point, they should check the log scale residuals from the OLS model. If they are heteroscedastic in x, then the standard OLS analysis will be biased, unless corrected on retransformation by incorporating the log scale variance function v(x) or by moving to the GLM

---

[14] If the log scale residuals are symmetric, consider the log normal. A test for skewness is: $n(s^2) / 6$, where s is the skewness of the log scale residuals, which is distributed as chi square with one degree of freedom. A test for log normality can be formed using the skewness and kurtosis of the log scale residuals: $n[(s^2 / 6) + ((k - 3)^2 / 24)]$, which is distributed as chi square with two degrees of freedom; D'Agostino and Pearson (1973).

But if the pdf is monotonically declining, based on either plots of the data or the estimated shape parameter c α 1 under a gamma assumption, then the gamma model would appear to be more appropriate. One test for this possibility is to use the sample mean and variance to estimate c for the variable $z = y/[\exp(x\beta)]$, which is the analog of a residual for the gamma. A moment based estimate of c is the unconditional sample mean squared, divided by the unconditional sample variance.

approach outlined above. If they are leptokurtotic (kurtosis > 3), then the GLM models considered may be quite problematic.[15]

All of these checks can be done with tests readily available or programmable in major statistical and econometric programs. We are convinced that the return on the time spent on such analysis can be very high – in terms of major biases or losses in precision avoided.

While we have considered results here for the case where y is strictly positive (e.g. part two of a two-part model), it will be interesting in future work to assess the performance of these various estimation strategies for the more general case where y is nonnegative. Because we have been working largely in a mean-variance framework, there is ostensibly nothing in the above analysis that would preclude application to data where the realizations of y are either positive or zero, as is common in many health economics applications. It is, of course, an empirical matter as to whether a one-part or a two-part model is a more suitable characterization of the data (Mullahy, 1998), with the parsimony offered by a one-part model being desirable in some circumstances if an adequate variance function can be found to yield precise estimates. Assessing the relative merits of a conventional two-part model (Duan, Manning, et al., 1983: Manning, Duan, and Rogers 1987), the logit/gamma alternative (Blough et al., 1999) or some non-linear least-squares alternative (Mullahy, 1998) is a subject for further research. One of the major implications of the current research is that failure to closely approximate the true variance function could lead to major losses in precision for the GLM models.

---

[15] If the residuals are heteroscedastic in x, they will also be heavy-tailed. One way of generating a heavy-tailed distribution is to use a mixture model where the error term has zero expectation and different variance across observations. To rule out heteroscedasticity induced kurtosis, one can substitute the OLS residual divided by the square root of the estimated variance function. If this is leptokurtotic (kurtosis > 3), then the heteroscedastic version of OLS may be preferable to the GLM model. The choice will depend on how easy it is to model the variance function for log scale error versus how large the precision losses are as kurtosis rises.

REFERENCES

Blough, D.K., C.W. Madden, and M.C. Hornbrook. 1999. "Modeling risk using Generalized Linear Models." *Journal of Health Economics* 18: 153-171.

D'Agostino, R.B., and E.S.Pearson, 1973. "Tests for Departure from Normality: Empirical Results for the distribution of b_2 and b_1," *Biometrika* 60:613-622.

Duan, N. 1983. "Smearing Estimate: a Nonparametric Retransformation Method." *Journal of the American Statistical Association* 78: 605-610.

Duan, N., W.G. Manning, et al., 1983. "A Comparison of Alternative Models for the Demand for Medical Care." *Journal of Business and Economics Statistics* 1:115-126.

Jones, A. 1999. "Health Econometrics." in A. Culyer and J. Newhouse, eds., *Handbook of Health Economics*. Amsterdam: Elsevier (forthcoming).

Kennedy, P.1981, "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations." *American Economic Review* 71: 801.

Kennedy, P. 1983. "Logarithmic Dependent Variables and Prediction Bias." *Oxford Bulletin of Economics and Statistics* 45: 389-392.

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

Manning, W.G., N. Duan, and W.H. Rogers. 1987. "Monte Carlo Evidence on the Choice Between Sample Selection and Two-Part Models." *Journal of Econometrics* 35: 59-82.

Manning, W.G., J.P. Newhouse, N. Duan, et al., 1987. "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *American Economic Review*, 77(3):251-277.

Manning, W.G. 1998. "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem." *Journal of Health Economics* 17: 283-295.

Mullahy, J. 1998. "Much Ado about Two: Reconsidering Retransformation and the Two-part Model in Health Econometrics." *Journal of Health Economics* 17: 247-281.

Newhouse, J.P., et al., 1993. *Free-For-All: Health Insurance, Medical Costs, and Health Outcomes: The Results of the Health Insurance Experiment.* Cambridge: Harvard University Press.

Park, R. 1966. "Estimation with Heteroscedastic Error Terms." *Econometrica* 34: 888.

Taylor, J. M. 1986. "The Retransformed Mean after a Fitted Power Transformation." *Journal of the American Statistical Association* 81, 114-118.

Wooldridge, J.M. 1991. "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances." *Journal of Econometrics* 47: 5-46.

Table 1
Monte Carlo Simulation Design

A. Alternative Data Generating Models.

1. Alternative log normal models

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon$$

where x is uniform (0,1), ε is N(0,v) with variance v = 0.5, 1.0, 1.5, or 2.0, and $E(x'\varepsilon) = 0$. β_1 equals 1.0. β_0 is selected so that the E(y) = 1. Note: as the variance increases, the skewness and kurtosis of y increase.

2. Two alternative models with ε being heavy-tailed (kurtosis > 3). In the first, ε is a 90/10 mixture of normals with mean zero, and variances 1 and 3.3, respectively. In the second, the second variance is 4.6. The resulting kurtosis in ε is 4 and 5, respectively.

3. Gamma model with scale = $\exp(\beta_0 + \beta_1 x)$, where $\beta_1 = 1$, and β_0 is selected so that the E(y) = 1. The shape parameter c is 0.5, 1.0, or 4.0. The first and second have monotonically declining pdf's, conditional on x, while the last is bell shaped but skewed right. The second is an exponential variate. As the shape increases to infinity, the distribution approaches a normal.

4. An NLS-like structure where $y = [\exp(\beta_0 + \beta_1 x)] + \varepsilon$, with ε is N(0, 0.3).

5. Alternative heteroscedastic normal models. In the case above, ε is the product of a N(0,1) variable and either (1 + x) or its square root. The former has error variance that is linear in x, while the latter is quadratic in x. Again, $\beta_1 = 1$, and β_0 is selected so that the E(y) = 1.

Table 1 (cont'd)
Monte Carlo Simulation Design

B. Alternative Estimators.

1. OLS regression for ln(y) with a homoscedastic retransformation.

2. OLS regression for ln(y) with a heteroscedastic retransformation.

3. GLM for y with a log link, with a variance proportional to the E(y) -- a Poisson regression.

4. GLM for y with a log link, with a standard deviation proportional to the E(y) – a gamma regression.

5. Nonlinear least-squares by GLM for y with a log link, and an additive homoscedastic error term.

Except for the heteroscedastic case with standard deviation = 1+x, the covariate list includes an intercept and a single covariate x.

Table 2

Sample Statistics for Distribution

Log Normal Models	Log variance	mean	std.dev	skewness
	0.5	1	0.88	3.29
	1.0	1	1.39	6.59
	1.5	1	1.96	11.42
	2.0	1	2.62	17.94
Gamma Models	Shape			
	0.5	1	1.50	3.30
	1.0	1	1.08	2.42
	4.0	1	0.59	1.40
Heavy tailed distributions on log scale	Form			
	Mixed Normal 1*	1	16.44	126.44
	Mixed Normal 2**	1	17.13	174.63
Heteroscedastic in x on log scale	Log Variance			
	linear in x	1	2.27	18.79
	quadratic in x	1	5.11	55.17

* Kurtosis in ε = 4.0 on log scale

** Kurtosis in ε = 5.0 on log scale

Table 3
Effect of Skewness
on the Raw Scale
Coefficient on Slope of ln(E(y|x))

Generating Mechanism	Estimator	Mean	Std. Err.	95% Confidence Interval Lower	Upper
Log Normal Var = 0.5	LnOLS-Hom	1.0001	0.0237	0.9549	1.0457
	LnOLS-Het	1.0000	0.0264	0.9491	1.0537
	NLS	0.9998	0.0299	0.9407	1.0617
	Poisson	0.9998	0.0273	0.9461	1.0572
	Gamma	1.0000	0.0269	0.9476	1.0552
	True	1.0	------	------	------
Log Normal Var = 1.0	LnOLS-Hom	0.9996	0.0348	0.9322	1.0716
	LnOLS-Het	0.9985	0.0418	0.9157	1.0824
	NLS	0.9980	0.0505	0.9034	1.0953
	Poisson	0.9979	0.0462	0.9057	1.0867
	Gamma	0.9980	0.0447	0.9066	1.0859
	True	1.0	------	------	------
Log Normal Var = 1.5	LnOLS-Hom	1.000618	0.0428	0.9223	1.0891
	LnOLS-Het	1.00237	0.0567	0.8906	1.1125
	NLS	1.002461	0.0733	0.8557	1.1468
	Poisson	1.002091	0.0670	0.8671	1.1335
	Gamma	1.002465	0.0651	0.8740	1.1243
	True	1.0	------	------	------
Log Normal Var = 2.0	LnOLS-Hom	0.9991	0.0484	0.9035	1.0945
	LnOLS-Het	1.0013	0.0684	0.8640	1.1283
	NLS	1.0051	0.0982	0.8223	1.2140
	Poisson	1.0026	0.0882	0.8353	1.1739
	Gamma	1.0013	0.0845	0.8370	1.1734
	True	1.0	------	------	------

Table 4
Effect of Heavy Tails
on Log Scale
Coefficient on Slope of ln(E(y|x))

Generating Mechanism	Estimator	Mean	Std. Err.	95% Confidence Interval Lower	Upper
Log Normal	LnOLS-Hom	0.9996	0.0348	0.9322	1.0716
Var = 1.0	LnOLS-Het	0.9985	0.0418	0.9157	1.0824
k= 3	NLS	0.9980	0.0505	0.9034	1.0953
	Poisson	0.9979	0.0462	0.9057	1.0867
	Gamma	0.998083	0.0447	0.9066	1.0859
	True	1.0	------	------	------
Heavy Tailed	LnOLS-Hom	1.0002	0.0375	0.9274	1.0727
k = 4	LnOLS-Het	0.9994	0.0510	0.8973	1.1018
	NLS	1.0083	0.1737	0.7387	1.3421
	Poisson	1.0039	0.1426	0.7628	1.2883
	Gamma	1.0036	0.1320	0.7679	1.2544
	True	1.0	------	------	------
Heavy Tailed	LnOLS-Hom	1.000258	0.0396	0.9235	1.0765
k = 5	LnOLS-Het	0.999221	0.0593	0.8791	1.1176
	NLS	1.278733	5.0327	0.4137	1.9416
	Poisson	1.010899	0.3326	0.4566	1.6776
	Gamma	1.00994	0.2951	0.4344	1.5629
	True	1.0	------	------	------

Table 5
Effect of Shape
Coefficient on Slope of ln(E(y|x))

Generating Mechanism	Estimator	Mean	Std. Err.	95% Confidence Interval Lower	Upper
Gamma Shape = 0.5	LnOLS-Hom	1.000707	0.0748174	0.8528473	1.149035
	LnOLS-Het	0.997560	0.1733666	0.660885	1.33691
	NLS	0.999644	0.0526338	0.8939781	1.111083
	Poisson	0.999670	0.0481553	0.9044225	1.099688
	Gamma	1.000058	0.0473053	0.9108036	1.096514
	True	1.0	---	---	---
Gamma Shape = 1.0	LnOLS-Hom	0.999915	0.0449406	0.9128188	1.089832
	LnOLS-Het	0.998422	0.0502031	0.8991024	1.095391
	NLS	0.998720	0.0386417	0.9243465	1.073881
	Poisson	0.998785	0.035305	0.9292879	1.066697
	Gamma	0.998973	0.0342242	0.9300015	1.06794
	True	1.0	---	---	---
Gamma Shape = 4.0	LnOLS-Hom	1.000154	0.0186658	0.9657184	1.038561
	LnOLS-Het	1.000517	0.0176041	0.9663996	1.034754
	NLS	1.000467	0.0202586	0.9635911	1.041706
	Poisson	1.000405	0.0183169	0.9665445	1.037278
	Gamma	1.000432	0.0174984	0.9673809	1.035323
	True	1.0	---	---	---

Table 6
Effect of Heteroscedasticity
on the Log Scale
Coefficient on Slope of ln(E(y|x))

Generating Mechanism	Estimator	Mean	Std. Err.	95% Confidence Interval Lower	Upper
Log Normal	LnOLS-Hom	0.999621	0.0348148	0.9322314	1.071647
Var = 1.0	LnOLS-Het	0.998580	0.0418082	0.9157929	1.082485
	NLS	0.998001	0.0505482	0.9034469	1.095328
	Poisson	0.997919	0.0462138	0.9057372	1.08671
	Gamma	0.998083	0.0447803	0.906603	1.085904
	True	1.0	---------	---------	---------
Hetero	LnOLS-Hom	1.000069	0.0407556	0.9190375	1.079981
Var = 1+x	LnOLS-Het	1.499164	0.0546495	1.391562	1.608531
	NLS	1.499841	0.1025363	1.304943	1.721955
	Poisson	1.498189	0.0783845	1.341698	1.663669
	Gamma	1.498431	0.0661695	1.369247	1.639189
	True	1.5	---------	--------	--------
Hetero	LnOLS-Hom	0.999664	0.0539657	0.8966974	1.115192
Std = 1+x	LnOLS-Het	2.494235	0.0825744	2.324784	2.65801
	NLS	2.277278	2.438436	0.5461946	4.069843
	Poisson	2.270902	0.3723323	1.484793	2.982523
	Gamma	2.256181	0.1939853	1.855002	2.615747
	True	2.5	---------	---------	---------

Note: For the log normal case where the standard deviation of ε is 1+x, the slope is evaluated at x=0.5.

Table 7
Efficiency Effects
Coefficient on Slope of ln(E(y|x))

Generating Mechanism	Estimator	Mean	Std. Err.	95% Conf. Interval Lower	Upper
Log Normal Var = 1.0	LnOLS-Hom	0.999621	0.0348148	0.9322314	1.071647
	LnOLS-Het	0.998580	0.0418082	0.9157929	1.082485
	NLS	0.998001	0.0505482	0.9034469	1.095328
	Poisson	0.997919	0.0462138	0.9057372	1.08671
	Gamma	0.998083	0.0447803	0.906603	1.085904
	True	1.0	---------	---------	--------
Gamma Shape = 1.0	LnOLS-Hom	0.999915	0.0449406	0.9128188	1.089832
	LnOLS-Het	0.998422	0.0502031	0.8991024	1.095391
	NLS	0.998720	0.0386417	0.9243465	1.073881
	Poisson	0.998785	0.035305	0.9292879	1.066697
	Gamma	0.998973	0.0342242	0.9300015	1.06794
	True	1.0	---------	---------	--------
NLS Additive error	LnOLS-Hom	1.043221	0.0080515	1.027155	1.058511
	LnOLS-Het	0.993504	0.0074022	0.9790431	1.007654
	NLS	1.000138	0.0061217	0.9885183	1.012165
	Poisson	1.000133	0.0064606	0.9874387	1.012556
	Gamma	1.000118	0.0072247	0.9856906	1.013724
	True	1.0	---------	---------	--------
Hetero Var = 1+x	LnOLS-Hom	1.000069	0.0407556	0.9190375	1.079981
	LnOLS-Het	1.499164	0.0546495	1.391562	1.608531
	NLS	1.499841	0.1025363	1.304943	1.721955
	Poisson	1.498189	0.0783845	1.341698	1.663669
	Gamma	1.498431	0.0661695	1.369247	1.639189
	True	1.50	---------	--------	--------

Table 8
Comparisons of Alternative Estimators
Park Tests of Mean-Variance Relationship
Estimates of λ

Generating Mechanism	Estimator	Mean	Std. Err.	95% Conf. Interval Lower	Upper
Log Normal	Ln-OLS-Hom	2.0005	0.0723	1.8562	2.1471
Var. = 1.0	Ln-OLS-Het	1.9998	0.0726	1.8533	2.1408
	NLS	1.9994	0.0737	1.8499	2.1417
	Poisson	1.9995	0.0735	1.8480	2.1398
	Gamma	2.0000	0.0731	1.8486	2.1424
	True	2.0			
Gamma	Ln-OLS-Hom	2.0057	0.0788	1.8545	2.1602
Shape = 1.0	Ln-OLS-Het	2.0003	0.0675	1.8684	2.1305
	NLS	2.0033	0.0692	1.8632	2.1309
	Poisson	2.0031	0.0688	1.8630	2.1322
	Gamma	2.0032	0.0689	1.8671	2.1302
	True	2.0			
Heavy tailed	Ln-OLS-Hom	1.9964	0.0655	1.8653	2.1267
k = 5	Ln-OLS-Het	1.9951	0.0782	1.8271	2.1391
	NLS	2.1372	5.7609	1.2939	2.2001
	Poisson	2.1463	5.7867	1.3535	2.2201
	Gamma	2.1512	5.8152	1.3457	2.2252
	True	2.0			
Nonlinear	Ln-OLS-Hom	0.02159	0.0729	-0.1289	0.1661
Additive	Ln-OLS-Het	0.00319	0.0780	-0.1536	0.1532
Error	NLS	0.00301	0.0782	-0.1560	0.1548
	Poisson	0.00313	0.0782	-0.1567	0.1540
	Gamma	0.00317	0.0783	-0.1559	0.1567
	True	0.0			
Log Normal	Ln-OLS-Hom	2.7611	0.0763	2.6154	2.9196
Var = 1+x	Ln-OLS-Het	2.3801	0.0248	2.3299	2.4281
	NLS	2.3510	0.1259	2.1875	2.4304
	Poisson	2.3795	0.0299	2.3225	2.4360
	Gamma	2.3827	0.0283	2.3263	2.4370
	True	???			

Note: Estimate of slope λ from log OLS version of Park test.

Table 9

GLM Estimates of NHIS Doctor Visit Data: Alternative λ Values
(Heteroskedasticity-robust std. errors in parentheses)

	λ				
Variable	0	1	2	3	1.917887
CONSTANT	2.7492	2.7556	2.7778	2.8135	2.7831
	(0.1298)	(0.0881)	(0.0728)	(0.0760)	(0.0726)
AGE	-0.0091	-0.0078	-0.0068	-0.0062	-0.0070
	(0.0019)	(0.0012)	(0.0010)	(0.0011)	(0.0010)
MALE	-0.0471	-0.1059	-0.1446	-0.1630	-0.1393
	(0.0446)	(0.0304)	(0.0257)	(0.0266)	(0.0256)
EDUC	0.0372	0.0308	0.0239	0.0176	0.0239
	(0.0067)	(0.0047)	(0.0041)	(0.0046)	(0.0041)
WHITE	0.1324	0.1514	0.1625	0.1654	0.1622
	(0.0524)	(0.0380)	(0.0322)	(0.0334)	(0.0317)
MARRIED	-0.1478	-0.1358	-0.1132	-0.0940	-0.1173
	(0.0414)	(0.0295)	(0.0263)	(0.0282)	(0.0262)
XCELLENT	-1.6125	-1.5753	-1.5495	-1.5328	-1.5412
	(0.0477)	(0.0410)	(0.0393)	(0.0408)	(0.0391)
VERYGOOD	-1.3395	-1.3107	-1.2909	-1.2769	-1.2844
	(0.0449)	(0.0391)	(0.0374)	(0.0384)	(0.0370)
GOOD	-0.8561	-0.8468	-0.8420	-0.8386	-0.8386
	(0.0441)	(0.0405)	(0.0396)	(0.0405)	(0.0388)

Appendix Table 1
Simulation Results for
Coefficient on Slope of ln(E(y|x))

Generating Mechanism	Estimator	Mean	Std. Err.	95% Confidence Interval Lower	Upper
Log Normal Var = 0.5	LnOLS-Hom	1.000133	0.023751	0.9549728	1.045786
	LnOLS-Het	1.000017	0.0264442	0.9491883	1.053797
	NLS	0.999853	0.0299333	0.9407199	1.061769
	Poisson	0.999886	0.0273986	0.9461934	1.057224
	Gamma	1.000029	0.0269116	0.9476735	1.05529
Log Normal Var = 1.0	LnOLS-Hom	0.999621	0.0348148	0.9322314	1.071647
	LnOLS-Het	0.998580	0.0418082	0.9157929	1.082485
	NLS	0.998001	0.0505482	0.9034469	1.095328
	Poisson	0.997919	0.0462138	0.9057372	1.08671
	Gamma	0.998083	0.0447803	0.906603	1.085904
Log Normal Var = 1.5	LnOLS-Hom	1.000618	0.0428613	0.922383	1.089104
	LnOLS-Het	1.00237	0.0567259	0.890652	1.112596
	NLS	1.002461	0.0733045	0.8557558	1.146891
	Poisson	1.002091	0.0670108	0.8671428	1.133539
	Gamma	1.002465	0.065142	0.874099	1.124307
Log Normal Var = 2.0	LnOLS-Hom	0.999136	0.0484478	0.9035519	1.094575
	LnOLS-Het	1.001304	0.0684748	0.8640262	1.128377
	NLS	1.005102	0.0982818	0.8223573	1.214009
	Poisson	1.002657	0.0882449	0.8353381	1.173998
	Gamma	1.001392	0.0845442	0.8370611	1.173467
Heavy Tailed k = 4	LnOLS-Hom	1.000238	0.0375558	0.9274483	1.072745
	LnOLS-Het	0.999493	0.0510681	0.897309	1.101875
	NLS	1.008394	0.1737535	0.7387403	1.342172
	Poisson	1.003922	0.1426515	0.762818	1.28838
	Gamma	1.003618	0.1320229	0.7679142	1.254455
Heavy Tailed k = 5	LnOLS-Hom	1.000258	0.0396378	0.9235228	1.076574
	LnOLS-Het	0.999221	0.0593454	0.8791245	1.117663
	NLS	1.278733	5.032764	0.413769	1.941622
	Poisson	1.010899	0.3326238	0.456693	1.677678
	Gamma	1.00994	0.2951944	0.434469	1.562959

Appendix Table 1 (cont'd)
Simulation Results for β_1

Generating Mechanism	Estimator	Mean	Std. Err.	95% Confidence Interval Lower	Upper
Gamma Shape = 0.5	LnOLS-Hom	1.000707	0.0748174	0.8528473	1.149035
	LnOLS-Het	0.997560	0.1733666	0.660885	1.33691
	NLS	0.999644	0.0526338	0.8939781	1.111083
	Poisson	0.999670	0.0481553	0.9044225	1.099688
	Gamma	1.000058	0.0473053	0.9108036	1.096514
Gamma Shape = 1.0	LnOLS-Hom	0.999915	0.0449406	0.9128188	1.089832
	LnOLS-Het	0.998422	0.0502031	0.8991024	1.095391
	NLS	0.998720	0.0386417	0.9243465	1.073881
	Poisson	0.998785	0.035305	0.9292879	1.066697
	Gamma	0.998973	0.0342242	0.9300015	1.06794
Gamma Shape = 4.0	LnOLS-Hom	1.000154	0.0186658	0.9657184	1.038561
	LnOLS-Het	1.000517	0.0176041	0.9663996	1.034754
	NLS	1.000467	0.0202586	0.9635911	1.041706
	Poisson	1.000405	0.0183169	0.9665445	1.037278
	Gamma	1.000432	0.0174984	0.9673809	1.035323
NLS Additive error	LnOLS-Hom	1.043221	0.0080515	1.027155	1.058511
	LnOLS-Het	0.993504	0.0074022	0.9790431	1.007654
	NLS	1.000138	0.0061217	0.9885183	1.012165
	Poisson	1.000133	0.0064606	0.9874387	1.012556
	Gamma	1.000118	0.0072247	0.9856906	1.013724
Hetero Var = 1+x	LnOLS-Hom	1.000069	0.0407556	0.9190375	1.079981
	LnOLS-Het	1.499164	0.0546495	1.391562	1.608531
	NLS	1.499841	0.1025363	1.304943	1.721955
	Poisson	1.498189	0.0783845	1.341698	1.663669
	Gamma	1.498431	0.0661695	1.369247	1.639189
Hetero Std = 1+x	LnOLS-Hom	0.999664	0.0539657	0.8966974	1.115192
	LnOLS-Het	2.494235	0.0825744	2.324784	2.65801
	NLS	2.277278	2.438436	0.5461946	4.069843
	Poisson	2.270902	0.3723323	1.484793	2.982523
	Gamma	2.256181	0.1939853	1.855002	2.615747

"Mean" evaluated at x=0.50 for log normal model with heteroscedasticity, std=1+x.

Appendix Table 2
Mean Squared Error (MSE) and Absolute Prediction Error (APE)

Generating Mechanism	Estimators Compared A	B	MSE A < B	APE A < B
Log Normal Var = 0.5	LnOLS-Hom	LnOLS-Het	915	576
	LnOLS-Hom	NLS	0	614
	LnOLS-Hom	Poisson	145	580
	LnOLS-Hom	Gamma	356	578
	LnOLS-Het	LnOLS-Hom	85	424
	LnOLS-Het	NLS	0	576
	LnOLS-Het	Poisson	0	550
	LnOLS-Het	Gamma	0	537
	NLS	LnOLS-Hom	997	386
	NLS	LnOLS-Het	1000	424
	NLS	Poisson	927	395
	NLS	Gamma	977	432
	Poisson	LnOLS-Hom	848	420
	Poisson	LnOLS-Het	1000	450
	Poisson	Gamma	942	477
	Gamma	LnOLS-Hom	638	422
	Gamma	LnOLS-Het	1000	463
	Gamma	NLS	0	568
	Gamma	Poisson	28	523
Log Normal Var = 1.0	LnOLS-Hom	LnOLS-Het	837	577
	LnOLS-Hom	NLS	0	671
	LnOLS-Hom	Poisson	138	636
	LnOLS-Hom	Gamma	289	615
	LnOLS-Het	LnOLS-Hom	163	423
	LnOLS-Het	NLS	0	611
	LnOLS-Het	Poisson	0	596
	LnOLS-Het	Gamma	1	566
	NLS	LnOLS-Hom	999	329
	NLS	LnOLS-Het	1000	389
	NLS	Poisson	934	385
	NLS	Gamma	985	415
	Poisson	LnOLS-Hom	856	364
	Poisson	LnOLS-Het	1000	404
	Poisson	Gamma	941	438
	Gamma	LnOLS-Hom	710	385
	Gamma	LnOLS-Het	999	434
	Gamma	NLS	0	585
	Gamma	Poisson	36	562

Appendix Table 2 (cont'd)
Mean Squared Error (MSE) and Absolute Prediction Error (APE)

Generating Mechanism	Estimators Compared A	B	MSE A < B	APE A < B
Log Normal Var = 1.5	LnOLS-Hom	LnOLS-Het	829	638
	LnOLS-Hom	NLS	0	691
	LnOLS-Hom	Poisson	128	666
	LnOLS-Hom	Gamma	258	669
	LnOLS-Het	LnOLS-Hom	171	362
	LnOLS-Het	NLS	0	609
	LnOLS-Het	Poisson	0	599
	LnOLS-Het	Gamma	7	592
	NLS	LnOLS-Hom	999	309
	NLS	LnOLS-Het	1000	391
	NLS	Poisson	936	396
	NLS	Gamma	972	433
	Poisson	LnOLS-Hom	869	334
	Poisson	LnOLS-Het	1000	401
	Poisson	Gamma	935	479
	Gamma	LnOLS-Hom	742	331
	Gamma	LnOLS-Het	993	408
	Gamma	NLS	2	567
	Gamma	Poisson	34	521
Log Normal Var = 2.0	LnOLS-Hom	LnOLS-Het	813	637
	LnOLS-Hom	NLS	0	704
	LnOLS-Hom	Poisson	116	705
	LnOLS-Hom	Gamma	261	684
	LnOLS-Het	LnOLS-Hom	187	363
	LnOLS-Het	NLS	0	625
	LnOLS-Het	Poisson	0	593
	LnOLS-Het	Gamma	3	591
	NLS	LnOLS-Hom	998	296
	NLS	LnOLS-Het	1000	375
	NLS	Poisson	955	400
	NLS	Gamma	984	441
	Poisson	LnOLS-Hom	875	295
	Poisson	LnOLS-Het	1000	407
	Poisson	Gamma	941	471
	Gamma	LnOLS-Hom	737	316
	Gamma	LnOLS-Het	997	409
	Gamma	NLS	0	559
	Gamma	Poisson	33	529

Appendix Table 2 (cont'd)
Mean Squared Error (MSE) and Absolute Prediction Error (APE)

Generating Mechanism	Estimators Compared A	B	MSE A < B	APE A < B
Heavy Tailed k = 4	LnOLS-Hom	LnOLS-Het	824	640
	LnOLS-Hom	NLS	0	818
	LnOLS-Hom	Poisson	105	823
	LnOLS-Hom	Gamma	258	822
	LnOLS-Het	LnOLS-Hom	176	360
	LnOLS-Het	NLS	0	766
	LnOLS-Het	Poisson	0	778
	LnOLS-Het	Gamma	10	786
	NLS	LnOLS-Hom	998	182
	NLS	LnOLS-Het	1000	234
	NLS	Poisson	952	382
	NLS	Gamma	985	418
	Poisson	LnOLS-Hom	891	177
	Poisson	LnOLS-Het	1000	222
	Poisson	Gamma	950	459
	Gamma	LnOLS-Hom	739	178
	Gamma	LnOLS-Het	990	214
	Gamma	NLS	1	582
	Gamma	Poisson	37	541
Heavy Tailed k = 5	LnOLS-Hom	LnOLS-Het	797	669
	LnOLS-Hom	NLS	0	900
	LnOLS-Hom	Poisson	107	907
	LnOLS-Hom	Gamma	237	900
	LnOLS-Het	LnOLS-Hom	203	331
	LnOLS-Het	NLS	0	863
	LnOLS-Het	Poisson	0	862
	LnOLS-Het	Gamma	10	855
	NLS	LnOLS-Hom	1000	100
	NLS	LnOLS-Het	1000	137
	NLS	Poisson	950	382
	NLS	Gamma	986	420
	Poisson	LnOLS-Hom	889	93
	Poisson	LnOLS-Het	1000	138
	Poisson	Gamma	944	457
	Gamma	LnOLS-Hom	761	100
	Gamma	LnOLS-Het	990	145
	Gamma	NLS	4	580
	Gamma	Poisson	35	543

Appendix Table 2 (cont'd)
Mean Squared Error (MSE)and Absolute Prediction Error (APE)

Generating Mechanism	Estimators Compared A	B	MSE A < B	APE A < B
Gamma	LnOLS-Hom	LnOLS-Het	1000	759
Shape = 0.5	LnOLS-Hom	NLS	0	379
	LnOLS-Hom	Poisson	96	340
	LnOLS-Hom	Gamma	199	326
	LnOLS-Het	LnOLS-Hom	0	241
	LnOLS-Het	NLS	0	183
	LnOLS-Het	Poisson	0	173
	LnOLS-Het	Gamma	0	176
	NLS	LnOLS-Hom	1000	621
	NLS	LnOLS-Het	1000	817
	NLS	Poisson	931	395
	NLS	Gamma	965	433
	Poisson	LnOLS-Hom	902	660
	Poisson	LnOLS-Het	1000	827
	Poisson	Gamma	933	476
	Gamma	LnOLS-Hom	801	674
	Gamma	LnOLS-Het	1000	824
	Gamma	NLS	0	567
	Gamma	Poisson	30	524
Gamma	LnOLS-Hom	LnOLS-Het	1000	532
Shape = 1.0	LnOLS-Hom	NLS	0	434
	LnOLS-Hom	Poisson	105	380
	LnOLS-Hom	Gamma	255	364
	LnOLS-Het	LnOLS-Hom	0	468
	LnOLS-Het	NLS	0	405
	LnOLS-Het	Poisson	0	358
	LnOLS-Het	Gamma	0	340
	NLS	LnOLS-Hom	995	566
	NLS	LnOLS-Het	1000	595
	NLS	Poisson	927	406
	NLS	Gamma	965	435
	Poisson	LnOLS-Hom	890	620
	Poisson	LnOLS-Het	1000	642
	Poisson	Gamma	922	483
	Gamma	LnOLS-Hom	740	636
	Gamma	LnOLS-Het	1000	660
	Gamma	NLS	0	565
	Gamma	Poisson	29	517

Appendix Table 2 (cont'd)
Mean Squared Error (MSE)and Absolute Prediction Error (APE)

Generating Mechanism	Estimators Compared A	B	MSE A < B	APE A < B
Gamma Shape = 4.0	LnOLS-Hom	LnOLS-Het	962	453
	LnOLS-Hom	NLS	0	557
	LnOLS-Hom	Poisson	140	481
	LnOLS-Hom	Gamma	363	449
	LnOLS-Het	LnOLS-Hom	38	547
	LnOLS-Het	NLS	0	610
	LnOLS-Het	Poisson	0	559
	LnOLS-Het	Gamma	3	476
	NLS	LnOLS-Hom	996	443
	NLS	LnOLS-Het	1000	390
	NLS	Poisson	925	352
	NLS	Gamma	975	387
	Poisson	LnOLS-Hom	852	519
	Poisson	LnOLS-Het	1000	441
	Poisson	Gamma	935	423
	Gamma	LnOLS-Hom	633	551
	Gamma	LnOLS-Het	997	524
	Gamma	NLS	0	613
	Gamma	Poisson	24	577
Non Linear Additive error	LnOLS-Hom	LnOLS-Het	0	10
	LnOLS-Hom	NLS	0	1
	LnOLS-Hom	Poisson	0	1
	LnOLS-Hom	Gamma	0	2
	LnOLS-Het	LnOLS-Hom	1000	990
	LnOLS-Het	NLS	0	309
	LnOLS-Het	Poisson	0	322
	LnOLS-Het	Gamma	36	331
	NLS	LnOLS-Hom	1000	999
	NLS	LnOLS-Het	1000	691
	NLS	Poisson	952	562
	NLS	Gamma	992	617
	Poisson	LnOLS-Hom	1000	999
	Poisson	LnOLS-Het	1000	678
	Poisson	Gamma	949	648
	Gamma	LnOLS-Hom	1000	998
	Gamma	LnOLS-Het	964	669
	Gamma	NLS	0	383
	Gamma	Poisson	33	352

Appendix Table 2 (cont'd)
Mean Squared Error (MSE) and Absolute Prediction Error (APE)

Generating Mechanism	Estimators Compared A	B	MSE A < B	APE A < B
Log Normal Var = 1+x	LnOLS-Hom	LnOLS-Het	0	0
	LnOLS-Hom	NLS	0	0
	LnOLS-Hom	Poisson	0	0
	LnOLS-Hom	Gamma	0	0
	LnOLS-Het	LnOLS-Hom	1000	1000
	LnOLS-Het	NLS	0	735
	LnOLS-Het	Poisson	0	677
	LnOLS-Het	Gamma	12	612
	NLS	LnOLS-Hom	1000	1000
	NLS	LnOLS-Het	1000	265
	NLS	Poisson	964	276
	NLS	Gamma	989	309
	Poisson	LnOLS-Hom	1000	1000
	Poisson	LnOLS-Het	1000	323
	Poisson	Gamma	925	376
	Gamma	LnOLS-Hom	1000	1000
	Gamma	LnOLS-Het	988	388
	Gamma	NLS	0	691
	Gamma	Poisson	59	624
Log Normal std = 1+x	LnOLS-Hom	LnOLS-Het	0	0
	LnOLS-Hom	NLS	0	258
	LnOLS-Hom	Poisson	0	76
	LnOLS-Hom	Gamma	0	14
	LnOLS-Het	LnOLS-Hom	1000	1000
	LnOLS-Het	NLS	0	699
	LnOLS-Het	Poisson	0	550
	LnOLS-Het	Gamma	35	499
	NLS	LnOLS-Hom	1000	742
	NLS	LnOLS-Het	1000	301
	NLS	Poisson	988	230
	NLS	Gamma	999	258
	Poisson	LnOLS-Hom	1000	924
	Poisson	LnOLS-Het	1000	450
	Poisson	Gamma	926	371
	Gamma	LnOLS-Hom	1000	986
	Gamma	LnOLS-Het	965	501
	Gamma	NLS	0	742
	Gamma	Poisson	71	629