ELSEVIER

Journal of Health Economics 20 (2001) 461–494

JOURNAL OF HEALTH ECONOMICS

www.elsevier.nl/locate/econbase

Estimating log models: to transform or not to transform?☆

Willard G. Manning [a,*], John Mullahy [b]

[a] *Department of Health Studies, Biological Sciences Division, Harris School of Public Policy Studies, The University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637, USA*

[b] *Departments of Preventive Medicine and Economics, University of Wisconsin and National Bureau of Economic Research, Madison, WI 53705, USA*

Received 1 July 2000; received in revised form 1 March 2001; accepted 20 March 2001

Abstract

Health economists often use log models to deal with skewed outcomes, such as health utilization or health expenditures. The literature provides a number of alternative estimation approaches for log models, including ordinary least-squares on ln(y) and generalized linear models. This study examines how well the alternative estimators behave econometrically in terms of bias and precision when the data are skewed or have other common data problems (heteroscedasticity, heavy tails, etc.). No single alternative is best under all conditions examined. The paper provides a straightforward algorithm for choosing among the alternative estimators. Even if the estimators considered are consistent, there can be major losses in precision from selecting a less appropriate estimator.

JEL classification: C1 econometric and statistical methods: general; C5 econometric modeling

Keywords: Health econometrics; Transformation; Retransformation; Log models

1. Introduction

Health economists need little convincing that many of the outcomes with which they are concerned are awkward to analyze empirically; see Jones (2000) 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

☆ An earlier version of this paper was presented at the Second World Conference of the International Health Economics Association, Rotterdam, The Netherlands, 6–9 June 1999, and published as NBER Technical Report 0246.

* Corresponding author. Tel.: +1-773-834-1971; fax: +1-773-702-1979.
E-mail address: w-manning@uchicago.edu (W.G. Manning).

0167-6296/01/$ – see front matter
PII: S0167-6296(01)00086-8

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 et al., 1983; Jones, 2000; Manning, 1998; Mullahy, 1998; Blough et al., 1999). For count variables, such as utilization, there is an additional literature based on Poisson and negative binomial models (Jones, 2000; Cameron and Trivedi, 1998). A few investigators have also examined the use of duration models for health expenditures and length-of-stay; for a recent review, see Jones (2000, Section 8).

In this paper, we focus our attention on the positive parts of health economic outcomes where we are often concerned with the impact of out-of-pocket price, income, health status or some other economic or health covariates on the expenditures or visits by users of health care or the impact on some other positive economic outcome. The twin primary concerns are to obtain unbiased and precise estimates of the impact of those covariates in the face of the third of the three characteristics mentioned above — positively skewed dependent variables. The recent literature has suggested three different approaches to addressing this problem (Manning, 1998; Mullahy, 1998; Blough et al., 1999). These articles did not provide evidence on how well their estimators would behave under a range of data conditions, nor did they provide an algorithm for choosing among the alternatives. In this paper, we try to fill both of these gaps, and to illustrate the approaches using examples from health care utilization and earnings.

This paper provides some simulation-based evidence on the finite-sample behavior of some of the estimators designed to look at the effect of a set of covariates x on the expected outcome, $E(y)$, when y is strictly positive, under a range of data problems encountered in every day practice. We assume that the researcher wants to make a statement about mean or total outcomes or expenditures, rather then median outcomes or expenditures. We work largely within the two classes of estimators: two derived from least-squares (LS) estimators for the $\ln(y)$, and some of the generalized linear models (GLM) with log links, which can simply be viewed as differentially weighted nonlinear least-squares estimators. We consider the first- and second-order behavior — bias and precision — of the least-squares and GLM estimators under alternative assumptions about the data generating processes. While these two classes of models — the LS-based and GLM — overlap for some model assumptions, neither is a proper subset of the other. Thus, we cannot nest the choices in a broader class of models, and test which member applies.

We investigate the performance of two variants of the traditional OLS model for the $\ln(y)$. Although technically, these are models for the expectation of $\ln(y)$, rather than for the natural log of the expectation, they are interesting for two reasons. First, OLS for $\ln(y)$ is one of the most prevalently used (and most prevalently misused) models for analyzing such data. Second, it is possible to go from the $E(\ln(y|x))$ to the $\ln(E(y|x))$ by retransformation (Duan, 1983; Manning, 1998). The GLM models considered here provide estimates of the $\ln(E(y|x))$ and $E(y|x)$ directly, without any requirement for retransformation.

The results indicate that there can be important tradeoffs among the estimators in terms of precision and bias. The LS-based methods can be biased in the face of heteroscedasticity if not appropriately retransformed (Manning, 1998; Mullahy, 1998). The GLM models can yield very imprecise estimates if the log-scale error is heavy-tailed. Even if the estimators

considered are consistent, there can be major losses in precision from selecting a less appropriate estimator. Choosing a less appropriate estimator can cause precision losses equivalent to the loss of one half or more of one's sample.

We develop a method for determining which estimation method to choose for any application using tests that are relatively easy to implement. The method relies on estimating both the OLS model for $\ln(y) = xd+\varepsilon$, and one of the GLM models for the $\ln(E(y|x)) = x\beta$, and generating log-scale and raw-scale residuals for the two models, respectively. Tests based on these two sets of residuals will indicate whether to use OLS on $\ln(y)$ or which GLM model to use for the $\ln(E(y|x))$. If the OLS residuals on the log-scale are heteroscedastic in some x, then one should employ one of the GLM models or do a heteroscedastic retransformation to avoid the bias in statements about $E(y|x)$. We provide a simple extension of Park's (1966) test applied to the raw-scale residuals from the GLM model to determine which specific GLM model to use. Even in the absence of heteroscedasticity, there are cases where the GLM approach is more precise than OLS on $\ln(y)$. We provide a simple test using the OLS residuals for one of these cases. If the OLS residuals on the log-scale are heavier tailed than a normal, then one should employ OLS for $\ln(y)$ to reduce the precision losses. If the log-scale residuals from the OLS model are symmetric or if the variances are large (≥ 1), then OLS on $\ln(y)$ is indicated.

In either of the cases of the GLM or suitably retransformed OLS for $\ln(y)$ estimators, all of the usual interpretations of the coefficients from a log model will be retained, while avoiding the bias and precision problems that can arise. The models considered are easy to estimate given modern software packages, and the tests are relatively straightforward.

The plan for the paper is as follows. Section 2 describes the general modeling approaches that we consider. Section 3 presents our simulation framework. Section 4 summarizes the results of the simulations and two empirical examples that focus on the outcomes of annual physician visits and annual earnings; the latter indicates that these modeling issues are not limited to health economics and health services research. Section 5 contains our proposed algorithm for choosing among the competing estimators for log models.

2. Modeling framework

In what follows, we adopt the perspective that the purpose of the analysis is to say something about how the expected outcome, $E(y|x)$, responds to shifts in a set of covariates x.[1] Whether $E(y|x)$ will always be the most interesting feature of the joint distribution $\phi(y, x)$ to analyze is, of course, a situation-specific issue. However, the prominence 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 1 of a two-part model") are ignored here, but will be addressed in future work. We also do not consider problems of censoring or unequal periods of observation.

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

Our modeling framework includes two classes of estimators: generalized linear models (GLM) with a logarithmic link function, and least-squares for models for logged dependent variables.[2] The specific GLM models estimate the $\ln(E(y|x))$ directly, while the least-squares model estimate $E(\ln(y)|x)$, which can at least in principle be converted to $E(y|x)$ by a suitable retransformation. As we have stressed elsewhere (Manning, 1998; Mullahy, 1998), it is essential to distinguish these related but distinct models.

2.1. 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 regression model is

$$\ln(y) = x\delta + \varepsilon \tag{1}$$

where we assumed that $E(\varepsilon) = 0$ and $E(x'\varepsilon) = 0$; the error term ε need not be i.i.d. If the error term is normally distributed $N(0, \sigma_\varepsilon^2)$, 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 $\exp(\varepsilon)$ has constant mean and variance, then $E(y|x) = s\exp(x\delta)$, where $s = E(\exp(\varepsilon))$.[3] 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. $E(\exp(\varepsilon))$ is some function $f(x)$, then $E(y|x) = f(x)\exp(x, \delta)$, or, equivalently,

$$\ln(E(y|x)) = x\delta + \ln(f(x)) \tag{2}$$

and in the log normal case,

$$\ln(E(y|x)) = x\delta + 0.5\sigma_\varepsilon^2(x) \tag{3}$$

where the last term in Eq. (3) is the error variance as a function of x on the log-scale.[4]

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

2.2. 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)

[2] The same issues that we raise for log models also apply to all models with nonlinear transformations of dependent variables (such as Box–Cox models) or nonlinear link functions in GLM. In those cases, the choice will be between the Box–Cox transformation of the dependent variable y or a GLM model with a power link function. Here we focus on the log version because of its widespread use.

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

[4] Although the log transformation can resolve heteroscedasticity on the raw-scale, it seems unlikely that heteroscedasticity on the log-scale will remove it on the raw-scale, unless $\sigma_\varepsilon^2(x) = -2x\beta$.

or log link relationship:

$$\ln(E(y|x)) = x\beta \tag{4a}$$

or

$$E(y|x) = \exp(x\beta) = \mu(x;\beta) \tag{4b}$$

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

$$\text{var}(y|x) = \sigma^2 v(x) \tag{5}$$

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 $v(y|x) = \kappa_1 \mu(x)$, where $\kappa_1 > 0$; that is the variance is proportional to the mean, which is itself a function of x; $\kappa_1 > 1$ indicates the degree of "overdispersion". The third has a gamma structure with $v(y|x) = \kappa_2(\mu(x))^2$, where $\kappa_2 > 0$; that is, the standard deviation is proportional to the mean. [5]

Within this class of power-proportional variance functions, it is useful to think more generally of the variance function $v(y|x)$ being

$$v(y\,|x\,) = \kappa(\mu(x\beta))^{\lambda} \tag{6}$$

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 gamma, the homoscedastic log normal, the Weibull, and the Chi-square, with the suitable specification of a distribution. [6] In the case $\lambda = 3$, we get the inverse Gaussian (or Wald) distribution; we do not consider that estimator here. 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 generalized method of moments (GMM) estimation but what is more generally spoken of by statisticians as GLM modeling using

[5] We do not consider two other GLM models. The first is the inverse Gaussian (Wald) distribution for situations where the variance function is proportional to the cube of the mean function. The second is the negative Binomial distribution, which can be generated as a gamma mixture of Poissons. Its variance function is a specific quadratic function of the mean. This distribution has been widely used for count data.

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

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

and if u is taken to be homoscedastic, then it is indeed natural to suggest that $\text{var}(y|x)$ is proportional to $E(y|x)$-squared. Thus, just as the homoscedastic 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.

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

$$\sum_{i=1}^{N} \frac{\partial \mu(x_i;\beta)}{\partial \beta}(V(y_i\,|x_i))^{-1}(y_i - \mu(x_i;\beta)) = 0 \tag{7}$$

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

3. 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 raw-scale variable, (2) heavy-tailed distributions (even after the use of log transformations to reduce skewness on the raw-scale), (3) pdfs that are monotonically declining, rather than bell-shaped, (4) data with nonlinear responses but additive errors and (5) log-scale error terms that are heteroscedastic. We do not deal with either truncation or censoring.

3.1. Alternative data generating processes

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 number of inpatient days is 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. 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 coefficient of skewness (S) for this data generating mechanism is

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

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 coefficient of skewness of $\exp(\varepsilon)$ varied from 2.9 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, $E(x'\varepsilon) = 0$, and β_1 equals 1.0. The value for the intercept β_0 is selected so that $E(y|x) = 1$.

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

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

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

Some studies deal with dependent measures and error terms that are heavier tailed (on the log-scale) than even the log normal.[7] 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.

We also consider data generating processes 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 right. The pdf for the gamma variate y is

$$f(y) = \left(\frac{y}{b}\right)^{c-1} \frac{\exp(-y/b)}{b\Gamma(c)} \tag{11}$$

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 $E(y|x) = 1$. The shape parameter c is 0.5, 1.0, or 4.0. The first and second values of the shape parameter yield monotonically declining pdfs, 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 coefficient of skewness S on the raw-scale is a declining function of c, $S = 2c^{-0.5}$ (conditional on the covariates).

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

$$y = \exp(x\beta) + \varepsilon \tag{12}$$

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 $E(\ln(y)|x)$. Estimates based on OLS on the log-scale can provide a biased

[7] For example, the residual for Edward Norton's (personal communication) study of (log)length of stay for Medicaid psychiatric inpatient care has a log-scale coefficient of kurtosis (k) 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 inpatient costs (Meltzer et al., 2000).

assessment of the impact of the covariate x on $E(y|x)$; see Manning (1998) for a discussion. In this case, the constant variance v in Eq. (10), is replaced by some log-scale variance function $v(x)$. The expectation of y on the raw-scale becomes

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

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

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

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 $E(y|x) = 1$.

Table 1 summarizes the data generating mechanisms that we consider.

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 unconditional $E(y) = 1$. Note: as the variance increases, the skewness and kurtosis of y increase.
- (2) Two alternative models with ε being heavy-tailed (coefficient of 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 coefficient of 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 unconditional $E(y) = 1$. The shape parameter c is 0.5, 1.0, or 4.0. The first and second have monotonically declining pdfs, 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 Eq. (1), 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 unconditional $E(y) = 1$.

(B) Alternative estimators and STATA 7.0 estimation commands
- (1) OLS regression for $\ln(y)$ with a homoscedastic retransformation (ln OLS-Hom) reg $\ln(y)$ x.
- (2) OLS regression for $\ln(y)$ with a heteroscedastic retransformation (ln OLS-Het) reg $\ln(y)$ x.
- (3) GLM for y with a log link, with a variance proportional to the $E(y|x)$:
  a Poisson regression with over dispersion (Poisson):
  glm y x, link(log) family(Poisson).
- (4) GLM for y with a log link, with a standard deviation proportional to the $E(y|x)$:
  a gamma regression (gamma):
  glm y x, link(log) family(gamma).
- (5) Nonlinear least-squares by GLM for y with a log link, and an additive homoscedastic error term (NLS):
  glm y x, link(log) family(Gaussian).

Except for the heteroscedastic case with standard deviation $= 1 + x$, the covariate list includes an intercept and a single covariate x. "ln(y)" stands for the name of the log-scale variable, "y" is the name of the raw-scale variable and "x" stands for the list of covariates.

3.2. 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 $E(y|x)$ or x. This is basically the nonlinear least-squares (NLS) estimator proposed by Mullahy (1998). The second GLM estimator assumes that the raw-scale variance is proportional to the $E(y|x)$, which is a Poisson-like assumption with overdispersion, but without the discrete nature of the usual Poisson variate. The third GLM estimator assumes that the raw-scale standard deviation is proportional to $E(y|x)$, which is a gamma-like assumption similar to the model used by Blough et al. (1999). In all three GLM models,

$$E(y|x) = \exp(\beta_0 + \beta_1 x) \tag{15}$$

We do not include any of the maximum likelihood estimators in our study. In practice, the analyst may not know which distribution function to employ in an MLE model. Mis-specification of the likelihood function can lead to inconsistent estimates of either the parameters of interest (the β's) or the associated inference statistics. Using the quasi-likelihood approach for GLM only requires that the mean function be correctly specified to obtain consistent estimates. Incorrectly specifying the variance function or the distribution function leads to efficiency losses. The inferences can be corrected using robust (sandwich) estimators for the variance–covariance matrix. Thus, the quasi-likelihood approach protects against some of the problems that can arise from a mis-specified distribution function.

Gourieroux et al. (1984) demonstrate how "pseudo-maximum likelihood estimators" of parametric models having finite variances will in general be consistent so long as the first-order conditional moments (i.e. conditional means) are correctly parameterized. The examples they use are from linear exponential families, focusing specifically on Poisson-type exponential conditional mean specifications that may be embedded in overdispersed Poisson models. The fundamental notion here is that even if a log likelihood function is per se mis-specified, so long as its corresponding score equations have zero expectation under the true data generating process, the resultant parameter estimates will be consistent and asymptotically normal; this is essentially the same line of reasoning that is the basis of the consistency and asymptotic normality results for GLM estimators. The quasi-generalized pseudo-maximum likelihood approach suggested by Gourieroux et al., which affords efficiency enhancements by utilizing second-moment information, is analogous to the quasi-likelihood approach that is the basis of the efficiency improvements offered by GLM.

Because the OLS estimates are for the $E(\ln(y)|x)$, we *retransform* the log-scale estimates to obtain raw-scale estimates of $E(y|x)$. 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 $E(y|x)$. The smearing estimator for $E(\exp(\varepsilon))$ is the average of the exponentiated (log-scale) residuals from the $\ln(y)$ regression.[8] If the log-scale errors are not heteroscedastic in some function of x or of $E(y|x)$, then the smearing estimate provides a consistent estimate of $E(\exp(\varepsilon))$. 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 error variance.

We also generate predictions based on heteroscedastic retransformation as follows:

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

When the variance is $1 + x$, we omitted the x-squared term from a regression of the 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:

$$\text{glm}\, x\, y, \text{link}(\log)\, \text{family}(\cdot)$$

where the dot represents Gaussian, Poisson, or gamma.[9]

3.3. Design and evaluation

Each model is evaluated on 1000 random samples, with each 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 and 95% simulation 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 residual error on the raw-scale on the estimation sample replicate. For each replicate r, the

$$\text{MSE} = \frac{1}{N}\sum (y_{ri} - \hat{y}_{ri})^2$$

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. A more precise estimator should be closer to the true value.

[8] We did not use the normal theory retransformation from Eq. (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.

[9] In practice, we recommend the use of STATA's "xtgee" or "glm" command with the robust option, because they accommodate estimation of the robust covariance matrix (the GLM analog of the Huber/White corrected estimate for OLS), while the older versions of GLM do not.

If a model has low MSE and high APE, then there is strong evidence that that estimator has overfitted the estimation sample. The 95% simulation intervals are based on the 0.025 and 0.975th percentiles of the estimates, rather than using the normal theory estimate derived from the standard deviation of the estimates across replicates. [10] 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 $\hat{p}$ where 'A' is lower than 'B' should be 0.5 under the null that the two estimators are equally good, and the variance of $\hat{p}$ is $p(1-p)/\sqrt{n}$.

3.4. Diagnostics for variance functions (Park tests)

The results below will 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 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

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

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

$$m(y_i, x_1; \beta, \alpha, \lambda) = [(y_i - \exp(x_i\beta))^2 - \alpha \exp(\lambda x_i \beta)] \tag{18}$$

such that $E[m(\cdot)] = 0$ 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 original 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 Eq. (17) is

$$\ln(y_i - \hat{y}_i)^2 = \lambda_0 + \lambda_1 \ln(\hat{y}_i) + v_i \tag{19}$$

where $\hat{y}_i = \exp(x_i\hat{\beta})$ from one of the GLM specifications, or $\exp(x_i\hat{\beta} + 0.5\hat{\sigma}^2(x))$ from the log normal specifications. The estimate of 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. [11]

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

[10] Not all of the estimated β's from our simulations had distributions that were well approximated by a normal distribution. To avoid biased comparisons, we relied on non-parametric estimates of the 95% simulation intervals.

[11] The modified version of the Park test can also be estimated as a GLM with log link where the dependent variable is $(y_i - \hat{y}_i)^2$ and the explanatory variable is $x\hat{\beta}$ from the initial GLM of y on x. This version requires the use of a robust variance covariance matrix for $\hat{\lambda}$ to yield consistent inferences.

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 proportional to $[\exp(x_i\hat{\beta})]^{\hat{\lambda}}$. Blough et al. (1999) provides an alternative but related test specifically for the gamma alternative.

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.

4. Results: simulations and empirical examples

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.

Table 2
Sample statistics for distributions[a]

	log variance ε	Mean	S.D.	Skewness
Log normal models	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[b]	1	16.44	126.44
	Mixed normal 2[c]	1	17.13	174.63
Heteroscedastic in *x* on log-scale	log variance ε			
	Linear in *x*	1	2.27	18.79
	Quadratic in *x*	1	5.11	55.17

[a] Estimates averaged over *x* with *x* uniform (0, 1).
[b] Kurtosis in ε = 4.0 on log-scale.
[c] 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	S.E.	95% simulation interval	
				Lower	Upper
log normal (variance = 0.5)	ln OLS-Hom	1.0001	0.0237	0.9549	1.0457
	ln OLS-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 (variance = 1.0)	ln OLS-Hom	0.9996	0.0348	0.9322	1.0716
	ln OLS-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 (variance = 1.5)	ln OLS-Hom	1.0006	0.0428	0.9223	1.0891
	ln OLS-Het	1.0024	0.0567	0.8906	1.1125
	NLS	1.0025	0.0733	0.8557	1.1468
	Poisson	1.0021	0.0670	0.8671	1.1335
	Gamma	1.0025	0.0651	0.8740	1.1243
	True	1.0			
log normal (variance = 2.0)	ln OLS-Hom	0.9991	0.0484	0.9035	1.0945
	ln OLS-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			

4.1. 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 $\ln(E(y|x))$ 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, [12] the NLS, Poisson-like, and gamma models all produce consistent estimates of the slope β_1.

Thus, if consistency is 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 error variance increases. At a variance of 0.5 on the

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

Table 4
Effect of heavy tails on log-scale coefficient on slope of $\ln(E(y|x))$

Generating mechanism	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
log normal (variance = 1.0; $k = 3$)	ln OLS-Hom	0.9996	0.0348	0.9322	1.0716
	ln OLS-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	0.9981	0.0447	0.9066	1.0859
	True	1.0			
Heavy-tailed ($k = 4$)	ln OLS-Hom	1.0002	0.0375	0.9274	1.0727
	ln OLS-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 ($k = 5$)	ln OLS-Hom	1.0001	0.0396	0.9235	1.0765
	ln OLS-Het	0.9992	0.0593	0.8791	1.1176
	NLS	1.2787	5.0327	0.4137	1.9416
	Poisson	1.0109	0.3326	0.4566	1.6776
	Gamma	1.0099	0.2951	0.4344	1.5629
	True	1.0			

log-scale, the gamma standard error is roughly 13% larger, and it would take a sample size 28% [$0.28 = (1.133^2 - 1)$] 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% 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 if the underlying model is truly log normal with constant (log-scale) error variance.

4.2. 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% 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 5.0. [13]

[13] 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 by far 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 the coefficient of kurtosis of the log-scale error.

4.3. 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% larger if the shape $c = 4.0$.

Thus, the efficiency losses (relative to the gamma-based GLM 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.

Table 5
Effect of shape coefficient on slope of $\ln(E(y|x))$

Generating mechanism (gamma)	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
Shape = 0.5	ln OLS-Hom	1.0007	0.0748	0.8528	1.1490
	ln OLS-Het	0.9976	0.1733	0.6609	1.3369
	NLS	0.9996	0.0526	0.8940	1.1111
	Poisson	0.9997	0.0482	0.9044	1.0997
	Gamma	1.0001	0.0473	0.9108	1.0965
	True	1.0			
Shape = 1.0	ln OLS-Hom	0.9999	0.0449	0.9128	1.0898
	ln OLS-Het	0.9984	0.0502	0.8991	1.0954
	NLS	0.9987	0.0386	0.9243	1.0739
	Poisson	0.9988	0.0353	0.9293	1.0667
	Gamma	0.9990	0.0342	0.9300	1.0679
	True	1.0			
Shape = 4.0	ln OLS-Hom	1.0002	0.0187	0.9657	1.0386
	ln OLS-Het	1.0005	0.0176	0.9664	1.0348
	NLS	1.0005	0.0203	0.9636	1.0417
	Poisson	1.0004	0.0183	0.9665	1.0373
	Gamma	1.0004	0.0175	0.9674	1.0353
	True	1.0			

4.4. 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 statistically significantly biased estimate of the true value, but one that is not appreciably biased — the bias is only on the order of 4%. 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 6.

4.5. Heteroscedasticity

As the earlier discussion indicated, heteroscedasticity that depends on x can lead to biased estimates of the impact of x on the $E(y|x)$ if OLS is used on $\ln(y)$ without an appropriate heteroscedastic retransformation. Table 7 indicates that GLM models capture consistently the effect of x on $\ln(E(y|x))$ 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 $\ln(E(y|x))$. 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% larger to give the same precision as the heteroscedastic retransformation version of OLS, and the NLS would require a sample 250% 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 $\partial \ln(E(y|x))/\partial x$ will be biased. If the square of x is added to the list of regressors, [14] 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% simulation 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% simulation interval is [0.002, 0.97]. [15]

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

[15] 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 $\ln(E(y|x))$ 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.

Table 6
Simulation results[a]

Generating mechanism	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
Simulation results for coefficient on slope of $\ln(E(y\|x))$					
log normal (variance = 0.5)	ln OLS-Hom	1.0001	0.0238	0.9550	1.0458
	ln OLS-Het	1.0000	0.0264	0.9492	1.0538
	NLS	0.9999	0.0299	0.9407	1.0618
	Poisson	0.9999	0.0274	0.9462	1.0572
	Gamma	1.0000	0.0269	0.9477	1.0553
log normal (variance = 1.0)	ln OLS-Hom	0.9996	0.0348	0.9322	1.0716
	ln OLS-Het	0.9986	0.0418	0.9158	1.0829
	NLS	0.9980	0.0505	0.9034	1.0953
	Poisson	0.9979	0.0462	0.9057	1.0867
	Gamma	0.9981	0.0448	0.9066	1.0859
log normal (variance = 1.5)	ln OLS-Hom	1.0006	0.0427	0.9224	1.0891
	ln OLS-Het	1.0024	0.0567	0.8907	1.1126
	NLS	1.0025	0.0733	0.8558	1.1469
	Poisson	1.0021	0.0670	0.8671	1.1335
	Gamma	1.0025	0.0651	0.8741	1.1243
log normal (variance = 2.0)	ln OLS-Hom	0.9991	0.0484	0.9036	1.0946
	ln OLS-Het	1.0013	0.0685	0.8640	1.1284
	NLS	1.0051	0.0983	0.8224	1.2140
	Poisson	1.0027	0.0882	0.8353	1.1740
	Gamma	1.0014	0.0845	0.8371	1.1735
Heavy-tailed ($k = 4$)	ln OLS-Hom	1.0002	0.0376	0.9274	1.0727
	ln OLS-Het	0.9995	0.0511	0.8973	1.1019
	NLS	1.0084	0.1738	0.7387	1.3422
	Poisson	1.0039	0.1427	0.7628	1.2884
	Gamma	1.0036	0.1320	0.7679	1.2545
Heavy-tailed ($k = 5$)	ln OLS-Hom	1.0003	0.0396	0.9235	1.0766
	ln OLS-Het	0.9992	0.0593	0.8791	1.1177
	NLS	1.2787	5.0328	0.4138	1.9416
	Poisson	1.0109	0.3326	0.4567	1.6777
	Gamma	1.0099	0.2952	0.4345	1.5630
Simulation results for β_1					
Gamma (shape = 0.5)	ln OLS-Hom	1.0007	0.0748	0.8528	1.1490
	ln OLS-Het	0.9976	0.1734	0.6609	1.3369
	NLS	0.9996	0.0526	0.8940	1.1111
	Poisson	0.9997	0.0482	0.9044	1.0997
	Gamma	1.0001	0.0473	0.9108	1.0965
Gamma (shape = 1.0)	ln OLS-Hom	0.9999	0.0449	0.9128	1.0898
	ln OLS-Het	0.9984	0.0502	0.8991	1.0954
	NLS	0.9987	0.0386	0.9243	1.0739
	Poisson	0.9988	0.0353	0.9293	1.0667
	Gamma	0.9990	0.0342	0.9300	1.0679

Table 6 (*Continued*)

Generating mechanism	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
Gamma (shape = 4.0)	ln OLS-Hom	1.0002	0.0187	0.9657	1.0386
	ln OLS-Het	1.0005	0.0176	0.9664	1.0348
	NLS	1.0005	0.0203	0.9636	1.0417
	Poisson	1.0004	0.0183	0.9665	1.0373
	Gamma	1.0004	0.0175	0.9674	1.0353
NLS additive error	ln OLS-Hom	1.0432	0.0081	1.0272	1.0585
	ln OLS-Het	0.9935	0.0074	0.9790	1.0077
	NLS	1.0001	0.0061	0.9885	1.0122
	Poisson	1.0001	0.0065	0.9874	1.0126
	Gamma	1.0001	0.0072	0.9857	1.0137
Heteroscedasticity (variance = $1 + x$)	ln OLS-Hom	1.0001	0.0408	0.9190	1.0800
	ln OLS-Het	1.4992	0.0546	1.3916	1.6085
	NLS	1.4998	0.1025	1.3049	1.7220
	Poisson	1.4982	0.0784	1.3417	1.6637
	Gamma	1.4984	0.0662	1.3692	1.6392
Heteroscedasticity (S.D. = $1 + x$)	ln OLS-Hom	0.9997	0.0540	0.8967	1.1152
	ln OLS-Het	2.4942	0.0826	2.3248	2.6580
	NLS	2.2773	2.4384	0.5462	4.0698
	Poisson	2.2709	0.3723	1.4848	2.9825
	Gamma	2.2562	0.1940	1.8550	2.6157

[a] "Mean" evaluated at $x = 0.50$ for log normal model with heteroscedasticity S.D. = $1 + x$.

As in 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. As expected for the cases that were not truly heteroscedastic, the heteroscedastic retransformation method yields less precise estimates than the homoscedastic version. Except for the cases that were truly heteroscedastic, 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. Tables 6 and 8 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 or distribution function can lead to a substantial loss in precision.

Table 7
Effect of heteroscedasticity on the log-scale on slope of $\ln(E(y|x))$[a]

Generating mechanism	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
log normal (variance = 1.0)	ln OLS-Hom	0.9996	0.0348	0.9322	1.0716
	ln OLS-Het	0.9986	0.0418	0.9158	1.0825
	NLS	0.9980	0.0505	0.9034	1.0953
	Poisson	0.9979	0.0462	0.9057	1.0867
	Gamma	0.9981	0.0448	0.9066	1.0859
	True	1.0			
Heteroscedasticity (variance = $1 + x$)	ln OLS-Hom	1.0001	0.0408	0.9190	1.0800
	ln OLS-Het	1.4992	0.0546	1.3916	1.6085
	NLS	1.4998	0.1025	1.3049	1.7220
	Poisson	1.4982	0.0784	1.3417	1.6637
	Gamma	1.4984	0.0662	1.3692	1.6392
	True	1.5			
Heteroscedasticity (S.D. = $1 + x$)	ln OLS-Hom	0.9997	0.0540	0.8967	1.1152
	ln OLS-Het	2.4942	0.0826	2.3248	2.6580
	NLS	2.2773	2.4384	0.5462	4.0698
	Poisson	2.2709	0.3723	1.4848	2.9825
	Gamma	2.2562	0.1940	1.8550	2.6157
	True	2.5			

[a] For the log normal case where the standard deviation of ε is $1 + x$, the slope is evaluated at $x = 0.5$.

4.6. *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 with y as the dependent measure would overfit the extreme cases. That is, the estimate of the β's would be overly influenced by extreme cases and not reflect well the true values β_0 and β_1. Some of the generalized linear models, especially the NLS, may exhibit a similar propensity to overfit extreme cases, because they will not 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.[16] If overfitting 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 compare alternative estimators in terms of which model had lower average MSEs or lower APEs; see Table 9. For the within-sample measure of MSE, NLS generally has smaller MSEs 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 retransformed OLS model is closer to true, followed by the

[16] APE = absolute(estimated β_1 − true β_1) averaged over replications.

Table 8
Efficiency effects coefficient on slope of $\ln(E(y|x))$

Generating mechanism	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
log normal (variance = 1.0)	ln OLS-Hom	0.9996	0.0348	0.9322	1.0716
	ln OLS-Het	0.9986	0.0418	0.9158	1.0825
	NLS	0.9980	0.0505	0.9034	1.0953
	Poisson	0.9979	0.0462	0.9057	1.0867
	Gamma	0.9981	0.0448	0.9066	1.0859
	True	1.0			
Gamma (shape = 1.0)	ln OLS-Hom	0.9999	0.0449	0.9128	1.0898
	ln OLS-Het	0.9984	0.0502	0.8991	1.0954
	NLS	0.9987	0.0386	0.9243	1.0739
	Poisson	0.9988	0.0353	0.9293	1.0667
	Gamma	0.9990	0.0342	0.9300	1.0679
	True	1.0			
NLS additive error	ln OLS-Hom	1.0432	0.0081	1.0272	1.0585
	ln OLS-Het	0.9935	0.0074	0.9790	1.0077
	NLS	1.0001	0.0061	0.9885	1.0122
	Poisson	1.0001	0.0065	0.9874	1.0126
	Gamma	1.0001	0.0072	0.9857	1.0137
	True	1.0			
Heteroscedasticity (variance = $1 + x$)	ln OLS-Hom	1.0001	0.0408	0.9190	1.0800
	ln OLS-Het	1.4992	0.0546	1.3916	1.6085
	NLS	1.5000	0.1025	1.3050	1.7220
	Poisson	1.4982	0.0784	1.3417	1.6637
	Gamma	1.4984	0.0662	1.3692	1.6392
	True	1.50			

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 any heteroscedasticity, but the best of all (on APE grounds) if there is no heteroscedasticity.

In any event, the within-sample measure of goodness-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.

4.7. 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

Table 9
Mean squared error (MSE) and absolute prediction error (APE)

Generating mechanism	Estimators A	Compared B	Number of replications[a]	
			MSE(A) < MSE(B)	APE(A) < APE(B)
log normal (variance = 0.5)	ln OLS-Hom	ln OLS-Het	915	576
	ln OLS-Hom	NLS	0	614
	ln OLS-Hom	Poisson	145	580
	ln OLS-Hom	Gamma	356	578
	ln OLS-Het	ln OLS-Hom	85	424
	ln OLS-Het	NLS	0	576
	ln OLS-Het	Poisson	0	550
	ln OLS-Het	Gamma	0	537
	NLS	ln OLS-Hom	997	386
	NLS	ln OLS-Het	1000	424
	NLS	Poisson	927	395
	NLS	Gamma	977	432
	Poisson	ln OLS-Hom	848	420
	Poisson	ln OLS-Het	1000	450
	Poisson	Gamma	942	477
	Gamma	ln OLS-Hom	638	422
	Gamma	ln OLS-Het	1000	463
	Gamma	NLS	0	568
	Gamma	Poisson	28	523
log normal (variance = 1.0)	ln OLS-Hom	ln OLS-Het	837	577
	ln OLS-Hom	NLS	0	671
	ln OLS-Hom	Poisson	138	636
	ln OLS-Hom	Gamma	289	615
	ln OLS-Het	ln OLS-Hom	163	423
	ln OLS-Het	NLS	0	611
	ln OLS-Het	Poisson	0	596
	ln OLS-Het	Gamma	1	566
	NLS	ln OLS-Hom	999	329
	NLS	ln OLS-Het	1000	389
	NLS	Poisson	934	385
	NLS	Gamma	985	415
	Poisson	ln OLS-Hom	856	364
	Poisson	ln OLS-Het	1000	404
	Poisson	Gamma	941	438
	Gamma	ln OLS-Hom	710	385
	Gamma	ln OLS-Het	999	434
	Gamma	NLS	0	585
	Gamma	Poisson	36	562
log normal (variance = 1.5)	ln OLS-Hom	ln OLS-Het	829	638
	ln OLS-Hom	NLS	0	691
	ln OLS-Hom	Poisson	128	666
	ln OLS-Hom	Gamma	258	669
	ln OLS-Het	ln OLS-Hom	171	362
	ln OLS-Het	NLS	0	609
	ln OLS-Het	Poisson	0	599
	ln OLS-Het	Gamma	7	592

Table 9 (*Continued*)

Generating mechanism	Estimators A	Compared B	Number of replications[a]	
			MSE(A) < MSE(B)	APE(A) < APE(B)
	NLS	ln OLS-Hom	999	309
	NLS	ln OLS-Het	1000	391
	NLS	Poisson	936	396
	NLS	Gamma	972	433
	Poisson	ln OLS-Hom	869	334
	Poisson	ln OLS-Het	1000	401
	Poisson	Gamma	935	479
	Gamma	ln OLS-Hom	742	331
	Gamma	ln OLS-Het	993	408
	Gamma	NLS	2	567
	Gamma	Poisson	34	521
log normal (variance = 2.0)	ln OLS-Hom	ln OLS-Het	813	637
	ln OLS-Hom	NLS	0	704
	ln OLS-Hom	Poisson	116	705
	ln OLS-Hom	Gamma	261	684
	ln OLS-Het	ln OLS-Hom	187	363
	ln OLS-Het	NLS	0	625
	ln OLS-Het	Poisson	0	593
	ln OLS-Het	Gamma	3	591
	NLS	ln OLS-Hom	998	296
	NLS	ln OLS-Het	1000	375
	NLS	Poisson	955	400
	NLS	Gamma	984	441
	Poisson	ln OLS-Hom	875	295
	Poisson	ln OLS-Het	1000	407
	Poisson	Gamma	941	471
	Gamma	ln OLS-Hom	737	316
	Gamma	ln OLS-Het	997	409
	Gamma	NLS	0	559
	Gamma	Poisson	33	529
Heavy-tailed ($k = 4$)	ln OLS-Hom	ln OLS-Het	824	640
	ln OLS-Hom	NLS	0	818
	ln OLS-Hom	Poisson	105	823
	ln OLS-Hom	Gamma	258	822
	ln OLS-Het	ln OLS-Hom	176	360
	ln OLS-Het	NLS	0	766
	ln OLS-Het	Poisson	0	778
	ln OLS-Het	Gamma	10	786
	NLS	ln OLS-Hom	998	182
	NLS	ln OLS-Het	1000	234
	NLS	Poisson	952	382
	NLS	Gamma	985	418
	Poisson	ln OLS-Hom	891	177
	Poisson	ln OLS-Het	1000	222
	Poisson	Gamma	950	459
	Gamma	ln OLS-Hom	739	178
	Gamma	ln OLS-Het	990	214

Table 9 (*Continued*)

Generating mechanism	Estimators A	Compared B	Number of replications[a]	
			MSE(A) < MSE(B)	APE(A) < APE(B)
	Gamma	NLS	1	582
	Gamma	Poisson	37	541
Heavy-tailed ($k = 5$)	ln OLS-Hom	ln OLS-Het	797	669
	ln OLS-Hom	NLS	0	900
	ln OLS-Hom	Poisson	107	907
	ln OLS-Hom	Gamma	237	900
	ln OLS-Het	ln OLS-Hom	203	331
	ln OLS-Het	NLS	0	863
	ln OLS-Het	Poisson	0	862
	ln OLS-Het	Gamma	10	855
	NLS	ln OLS-Hom	1000	100
	NLS	ln OLS-Het	1000	137
	NLS	Poisson	950	382
	NLS	Gamma	986	420
	Poisson	ln OLS-Hom	889	93
	Poisson	ln OLS-Het	1000	138
	Poisson	Gamma	944	457
	Gamma	ln OLS-Hom	761	100
	Gamma	ln OLS-Het	990	145
	Gamma	NLS	4	580
	Gamma	Poisson	35	543
Gamma (shape = 0.5)	ln OLS-Hom	ln OLS-Het	1000	759
	ln OLS-Hom	NLS	0	379
	ln OLS-Hom	Poisson	96	340
	ln OLS-Hom	Gamma	199	326
	ln OLS-Het	ln OLS-Hom	0	241
	ln OLS-Het	NLS	0	183
	ln OLS-Het	Poisson	0	173
	ln OLS-Het	Gamma	0	176
	NLS	ln OLS-Hom	1000	621
	NLS	ln OLS-Het	1000	817
	NLS	Poisson	931	395
	NLS	Gamma	965	433
	Poisson	ln OLS-Hom	902	660
	Poisson	ln OLS-Het	1000	827
	Poisson	Gamma	933	476
	Gamma	ln OLS-Hom	801	674
	Gamma	ln OLS-Het	1000	824
	Gamma	NLS	0	567
	Gamma	Poisson	30	524
Gamma (shape = 1.0)	ln OLS-Hom	ln OLS-Het	1000	532
	ln OLS-Hom	NLS	0	434
	ln OLS-Hom	Poisson	105	380
	ln OLS-Hom	Gamma	255	364
	ln OLS-Het	ln OLS-Hom	0	468
	ln OLS-Het	NLS	0	405

Table 9 (*Continued*)

Generating mechanism	Estimators A	Compared B	Number of replications[a]	
			MSE(A) < MSE(B)	APE(A) < APE(B)
	ln OLS-Het	Poisson	0	358
	ln OLS-Het	Gamma	0	340
	NLS	ln OLS-Hom	995	566
	NLS	ln OLS-Het	1000	595
	NLS	Poisson	927	406
	NLS	Gamma	965	435
	Poisson	ln OLS-Hom	890	620
	Poisson	ln OLS-Het	1000	642
	Poisson	Gamma	922	483
	Gamma	ln OLS-Hom	740	636
	Gamma	ln OLS-Het	1000	660
	Gamma	NLS	0	565
	Gamma	Poisson	29	517
Gamma (shape = 4.0)	ln OLS-Hom	ln OLS-Het	962	453
	ln OLS-Hom	NLS	0	557
	ln OLS-Hom	Poisson	140	481
	ln OLS-Hom	Gamma	363	449
	ln OLS-Het	ln OLS-Hom	38	547
	ln OLS-Het	NLS	0	610
	ln OLS-Het	Poisson	0	559
	ln OLS-Het	Gamma	3	476
	NLS	ln OLS-Hom	996	443
	NLS	ln OLS-Het	1000	390
	NLS	Poisson	925	352
	NLS	Gamma	975	387
	Poisson	ln OLS-Hom	852	519
	Poisson	ln OLS-Het	1000	441
	Poisson	Gamma	935	423
	Gamma	ln OLS-Hom	633	551
	Gamma	ln OLS-Het	997	524
	Gamma	NLS	0	613
	Gamma	Poisson	24	577
Nonlinear additive error	ln OLS-Hom	ln OLS-Het	0	10
	ln OLS-Hom	NLS	0	1
	ln OLS-Hom	Poisson	0	1
	ln OLS-Hom	Gamma	0	2
	ln OLS-Het	ln OLS-Hom	1000	990
	ln OLS-Het	NLS	0	309
	ln OLS-Het	Poisson	0	322
	ln OLS-Het	Gamma	36	331
	NLS	ln OLS-Hom	1000	999
	NLS	ln OLS-Het	1000	691
	NLS	Poisson	952	562
	NLS	Gamma	992	617
	Poisson	ln OLS-Hom	1000	999
	Poisson	ln OLS-Het	1000	678
	Poisson	Gamma	949	648

Table 9 (*Continued*)

Generating mechanism	Estimators A	Compared B	Number of replications[a]	
			MSE(A) < MSE(B)	APE(A) < APE(B)
	Gamma	ln OLS-Hom	1000	998
	Gamma	ln OLS-Het	964	669
	Gamma	NLS	0	383
	Gamma	Poisson	33	352
log normal (variance $= 1 + x$)	ln OLS-Hom	ln OLS-Het	0	0
	ln OLS-Hom	NLS	0	0
	ln OLS-Hom	Poisson	0	0
	ln OLS-Hom	Gamma	0	0
	ln OLS-Het	ln OLS-Hom	1000	1000
	ln OLS-Het	NLS	0	735
	ln OLS-Het	Poisson	0	677
	ln OLS-Het	Gamma	12	612
	NLS	ln OLS-Hom	1000	1000
	NLS	ln OLS-Het	1000	265
	NLS	Poisson	964	276
	NLS	Gamma	989	309
	Poisson	ln OLS-Hom	1000	1000
	Poisson	ln OLS-Het	1000	323
	Poisson	Gamma	925	376
	Gamma	ln OLS-Hom	1000	1000
	Gamma	ln OLS-Het	988	388
	Gamma	NLS	0	691
	Gamma	Poisson	59	624
log normal (S.D. $= 1 + x$)	ln OLS-Hom	ln OLS-Het	0	0
	ln OLS-Hom	NLS	0	258
	ln OLS-Hom	Poisson	0	76
	ln OLS-Hom	Gamma	0	14
	ln OLS-Het	ln OLS-Hom	1000	1000
	ln OLS-Het	NLS	0	699
	ln OLS-Het	Poisson	0	550
	ln OLS-Het	Gamma	35	499
	NLS	ln OLS-Hom	1000	742
	NLS	ln OLS-Het	1000	301
	NLS	Poisson	988	230
	NLS	Gamma	999	258
	Poisson	ln OLS-Hom	1000	924
	Poisson	ln OLS-Het	1000	450
	Poisson	Gamma	926	371
	Gamma	ln OLS-Hom	1000	986
	Gamma	ln OLS-Het	965	501
	Gamma	NLS	0	742
	Gamma	Poisson	71	629

[a] Out of 1000 replicates.

the estimates from each model, we can construct raw-scale residuals for each estimator. Then we 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, as in Eq. (19).

Table 10 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

Table 10
Comparisons of alternative estimators Park tests of mean–variance relationship estimates of λ[a]

Generating mechanism	Estimator	Mean	S.E.	95% simulation interval	
				Lower	Upper
log normal (variance = 1.0)	ln-OLS-Hom	2.0005	0.0723	1.8562	2.1471
	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 (shape = 1.0)	ln-OLS-Hom	2.0057	0.0788	1.8545	2.1602
	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 ($k = 5$)	ln-OLS-Hom	1.9964	0.0655	1.8653	2.1267
	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 additive error	ln-OLS-Hom	0.02159	0.0729	−0.1289	0.1661
	ln-OLS-Het	0.00319	0.0780	−0.1536	0.1532
	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 (variance = $1 + x$)	ln-OLS-Hom	2.7611	0.0763	2.6154	2.9196
	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	???			

[a] Estimate of slope λ from log OLS version of Park test.

variable y, the estimates of λ centered quite tightly around the true value $\lambda = 2$. 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 $\lambda = 2$, 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 $\lambda = 2$, 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 (if the data exhibits heavy tails on the log-scale) 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 $\lambda = 2$, whereas the other estimators are not misled. 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 modified Park test such as Eq. (19) 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.

4.8. Empirical examples

To illustrate the empirical importance of the above issues in the context of real data, we examine two applications: one from health and one from labor. First, we estimate a set of 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 12 months prior to the survey, the number of visits is the outcome we consider here. The summary statistics for this outcome measure are mean = 6.42, median = 3, variance = 204.7, the coefficient of skewness = 9.79, and the coefficient of kurtosis = 158.6.

The model specification used here is identical to that used in Mullahy (1998), including as covariates: age in years (AGE), gender (MALE), years of completed schooling (EDUC), race (WHITE), marital status (MARRIED), and health status (EXCELLENT, VERYGOOD, GOOD). The models are estimated using λ-values of 0 (NLS class), 1 (Poisson class), and 2 (gamma 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). We include the homoscedastic version of the OLS model for $\ln(y)|x$.

The visit results are summarized in Table 11. 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 ($\lambda = 1, 2$) are not terribly inferior, while the $\lambda = 0$ case is generally much less precise than the others. No less importantly, note too that the

Table 11
GLM and log OLS estimates of NHIS doctor visit data: alternative λ-values[a]

Variable	λ				log OLS[b]
	0	1	2	1.9179	
AGE	−0.0091 (0.0019)	−0.0078 (0.0012)	−0.0068 (0.0010)	−0.0070 (0.0010)	−0.0026 (0.0005)
MALE	−0.0471 (0.0446)	−0.1059 (0.0304)	−0.1446 (0.0257)	−0.1393 (0.0256)	−0.1883 (0.0124)
EDUC	0.0372 (0.0067)	0.0308 (0.0047)	0.0239 (0.0041)	0.0239 (0.0041)	0.0245 (0.0025)
WHITE	0.1324 (0.0524)	0.1514 (0.0380)	0.1625 (0.0322)	0.1622 (0.0317)	0.1542 (0.0159)
MARRIED	−0.1478 (0.0414)	−0.1358 (0.0295)	−0.1132 (0.0263)	−0.1173 (0.0262)	−0.0406 (0.0131)
EXCELLENT	−1.6125 (0.0477)	−1.5753 (0.0410)	−1.5495 (0.0393)	−1.5412 (0.0391)	−1.3039 (0.0231)
VERYGOOD	−1.3395 (0.0449)	−1.3107 (0.0391)	−1.2909 (0.0374)	−1.2844 (0.0370)	−1.0803 (0.0229)
GOOD	−0.8561 (0.0441)	−0.8468 (0.0405)	−0.8420 (0.0396)	−0.8386 (0.0388)	−0.7515 (0.0234)

[a] Heteroscedasticity robust S.E. in parentheses.
[b] With homoscedastic retransformation.

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). The homoscedastic OLS model overstates the role of gender and understates the role of health status by moderate amounts.

Our second example examines the effect of education and race on annual earnings for those with any earnings in 1992. For this case, we use data from the 1992–1993 Panel Study on Income Dynamics (PSID). Of the 8321 adults on that file, we excluded 2120 for being too young (age < 25) or too old (age > 62). We excluded another 1463 for having no earnings or having missing earnings information. To avoid problems in separating the returns from education from the returns from other investment, we exclude another 579 individuals who were fully or partially self-employed. The analytic sample consists of 1808 men and 1810 women.

We follow the traditional specifications for earnings from the labor economics literature; we assume a proportional representation. For OLS, we examine the log of annual earnings, while for the generalized linear models, we employ a log link. Men and women are analyzed separately. Education enters as two variables: the PSID years of education (called "educ"), which is truncated at 17 years, and an indicator variable for 17 or more years of education (called "post"). Age enters as a quadratic. We include indicators for being black or other non-white racial group. To control for local market conditions, we include the area unemployment rate.

We employ the PSID's sampling weights to correct for the oversampling of low-income individuals, and robust Huber/White estimates of standard errors to correct the inferences. The generalized linear models were estimated using the procedure "rglm" from STATA 5.0.

The modified Park test indicates the raw-scale variance increases as the 2.03 power for men and 1.56 power for women. The men's estimate is significantly different from 0 or 1, but not from 2. This indicates that the gamma is the logical GLM alternative to OLS on log earnings for men. The women's estimate is significantly different from 0, 1 and 2, indicating that an iteratively reweighted nonlinear least-squares estimator may be a more efficient estimator alternative to OLS on log earnings for women. If a GLM estimator is used, it will be necessary to use a robust estimate of the variance–covariance matrix for the estimated parameters.

The log-scale residuals are too heavy-tailed for a log normal error. The coefficient of kurtosis is 15.73 for men and 6.89 for women. A conventional Park test on the log-scale residuals indicates that there is heteroscedasticity in race, but not age, education or unemployment rates. For men, blacks are much less variable than non-blacks.

Table 12 reports the coefficients, robust standard errors and *P*-values for years of education, the indicator for 17 or more years of education, and being black (relative to being white). The log OLS model overstates the effect on earnings of being black by about a third for males. The gamma model has a 23% lower standard error for education than does the log OLS model. For women, there are more modest reductions in standard errors on the education variables from using the reweighted nonlinear least-squares over either the other GLM approaches or the log OLS.

These real world examples further emphasize that the variance structure has considerable implications for the analyst's inferences in applied research.

Table 12
GLM and log OLS estimates for PSID earnings

Subsample gender[a]	Model[b]	Parameter	Estimate	S.E.	*P*-value
M	1	Educ	0.1123	0.0100	<0.001
M	2	Educ	0.1194	0.0100	<0.001
M	3	Educ	0.1175	0.0087	<0.001
M	4	Educ	0.1158	0.0077	<0.001
M	5	Educ	0.1157	0.0077	<0.001
F	1	Educ	0.1230	0.0166	<0.001
F	2	Educ	0.1289	0.0168	<0.001
F	3	Educ	0.1295	0.0151	<0.001
F	4	Educ	0.1303	0.0137	<0.001
F	5	Educ	0.1302	0.0143	<0.001
M	1	Post	0.0125	0.0708	0.8601
M	2	Post	0.0368	0.0767	0.6318
M	3	Post	0.0413	0.0737	0.5753
M	4	Post	0.0452	0.0708	0.5236
M	5	Post	0.0453	0.0707	0.5218
M	1	Post	0.1113	0.1069	0.2980
F	2	Post	0.0163	0.0978	0.8675
F	3	Post	0.0178	0.0928	0.8481
F	4	Post	0.0179	0.0887	0.8401
F	5	Post	0.0173	0.0904	0.8484
M	1	Black	−0.3438	0.0692	<0.001
M	2	Black	−0.2506	0.0566	<0.001
M	3	Black	−0.2479	0.0503	<0.001
M	4	Black	−0.2462	0.0464	<0.001
M	5	Black	−0.2462	0.0463	<0.001
F	1	Black	0.0369	0.0676	0.5854
F	2	Black	−0.0364	0.0505	0.4713
F	3	Black	−0.0214	0.0510	0.6754
F	4	Black	−0.0148	0.0530	0.7803
F	5	Black	−0.0165	0.0521	0.7513

[a] M: male; F: female.

[b] 1: OLS for log earnings with homoscedastic retransformation; 2: nonlinear least-squares (NLS) with log link; 3: Poisson regression with log link; 4: gamma regression with log link; 5: iteratively reweighted NLS based on modified Park test (one iteration).

5. Conclusions and an algorithm for choosing an estimator

Our results indicate that the choice of estimator for examining the $\ln(E(y|x))$ can have major implications for the empirical results if the estimator is not designed to deal with the specific data generating mechanism. Garden-variety distributional 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 appreciably biased. Moreover, when the pdf is not bell-shaped or a skewed bell-shaped, then the OLS-based models are notably less precise than some of the GLM alternatives, but remain consistent when the errors are homoscedastic.

The bias in the homoscedastic version of OLS can be corrected if one estimates the variance function for the log-scale error ε, and then uses that to obtain a retransformed prediction of $\ln(E(y|x))$; see Manning (1998). Although consistent, this approach can be quite 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 can be difficult for heteroscedasticity across multiple factors or continuous variables.

The GLM models, such as the NLS, Poisson-like, and gamma models, provide the alternative of directly estimating what most economists are really after, $E(y|x)$ or $\ln(E(y|x))$, without having to go through the process of estimating the variance function $v(x)$ for the log-scale error that is required for retransforming log OLS results. Unfortunately, if the true model is a heteroscedastic equation for $\ln(y)|x$, 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 have concentrated most of our attention on data generating mechanisms based on the log normal and on the gamma. [17] 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 $\partial \ln(E(y|x))/\partial x$, then we would recommend the GLM models. These estimators also would be easier to use than the heteroscedastic retransformation discussed by Manning (1998) and used in many of the papers from the Health Insurance Experiment (Manning et al., 1987a,b; Newhouse et al., 1993). However, we are often quite concerned about precision; a common difficulty that most health economists face is lack of precision due to the high variance in utilization and expenditures. We often need to find the most precise, as well as a consistent estimate of the response. Depending on the data, some GLM methods will be more precise than others. Unfortunately, we cannot rely on within-sample diagnostics to make the choice,

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

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 GLM estimators. If the log-scale residuals are heavy-tailed (the coefficient of 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 (coefficient of kurtosis about 3 or less), 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.[18] 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 approach outlined above. If they are leptokurtotic (coefficient of kurtosis > 3), then the GLM models considered may be quite problematic.[19] If the log-scale residuals are both leptokurtotic and heteroscedastic, then the performance of GLM and log OLS with heteroscedastic retransformation should be compared to see which is more precise and whether there is any residual bias in the OLS-based predictions.

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.

[18] If the log-scale residuals are symmetric, consider the log normal. A test for skewness is: $n(S^2)/6$, where S is the coefficient of 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 coefficients of skewness and kurtosis (k) 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 $c = 1$ 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.

[19] 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.

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.

While we have considered results here for the case where y is strictly positive (e.g. part 2 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 and may include cases where $y = 0$. 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 logit/OLS-based two-part model (Duan et al., 1983; Manning et al., 1987a), the logit/gamma alternative (Blough et al., 1999) or some non-linear least-squares alternative (Mullahy, 1998) is a subject for further research.

In this paper, we have focused on approaches that could be applied to either continuous or count data. Omitted from our estimators are the estimators for variants of the negative binomial distribution for count data, or the survival methods for duration data. Assessing the relative merits of these alternatives is clearly needed.

Acknowledgements

This research was supported in part by a grant from the National Institute of Alcohol Abuse and Alcoholism (NIAAA) under Grants AA10392 and AA10393. The opinions expressed are those of the authors, and not those of NIAAA, the National Bureau of Economic Research, the University of Chicago, or the University of Wisconsin. We would like to thank NIAAA and Janssen Pharmaceutica for their support of this research. We would like to thank Ashoke Bhattacharjya, Partha Deb, Tom DeLeire, Edward Norton, Daniel Polsky, Paul Rathouz, and an anonymous reviewer for their comments on earlier drafts.

References

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

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

D'Agostino, R.B., Pearson, E.S., 1973. Tests for departure from normality: empirical results for the distribution of b_2 and $\sqrt{b_1}$. Biometrika 60, 613–622.

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

Duan, N., Manning, W.G., et al., 1983. A comparison of alternative models for the demand for medical care. Journal of Business and Economics Statistics 1, 115–126.

Gourieroux, C., Montfort, A., Trognon, A., 1984. Pseudo-maximum likelihood methods: applications to Poisson models. Econometrica 52, 701–720.

Jones, A., 2000. Health econometrics. In: Culyer, A., Newhouse, J. (Eds.), Handbook of Health Economics. Elsevier, Amsterdam.

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

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

Manning, W.G., Duan, N., Rogers, W.H., 1987a. Monte Carlo evidence on the choice between sample selection and two-part models. Journal of Econometrics 35, 59–82.

Manning, W.G., Newhouse, J.P., et al., 1987b. Health insurance and the demand for medical care: evidence from a randomized experiment. American Economic Review 77 (3), 251–277.

Meltzer, D.O., Manning, W.G., Morrison, J., Guth, T., Hernanadez, A., Dhar, A., Jin, L., Levinson, W., 2000. Effects of hospitalist physicians on an academic general medicine service: results of a randomized trial. Draft.

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. In: The Results of the Health Insurance Experiment. Harvard University Press, Cambridge.

Park, R., 1966. Estimation with heteroscedastic error terms. Econometrica 34, 888.

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.