ELSEVIER Journal of Health Economics 17 (1998) 283–295

JOURNAL OF HEALTH ECONOMICS

The logged dependent variable, heteroscedasticity, and the retransformation problem

Willard G. Manning *

Department of Health Studies, The University of Chicago, Chicago, IL, USA

Received 1 December 1996; revised 1 July 1997; accepted 1 September 1997

Keywords: Heteroscedasticity; Retransformation; Logged dependent variable

1. Introduction

The use of a log transformed dependent variable has become commonplace in applied microeconomic work. In some cases, such as in the analysis of wages, the log transform has become the standard, while in other applied areas it is just considered good practice. Once the estimates from such a model have been obtained, the usual practice is to interpret the response to a particular variable (e.g., price or income) as being the exponential of the coefficient of that variable in the model. In some cases, these estimates of the impact of the variable are corrected for the fact that one is using an estimate, rather than the true value of the coefficient (Kennedy, 1981, 1983, 1992). [1] Very few of the applications make a full correction for the impact of heteroscedasticity on the estimated response (for examples, see Duan et al., 1983; Manning et al., 1987; McCuen et al., 1990;

[*] Corresponding author.

[1] If m is the estimate of the mean μ of log(y), then $E(\exp(\mu)) = \exp(m - 0.5\ \mathrm{Var}(m))$, where exp() is the exponential function. This correction for using the estimate m of μ may be important for small sample sizes, but becomes second order small for large ones.

PII S0167-6296(98)00025-3

Newhouse et al., 1981; Puma and Hoaglin, 1990; Showalter, 1994). Although many analysts will use either a generalized least squares estimator or the Huber/White consistent estimate of the variance–covariance matrix of the estimated coefficients, few make a direct adjustment to the predicted response. Unlike regression models on the raw, or untransformed scale, log model results are about geometric means, not arithmetic means. If an unlogged dependent variable is used, the estimated response is that for the arithmetic mean. In fact, there is a danger that log scale results may provide a very misleading, incomplete, and a biased estimate of the impact of a covariate on the arithmetic mean. If one wants to comment on the arithmetic mean response to some variable from a model with a logged dependent variable, then one must include a term that captures any heteroscedasticity in the error term on the log scale that is attributable to that variable.

The following sections will explore the role of heteroscedasticity in log models. After considering rationales for the log transformation in Section 2, I examine the effect of heteroscedasticity in log models with a normally distributed error term in Section 3. This is first done with a simple comparison of population means. Then, the model is extended to allow for other covariates and for a non-normal error term. In Section 4, the case of a non-normal error term is considered. In Section 5, the model is relaxed to allow for other power (or Box–Cox) transformations of the dependent variable. Section 6 deals with other estimators where retransformation is problematic, with or without heteroscedasticity. Section 8 illustrates the impact of this phenomenon using an example from the Health Insurance Experiment.

2. Rationales

The rationale for using a log transformed dependent measure can come from a variety of concerns: (1) a desire for multiplicative or proportional responses to a covariate of interest; (2) a desire to generate an estimate that easily yields an elasticity (as in the case of the log–log model); (3) as a consequence of working from certain classes of utility, demand, production, or cost functions (as in the cases of the Cobb–Douglas and translog formulations); (4) as a consequence of estimating the log of the odds ratio for grouped data from a logit model; or (5) a need to deal with dependent variables that are badly skewed to the right.

One widely used rationale for the log transform derives from the single parameter Box–Cox model, where the dependent variable is subjected to a power transformation in order to deal with the issues of skewed or non-normal data (Box and Cox, 1964). Specifically, one can select a power transform λ such that:

$$\frac{(y^{\lambda}-1)}{\lambda}=x\beta+\epsilon \text{ if } \lambda\neq 0 \tag{1a}$$

$$\ln(y)=x\beta+\epsilon \text{ if } \lambda=0 \tag{1b}$$

where y is the original (or untransformed) dependent variable, x is a row vector of covariates, ϵ is an additive error term that is independent of the covariates x, and β and λ are parameters to be estimated. If under a suitable transformation, the error term is normally distributed, then the model can be estimated by maximum likelihood. If the error term is not normally distributed, then one can chose λ to yield a more symmetric error, while retaining the linear response function. For example, in many analyzes of expenditures on health care, the expenditures for users are subject to a log transform to reduce, if not eliminate, the skewness inherent in health expenditure data; for an example with medical expenditures, see Duan et al. (1983). In such cases, estimates based on logged models are often much more precise and robust than direct analysis of the unlogged original dependent variable.

Although such estimates may be more precise and robust, no one is interested in log model results on the log scale per se. Congress does not appropriate log dollars. First Bank will not cash a check for log dollars. Instead, the log scale results must be retransformed to the original scale so that one can comment on the average or total response to a covariate x. There is a very real danger that the log scale results may provide a very misleading, incomplete, and biased estimate of the impact of covariates on the untransformed scale, which is usually the scale of ultimate interest.

In the following discussion, the major focus is on the effect of heteroscedasticity on estimates of the mean response. In many applications, the interest is what happens to a population of interest. However, in some applications, there is also an interest in what happens to an individual of interest. In both cases, an examination of $e^{x\beta}$, which ignores the role of heteroscedasticity, provides a biased estimate of the response of interest.

3. Expectation of y—the normal case

To see this, we need to write out the expectation of the untransformed dependent variable from a model using a logged dependent variable, that is:

$$\ln(y) = x\beta + \epsilon \tag{2}$$

where $E(\epsilon) = 0$ and $E(\epsilon|\mathrm{x}) = 0$. Although the error term on the log scale is independent of x, it may not exhibit constant error variance; that is, $E(\epsilon^2) \neq c$, a constant. If the error term has an expected value of zero, then $E(\ln(y)) = x\beta$. However, in most cases the expectation of y is a bit more complex:

$$\begin{aligned} E(y|x) &= e^{(x\beta)}E(e^{\epsilon}) \\ &\neq e^{x\beta} \end{aligned} \tag{3}$$

In the general case, the expected value of the unlogged dependent variable is:

$$E(y|x) = \int e^{(x\beta+\epsilon)}\mathrm{d}F(\epsilon) \tag{4}$$

where F is the cdf for ϵ. Eq. (4) can be factored two terms, one that has to do with the deterministic part of the model (on the log scale) and one that has to do with the error term.

$$E(y|x) = e^{x\beta}\int e^{\epsilon}\mathrm{d}F(\epsilon) \tag{5}$$

The last term (the integral) can be treated as a constant if the error term is homoscedastic in x and any other variables z.

To illustrate this point, let us consider the case of a log normal error term, where: $\epsilon \sim N(0, \sigma^2(x))$. In that case, the expected value of y is:

$$E(y|x) = e^{x\beta+0.5\sigma^2(x)} \tag{6a}$$

$$E(y|x) > e^{x\beta} \tag{6b}$$

where $\sigma^2(x)$ indicates that the variance is a function of x. This is a straightforward extension of the mean of a log normal variable being the exponential of the (log) mean plus one half of the (log) variance (Johnson et al., 1994, pp. 211–212). Thus, the expectation of y depends on the variance and heteroscedasticity on the log scale. The term $E(e^{x\beta})$ yields an estimate of the geometric mean, not the arithmetic mean of the response function. If the variable of interest x_i is not discrete, then the slope of the expected value is given by:

$$\frac{\partial E(y|x)}{\partial x_i} = E(y|x)\left[\beta_i + 0.5\frac{\partial\sigma^2}{\partial x_i}\right] \tag{7}$$

It is this derivative which should be used in the calculation of the elasticity of the mean response, rather than β_i alone.

In the simplest of cases, consider a two population problem, an individual may be assigned to treatment A or treatment B. For each group, assume that y is distributed as

$$\ln(y_{\mathrm{G}}) \sim N(\mu_{\mathrm{G}}, \sigma_{\mathrm{G}}^2) \tag{8}$$

where G = A or B. If the two populations differ in their (log) variances, then the two population (e.g., treatment) comparison exhibits the heteroscedastic retransformation problem that is of concern here. Using the formulas in Eq. (6a), the ratio of the expected value under the two treatments is given by:

$$\frac{E(y_{\mathrm{A}})}{E(y_{\mathrm{B}})} = e^{(\mu_{\mathrm{A}}-\mu_{\mathrm{B}})+0.5(\sigma_{\mathrm{A}}^2-\sigma_{\mathrm{B}}^2)} \tag{9}$$

because the error variance may differ by treatment group.[2] If the error terms were homoscedastic across treatment groups ($\sigma^2 =$ a constant), then (and only then) would the ratio of expected outcomes for the two treatments depend only on their means on the log scale.

4. Expectation of *y*—the non-normal case

If the error term is not normally distributed, there are two alternatives. If the error term is known to follow a specific distribution, then the expectation of the exponentiated error ($E(e^{\epsilon})$) can be derived directly. If the distribution is not known a priori, then one nonparametric alternative is the smearing estimator developed by Duan (1983), and which was applied in a number of the Health Insurance Experiment papers (Duan et al., 1983, 1984; Manning et al., 1987; Newhouse et al., 1981, 1993). The smearing estimator uses the average of the exponentiated residuals to estimate the expectation of exponentiated error term. Under the assumptions about ϵ given above, the smearing estimate provides a consistent estimate $E(e^{\epsilon}))$ when using the least squares residuals. If the error term is heteroscedastic by treatment group (A or B), then the ratio of the expected values in the two population problem is:

$$\frac{E(y_{\mathrm{A}})}{E(y_{\mathrm{B}})} = [e^{(\mu_{\mathrm{A}} - \mu_{\mathrm{B}})}]\frac{S_{\mathrm{A}}}{S_{\mathrm{B}}} \qquad (10)$$

where s_i is the smearing coefficient for subgroup i (see Appendix A).

One of the consequences of having heteroscedasticity in a transformed model is that the results for inferences of the mean response on the untransformed scale will be less precise than if the error where homoscedastic. In the homoscedastic case, the raw scale inferences on contrasts between two treatments will depend on both the variance in the difference of the two estimates of the two log means (μ_{A}, and μ_{B}), and also on the variance in the difference in the estimates of the two log variances.[3] In many applications, this second term could be quite large, given the variance is less precisely estimated than the mean.

[2] One way of thinking about heteroscedasticity by treatment group is that each subject has an unobserved propensity for higher y given by an error term e_i, where e is distributed as a $N(0,1)$ variate. Once a treatment group G has been selected then the outcome is:

$$\ln(y_i) = \mu_{\mathrm{G}} + \sigma_{\mathrm{G}} e_i$$

where the treatment determines both the mean and the spread in the distribution of outcomes.

[3] For the normal theory variant in Eq. (9), the variance of the ratio can be approximated using the delta method. In this case,

$$\mathrm{Var}\left(\frac{\bar{Y}_{\mathrm{A}}}{\bar{Y}_{\mathrm{B}}}\right) \approx \left(\frac{\bar{Y}_{\mathrm{A}}}{\bar{Y}_{\mathrm{B}}}\right)^2 \left[\left(\mathrm{Var}(\hat{\mu}_{\mathrm{A}}) + \mathrm{Var}(\hat{\mu}_{\mathrm{B}})\right) + 0.25\left(\mathrm{Var}(\hat{\sigma}_{\mathrm{A}}^2) + \mathrm{Var}(\hat{\sigma}_{\mathrm{B}}^2)\right)\right].$$

5. Alternative transformations

This issue of retransformation and heteroscedasticity is not unique to the case of a logged dependent variable. Any power transformation of y will raise this issue. If the square root transformation is used, then

$$\sqrt{y} = x\delta + \nu \tag{11}$$

where ν has zero mean and is independent of x. Then the expected value of y is:

$$E(y|x) = (x\delta)^2 + \sigma_\nu^2 \tag{12}$$

If the error term ν is heteroscedastic in x (or in treatment groups), then the effect of the heteroscedasticity on the arithmetic mean is additive, while for the log model, the effect is multiplicative. For an example, see Ettner et al. (1996).

For the general case of the Box–Cox model, given in Eqs. (1a) and (1b) above, the expectation of y is:

$$E(y|x) = \int [1 + \lambda(x\beta + \epsilon)]^{1/\lambda} \mathrm{d}F(\epsilon) \tag{13}$$

if λ is not equal to zero; if λ is equal to zero, then the expression in Eqs. (6a) and (6b) holds true. See Taylor (1986) for additional discussion Thus, the expression for $E(y)$ is easiest to calculate if $1/\lambda$ is an integer, such as in Eq. (12).

6. Related transformed model issues

Heteroscedasticity is not the only cause of re-transformation issues for logged (or other power transformed) models. There are a number of commonly used techniques for dealing with econometric problems that raise re-transformation issues, some of which exacerbate the issues raised earlier. These include transformations to do GLS, two stage least squares, and the LIML (Mills ratio) version of the Selection Model.

One commonly used methods for dealing with heteroscedasticity in estimating β is to transform the dependent and independent variables to remove the heteroscedasticity. This method produces estimates of β and the Var (β) that are identical to those produced by using weighted least squares directly on Eq. (2). However, when the dependent variable has already been transformed by the log (or any other power transformation), multiplying both $\ln(y)$ and x by an additional transform (T_i equal to the square root of the inverse of the variance) gives an efficient estimate of β, a consistent estimate of $V(\hat{\beta})$, but will give a biased estimate of the relevant variance σ^2. The reason is that the model is estimating a transform of σ^2, which should have an expectation of 1.0, rather than σ^2 itself.[4] In fact, if the functional form for the transform is correct, then the transformed error term $T_i \epsilon_i$ is homoscedastic, while ϵ may be quite heteroscedastic. Thus, if possible, either the model should be estimated by using the weights option in the regression package, or by using the estimate of β, $\ln(y)$, and x to calculate the true residual directly:

$$\hat{\epsilon}_i = \ln(y_i) - x_i \hat{\beta}_{\text{GLS}} \tag{14}$$

A similar issue arises in the use of two stage least squares and some other instrumental variables estimators. Many analysts will replace the endogenous explanatory variable by its predicted value, using all of the exogenous variables or the appropriate instruments as covariates. Although this approach produces unbiased estimates of β, the residual from such an approach provides an inconsistent estimates of the key parameters (e.g., the variance) of the distribution of the true error term ϵ. Instead, the revised error resulting from this substitution is

$$r = \epsilon + \left(Y_{\text{RHS}} - \hat{Y}_{\text{RHS}}\right)\beta_{Y_{\text{RHS}}} \tag{15}$$

where RHS indicates the right hand side endogenous variable, and Y_{RHS} is the predicted value for the right hand side variable, based on the available instruments. The variance of this revised error is more variable than is ϵ. It may also be heteroscedastic, even when ϵ is not. Either artifact could lead to a biased estimate of $E(y)$. Instead of using r, one should use an estimate based on

$$\hat{\epsilon}_i = \ln(y_i) - x_i \hat{\beta}_{\text{IV}} \tag{16}$$

where the x is the true set of covariates, including the original endogenous explanatory variable. The estimate in Eq. (16) is the basis for the standard way to provide the correct estimate of the variance covariance in two-stage least squares and other IV estimators. If the true error is heteroscedastic, then an additional correction using the residuals from Eq. (18) is required to obtain the expectation of y.

The use of the LIML or the inverse Mill's ratio version of the Selection Model raises similar problems. The use of the Mill's ratio actually may induce heteroscedasticity in the residuals when none is truly present in the underlying error term; see Greene (1993) (pp. 707–714) for the case of incidental truncation. To further complicate matters, the error term in the LIML equation that is being estimated is less variable than the true error variance absent the selection process.

[4] If the underlying true model is given by Eq. (2) with the error term having a variance σ_i^2 for the ith case, then the transformed model is $(T_i y_i) = (T_i x_i)\beta + (T_i \epsilon_i)$ where $T_i = 1/\sigma_i$. Then the variance of the transformed error $(T_i \epsilon_i) = \sigma_i^2/\sigma_i^2 = 1$, which is not σ_i^2.

To find the correct estimate of the (unselected) error variance, one must use FIML or back out the estimate of the true error variance using formulas such as those provided by Greene (Greene, 1993 p. 712).[5] If the true error term is heteroscedastic, then one must incorporate an additional correction to obtain the expected value of y.

Although technically not an issue of re-transformation, there is one common practice in estimating mean responses that will not provide a consistent estimate for log models. In the case of untransformed dependent variables, the mean response can be estimated by using the mean x multiplied by an unbiased estimate of β. This ability is one of the convenient properties of the OLS estimate that disappears when the dependent variable is transformed. With transformed dependent variables, the mean response is *not* equal to the response of the person with mean values for x's. Instead, the mean response is the mean of the retransformed estimate of y, which depends on the distribution of the x's, not just their mean. That is,

$$(\bar{y}|x) = \frac{1}{n}\Sigma\left[e^{x_i\hat{\beta}+0.5\hat{\sigma}^2}\right] \tag{17}$$

in the case that ϵ is distributed as a log normal; a similar expression can be developed if the smearing estimate is used, or if an alternative distribution is known to hold. The reason for needing to look at the average of the retransformed y is due to Jensen's Inequality, which states that if $f(x)$ is convex, then $E(f(x)) \geq f(E(x))$.

7. Effect on an individual prediction

In some cases, the analytical goal is to determine the effect of a variable on a particular individual, not the mean over a population of interest. If the error term

[5] Let the problem of selection be characterized by a bivariate normal model where the case

$$I_i = z_i\gamma + u_i$$

is observed with the equation of interest if some latent underlying index function I_i is positive: where z is a row vector of exogenous variables, including some that are excluded from x_i, and γ is a column vector of parameters to be estimated. In this case, (u,ϵ) are distributed as bivariate be written as normal [0, 0, 1, σ, ρ]. Then the LIML version of Eq. (2) for the cases observed with $\ln(y)$ is

$$\ln(y_i)x_i\beta + \rho\sigma_\epsilon\Lambda(z_i\gamma) + \nu_i$$

where $\Lambda(z_i\gamma)$ is $\phi(z_i\gamma)/\Phi(z_i\gamma)$, ϕ and Φ are the normal pdf and cdf, respectively. In this case, the relationship of ν to ϵ is given by

$$\text{Var}(\nu_i) = \sigma_\epsilon^2\left(1-\rho^2\delta_i\right)$$

where $\delta_i = \Lambda_i(\Lambda_i + z_i\gamma)$.

on the log scale is represented by $\epsilon_i = \sigma e_i$ (where e_i is $N(0,1)$), then Eq. (9) comparing the effect of treatments A and B at the individual level is:

$$\frac{y_{\mathrm{A},i}}{y_{\mathrm{B},i}} = e^{[(\mu_{\mathrm{A}} - \mu_{\mathrm{B}}) + (\sigma_{\mathrm{A}} - \sigma_{\mathrm{B}})e_i]} \tag{18}$$

Because the individual prediction depends on the individual's actual e_i, the formula in Eq. 20 differs from Eq. (9). If we were to average Eq. (18) over all individuals, the result would be Eq. (9). If there is heteroscedasticity present, then the relative effect of A compared to B depends on how far the individual is from zero in his error term. The relative effect is no longer a proportional shift that is the same for all individuals.

8. Example

Many of the Health Insurance Experiment (HIE) papers on expenditures and utilization relied on regression models with logged dependent variables. With standard deviations two to four times the mean, the log transformation was essential to finding estimates of the response of health care expenditures and utilization that were robust to the skewness in the data (Duan et al., 1983). In many of the analyses, the residual errors indicated the presence of appreciable heteroscedasticity by insurance plan, the primary covariate of interest. To avoid the potential bias described above, HIE analysts used a heteroscedastic smearing factor that reflected differences in the residual variance on the log scale (Newhouse et al., 1981; Manning et al., 1987; Newhouse et al., 1993). Because the error terms were not normally distributed, those papers used a heteroscedastic version of the smearing estimator developed by Duan (1983); the smearing coefficients were the average of the exponential of the model's residuals for that subgroup (e.g., by plan or other category).

To illustrate the problems, Table 1 reports the means, variances and other summary statistics on the log of medical expenditures (in 1995 US$) for those individuals who had any medical care expenditures during the year in question. For a description of the sample and the data, see Manning et al. (1987) or Newhouse et al. (1993). As the table indicates, the residuals on the log scale for users are more variable for the plans with greater cost-sharing, while the log means for these plans are lower. If the normal theory retransformation under homoscedasticity (Eq. (9), assuming equal variances) were to be used, the 95% plan [6] would have 30% lower expenditures than the free plan (= [exp(5.998 −

[6] The 95% plan required families to pay 95% of the medical bill up to a stop-loss that was either US$1000 (in 1970s US$) or a fraction of income, whichever was less. Beyond the stop-loss, the family paid nothing out-of-pocket for the remainder of that accounting year.

Table 1
Univariate statistics on log medical expenditures for users* health insurance experiment

Plan	Mean*	Variance**	Skewness	Kurtosis	Percentile					
					10%	25%	50%	75%	90%	95%
Free	6.349	2.083	0.202	3.385	4.541	5.469	6.289	7.166	8.290	8.976
25%	6.128	2.133	0.456	3.539	4.393	5.191	6.004	6.907	8.117	8.849
50%	6.039	2.003	0.513	4.067	4.265	5.134	5.969	6.766	7.906	8.628
95%	5.998	2.343	0.464	3.226	4.158	4.948	5.873	6.862	8.120	8.892
Individual deductible	6.153	2.384	0.305	2.993	4.237	5.068	6.039	7.093	8.322	8.876

*In 1995 US$, adjusted using the medical care component of the Consumer Price Index.
**Plan differences are significant at $F < 0.001$ for both mean and variance, by F-tests. Tests for variance based on Park Test.

6.349)] − 1). However, if the normal theory retransformation under heteroscedasticity were to be used (Eq. (9)), then the 95% plan would have a 20% lower expenditures (= [exp(5.998 − 6.349 + 0.5(2.343 − 2.083)] − 1). In this case, just using the log means of the plans overstates the plan response by a half. [Because the error term is not normally distributed, one should use a non-normal retransformation such as Duan's smearing factor (Duan, 1983). However, the qualitative pattern of the results still holds.] In the main HIE papers (Newhouse et al., 1981; Manning et al., 1987; and Newhouse et al., 1993), medical expenditures were split into outpatient only, and any inpatient because of the complex effect of covariates on inpatient expenditures and hence on the right tail of the medical care distribution. In that case, that split was preferable to making the correction above; see Duan et al. (1983) for additional details.

Table 1 also illustrates the effects of the HIE insurance plans on individuals at different points in the distribution of expenditures. If we were to use the difference in mean logs to estimate plan differences, then the 95% plan would have expenditures for users that were about 30% below those for users on the free plan at every percentile. However, allowing for heteroscedasticity across plans provides a different story. Moving from the lower quartile on the free plan to that on the 95% plan involves a reduction of 41%, while moving from the median on the free plan to that on the 95% plan involves a reduction of 34%. In contrast, at the 90th percentile, the 95% plan is 16% lower than the corresponding percentile for the free plan. The differences are even smaller at higher quantiles.

9. Conclusion—the effect of heteroscedasticity

In the case of least squares on an untransformed dependent variable, the possibility of heteroscedasticity should raise concerns about the efficiency of the

OLS estimate of β, and about the consistency of the OLS estimate of the variance of the OLS estimate of β. Most of us have learned to use GLS estimators to obtain efficient estimates of β and the correct inference statistics for the variance of the estimate of β. Failing to do that, we know that the Huber/White estimate of the variance–covariance matrix for the OLS estimate of β should be used to get consistent inference statistics.

In the case of the log (or any other transformed) dependent variable, the analyst needs to perform one additional set of tasks. The analyst needs to determine if the error term is heteroscedastic across treatment groups or depends on some combination of x's. If the error terms is heteroscedastic, then the analyst must ascertain the form of the heteroscedasticity, and then use that information to obtain an unbiased estimate of the retransformation factor in order to estimate the overall expected response of y to x.

Once the analyst transforms the dependent variable, he can no longer assume away or ignore the problem of heteroscedasticity or make appeals to Huber/White corrections as being complete. GLS estimates or Huber/White corrections should be used for reasons of efficiency and correct inference. However, the price of using a log (or any other) transform of the dependent variable is that the analyst *must* also learn more about the nature of the error structure than is commonly the case. Failure to carry out this additional step may lead to substantially biased estimates of the average effects of the treatments and other explanatory variables, and of their elasticities that depend on mean responses. Unfortunately, one cannot usually tell the direction or magnitude of the bias from ignoring heteroscedasticity on a priori grounds, because theory rarely tells us the nature of the heteroscedasticity. However, in the example discussed here, the effect of the ignoring heteroscedasticity was dramatic, whether one was concerned with the impact of insurance on populations or on individual's at different points in the distribution.

Acknowledgements

This paper has benefitted from the comments made by Bryan Dowd, Mike Finch, Richard Frank, John Mullahy, Tamara Stoner, and an anonymous reviewer. The original interest in this area began in collaboration with Naihua Duan, Carl Morris, and Joseph Newhouse on the Health Insurance Experiment. The current analysis and conclusions presented here are the author's alone.

Appendix A

The variance of the smearing estimate is

$$\mathrm{Var}(\bar{s}) = \frac{1}{n-1}\Sigma\left(e^{\hat{\epsilon}_i} - \left(\frac{1}{n}\Sigma e^{\hat{\epsilon}_i}\right)\right)^2$$

The delta method estimates the variance of the raw scale mean of y using the smearing estimator is

$$\operatorname{Var}(\bar{y}_{\mathrm{A}}) \approx (\bar{s}_{\mathrm{A}} e^{\hat{\mu}_{\mathrm{A}}}) \begin{pmatrix} \operatorname{Var}(\hat{\mu}_{\mathrm{A}}) & 0 \\ O & \operatorname{Var}(\bar{s}_{\mathrm{A}}) \end{pmatrix} \begin{pmatrix} \bar{s}_{\mathrm{A}} \\ e^{\hat{\mu}_{\mathrm{A}}} \end{pmatrix}$$

$$\approx (\bar{s}_{\mathrm{A}})^2 \operatorname{Var}(\hat{\mu}_{\mathrm{A}}) + (e^{\hat{\mu}_{\mathrm{A}}})^2 \operatorname{Var}(\bar{s}_{\mathrm{A}})$$

ignoring a small negative correlation between the estimate of the smearing term and the mean on the log scale; including, the covariance term would reduce the estimated variance.

The variance of the ratio in Eq. (10) is

$$\operatorname{Var}\left(\frac{\bar{y}_{\mathrm{A}}}{\bar{y}_{\mathrm{B}}}\right) \approx \left(\frac{e^{\hat{\mu}_{\mathrm{A}} - \hat{\mu}_{\mathrm{B}}}}{s_{\mathrm{B}}}\right)^2 \left(\bar{s}_{\mathrm{A}}, -\bar{s}_{\mathrm{A}}, 1, \frac{-\bar{s}_{\mathrm{A}}}{s_{\mathrm{B}}}\right)$$

$$\begin{pmatrix} \operatorname{Var}(\hat{\mu}_{\mathrm{A}}) & 0 & 0 & 0 \\ 0 & \operatorname{Var}(\hat{\mu}_{\mathrm{B}}) & 0 & 0 \\ 0 & 0 & \operatorname{Var}(\bar{s}_{\mathrm{A}}) & 0 \\ 0 & 0 & 0 & \operatorname{Var}(\bar{s}_{\mathrm{B}}) \end{pmatrix} \begin{pmatrix} \bar{s}_{\mathrm{A}} \\ -\bar{s}_{\mathrm{A}} \\ 1 \\ \dfrac{-\bar{s}_{\mathrm{A}}}{\bar{s}_{\mathrm{B}}} \end{pmatrix}$$

References

Box, G., Cox, D., 1964, An analysis of transformations. J. R. Stat. Soc. Ser. B, pp. 211–264.

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. J. Am. Stat. Assoc. 78, 605–610.

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P., 1983. A comparison of alternative models for the demand for medical care. J. Bus. Econ. Stat. 1, 115–126.

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P., 1984. Choosing between the sample selection model and the multi-part model. J. Bus. Econ. Stat. 2, 283–289.

Ettner, E.L., Frank, R.G., McGuire, T.G., Newhouse, J.P., Notman, E.H., 1996. Risk adjustment of mental health and substance abuse payments, Draft, Harvard School of Medicine.

Greene, W.H., 1993. Econometric Analysis, McMillan, New York.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, 2nd edn., Vol. 1, Wiley, New York.

Kennedy, P., 1992. A Guide to Econometrics, 3rd edn., MIT Press, Boston.

Kennedy, P., 1983. Logarithmic dependent variables and prediction bias. Oxford Bull. Econ. Stat. 45, 389–392.

Kennedy, P., 1981. Estimation with correctly interpreted dummy variables in semilogarithmic equations. Am. Econ. Rev. 71, 801.

Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., Leibowitz, A., Marquis, M.S., 1987. Health insurance and the demand for medical care: evidence from a randomized experiment. Am. Econ. Rev. 77, 251–277.

McCuen, R., Leahy, R., Johnson, P., 1990. Problems with logarithmic transformations in regression. J. Hydraulic Eng. 116, 414–428.

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

Newhouse, J.P., Manning, W.G. et al., 1981. Some interim results from a controlled trial in health insurance. New England J. Med. 305, 1501–1507.

Puma, M.J., Hoaglin, D.C., 1990. Food stamp payment error rates: can state-specific performance standards be developed?. J. Am. Stat. Assoc. 85, 891–899.

Showalter, M., 1994. A Monte Carlo investigation of the Box–Cox model and a nonlinear least squares alternative. Rev. Econ. Stat. 76, 560–570.

Taylor, J.M., 1986. The retransformed mean after a fitted power transformation. J. Am. Stat. Assoc. 81, 114–118.