

Standard errors for the retransformation problem with heteroscedasticity

Chunrong Ai^a, Edward C. Norton^{b,*}

^a *University of Florida, Florida, USA*

^b *Department of Health Policy and Administration, CB# 7400, McGavran-Greenberg Building,
University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-7400, USA*

Received 1 September 1999; received in revised form 1 March 2000; accepted 8 March 2000

Abstract

Economists often estimate models with a log-transformed dependent variable. The results from the log-transformed model are often retransformed back to the unlogged scale. Other studies have shown how to obtain consistent estimates on the original scale but have not provided variance equations for those estimates. In this paper, we derive the variance for three estimates — the conditional mean of y , the slope of y , and the average slope of y — on the retransformed scale. We then illustrate our proposed procedures with skewed health expenditure data from a sample of Medicaid eligible patients with severe mental illness. © 2000 Elsevier Science B.V. All rights reserved.

JEL classification: C20; I19

Keywords: Retransformation; Heteroscedasticity; Logged dependent variable

1. Introduction

Economists often estimate models with a log-transformed dependent variable. Justifications for using the log transformation include to deal with a dependent variable badly skewed to the right, and to interpret a covariate as either an elasticity or having a multiplicative response (Manning, 1998). The results from

* Corresponding author. Tel.: +1-919-966-8930; fax: +1-919-966-6961.
E-mail address: edward_norton@unc.edu (E.C. Norton).

the log-transformed model are often retransformed back to the unlogged scale to make inferences in the natural units of, say, dollars instead of log dollars. A log-transformed model is often the second part of the two-part model of health care expenditures when some individuals have zero expenditure (Duan et al., 1983, 1984).

It is now well known that the retransformed estimate of either the conditional mean or the effect of an independent variable on the dependent variable (the slope) must adjust for the distribution of the error term and for heteroscedasticity (Manning, 1998; Mullahy, 1998). Failure to account for the distribution of the error term and heteroscedasticity may lead to substantially biased estimates of the conditional mean and the slope on the original scale.

Many studies have carefully addressed the retransformation problem by explaining how inferences on the original scale may be biased, providing equations, and illustrating with examples (e.g. Duan et al., 1983; Manning, 1998; Mullahy, 1998). However, these studies have not fully addressed the standard errors of the estimates. Like any other estimates, the estimates on the original scale of the conditional mean and the slope are calculated with uncertainty. Without computing the correct standard errors, it would be easy to draw incorrect inferences about the effect of an independent variable on the dependent variable. When testing theory on the raw scale, the statistical significance of the corresponding parameter is not always a good guide. An insignificant variable on the log scale can still have a significant effect on the raw scale, for example if there is heteroscedasticity. A significant variable on the log scale, on the other hand, may be insignificant on the raw scale. Significance on the raw scale depends on retransforming both the predicted value and heteroscedasticity, and these two effects may offset each other. Theory on the raw scale must be tested on the raw scale, and to test theory one must have standard errors or confidence intervals. Like the estimates themselves, the standard errors require special equations that take into account both the distribution of the error term and the heteroscedasticity. We are not aware of any published formulas for computing the standard error of estimates based on the retransformed data (see Manning (1998) for estimates of the variance of the smearing estimate and the raw scale mean of y using the smearing estimator).

We derive the variance of three estimates based on results from a log transformed model retransformed back to the original scale: the conditional mean of y , the slope of y , and the average slope of y , where y is the unlogged dependent variable. The variance can then be used to construct standard errors. Because construction of confidence intervals requires the central limit theorem to provide the asymptotic normal approximation to the sampling distribution, and the normal approximation may not be accurate for a finite sample size, we focus on the standard errors. We derive variance formulas for normal linear, normal nonlinear, and nonparametric models, all with possible heteroscedasticity. The formulas are easy to implement in standard software packages, and a sample program in Stata is available upon request. We illustrate with skewed health

expenditure data from a sample of Medicaid eligible patients with severe mental illness. We derive formulas for the log transformation because it is by far the most common in the literature. Researchers should read the literature and investigate the data empirically to decide whether the log transformation is most appropriate for their data (Jones and Yen, 1999; Manning and Mullahy, 1999).

2. Framework and methods

Before deriving the variance formulas for the most general case, we start with the familiar normal linear model for a log-transformed dependent variable. The normal linear model provides intuition for the more general models, and is the model that is most often estimated by applied economists. Then we extend the model to the nonlinear case, which involves only a slight modification, before presenting the nonparametric model. The nonparametric model has the normal linear and nonlinear models as special cases, and we relate it to Duan's (1983) smearing estimator.

2.1. Model and notation

We start by introducing the general model and notation. Assume that the logarithm of the original dependent variable y is a (possibly nonlinear) function of K explanatory variables x , including a constant, and a vector of unknown parameters β . The model for a single observation (we suppress the subscript i) can be written as

$$\ln(y) = h(x, \beta) + \varepsilon,$$

where ε is an additive error term. We assume that no regressor can be expressed as a linear or nonlinear function of the other regressors, e.g. there are no higher-order or interaction terms. This assumption is not as restrictive as it appears since it can always be satisfied by redefining the function h . For example, if x includes a constant, income, and income squared, then the function $h(x, \beta)$ can be redefined as a function of a constant and income. The error term ε is the product of two elements

$$\varepsilon = v\sqrt{s(x, \alpha)}$$

where v is assumed to be independent of x and has zero mean and unit variance, the function $s(\cdot)$ allows for heteroscedasticity conditional on x , and α is a vector of unknown parameters. The first part of the paper assumes that v has a normal distribution, but later this assumption is dropped.

We are interested in three estimands based on the unlogged dependent variable y . The first is the mean of y conditional on x . This is the predicted value of the raw dependent variable for one specific x . The second is the derivative (local slope) of the mean with respect to one continuous explanatory variable x^k conditional on x , where x^k denotes the k th element of x . This is the change in the predicted value for a change in one independent variable, for one specific x . The third estimand depends on whether the regressors are treated as nonrandom or random. If the regressors are nonrandom, the third estimand is the simple average slope over all observations. If the regressors are random the third estimand is the average slope over the parent distribution of the regressors.

We introduce the notation that $\mu(x)$ is the mean of y conditional on x and that θ_1 and θ_2 denote the two average slopes. This notation will simplify later formulas. The function also depends on the true values of β and α , but these arguments will be suppressed for exposition. The estimands are as follows.

- Mean of y conditional on x

$$\mu(x) = E[y | x],$$

- Slope of mean with respect to x^k

$$\frac{\partial \mu(x)}{\partial x^k} = \frac{\partial E[y | x]}{\partial x^k},$$

- Average slope (two versions)

$$\text{for nonrandom } x: \theta_1 = \frac{1}{N} \sum_{i=1}^N \frac{\partial \mu(x_i)}{\partial x^k}.$$

$$\text{for random } x: \theta_2 = E \left[\frac{\partial \mu(x)}{\partial x^k} \right].$$

2.2. Normal linear model

To present the intuition behind our procedures, we begin with the simple normal linear model. The assumption of linearity implies that $h(x, \beta) = x'\beta$ and $s(x, \alpha) = x'\alpha$. Note that because elements of x are functionally independent, the functions $h(x, \beta)$ and $s(x, \alpha)$ are not only linear in the parameters but also linear in the regressors. This distinction becomes important when taking derivatives. If either $h(x, \beta)$ or $s(x, \alpha)$ is linear in parameters but nonlinear in regressors (e.g. higher-order or interaction terms), then the total derivative with respect to x^k involves multiple terms. The assumption of a linear variance is made only for illustration; an estimated linear variance could be negative. The reader will want to use the more general model presented later. The choice of model specification for the variance function is not always clear. One way to choose between two

nonnested models is to use the J -test for linear specifications, and the P_E test for nonlinear specifications (see Greene, 2000, pp. 302 and 441).

Consistent estimates for the slope and the average slope for a model nonlinear in the regressors can be obtained using the formulas in Section 2.3. We also assume that the random variable v has a standard normal distribution. The assumption of normality is attractive because: (1) it is often assumed in the literature, and (2) the retransformed estimands are simple to derive (Manning, 1998, pp. 285–287).

Under the assumptions of linearity and normality the first two estimands for fixed x are

$$\mu(x) = E[y|x] = e^{x'\beta + 0.5x'\alpha},$$

$$\frac{\partial \mu(x)}{\partial x^k} = (\beta_k + 0.5\alpha_k)\mu(x),$$

where the subscript k refers to the k th component of the parameter vector. Estimate the three estimands by replacing β and α with consistent estimates. To estimate α , regress the squared residuals from the ordinary least squares model on x , then multiply x by the estimated parameter vector $\hat{\alpha}$. Then the estimated mean of y conditional on x is the exponentiated predicted value of y , with a correction for the variance of the normal error term

$$\hat{\mu}(x) = e^{x'\hat{\beta} + 0.5x'\hat{\alpha}}.$$

The estimated slope with respect to one continuous variable x^k is the mean multiplied by the sum of two terms. The following is similar to Manning's (1998) Eq. (7), but we assume a particular form of the heteroscedasticity, namely $s(x, \hat{\alpha}) = x'\hat{\alpha}$, so the slope is

$$\frac{\partial \hat{\mu}(x)}{\partial x^k} = (\hat{\beta}_k + 0.5\hat{\alpha}_k)\hat{\mu}(x).$$

Finally, both versions of the third estimand are estimated by the sample average of the slope over all observations because the expectation is usually approximated by the sample average, leading to

$$\hat{\theta}_1 = \hat{\theta}_2 = \frac{1}{N} \sum_{i=1}^N \frac{\partial \hat{\mu}(x_i)}{\partial x^k}.$$

Under standard regularity conditions, the least squares estimator $\hat{\beta}$ is normally distributed. It is less well known, however, that the least squares estimator $\hat{\alpha}$ is asymptotically normally distributed. In the case of homoscedastic errors $\hat{\alpha}$ is estimated by the sample variance which has a chi-square distribution with degrees of freedom depending on the sample size N . The chi-square distribution with

many degrees of freedom looks like a normal distribution and so can be approximated by a normal distribution when sample size goes to infinity. Hence, the parameter estimator $\hat{\alpha}$ is asymptotically normally distributed. Moreover, $\hat{\beta}$ and $\hat{\alpha}$ are independent. Thus, applying the delta method gives the scalar variance of these three estimands

$$\omega_1(x) = \left(\frac{\partial \mu(x)}{\partial \beta'} \Sigma_{\beta} \frac{\partial \mu(x)}{\partial \beta} \right) + \left(\frac{\partial \mu(x)}{\partial \alpha'} \Sigma_{\alpha} \frac{\partial \mu(x)}{\partial \alpha} \right), \tag{1}$$

$$\omega_{2k}(x) = \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \Sigma_{\beta} \frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right) + \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha'} \Sigma_{\alpha} \frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right), \tag{2}$$

$$\omega_{3k} = \left(\left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \mu(x_i)}{\partial x^k \partial \beta'} \right) \Sigma_{\beta} \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \mu(x_i)}{\partial \beta \partial x^k} \right) \right) + \left(\left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \mu(x_i)}{\partial x^k \partial \alpha'} \right) \Sigma_{\alpha} \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \mu(x_i)}{\partial \alpha \partial x^k} \right) \right), \tag{3}$$

$$\omega_{4k} = \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \right] \Sigma_{\beta} E \left[\frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right] \right) + \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha'} \right] \Sigma_{\alpha} E \left[\frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right] \right) + \Sigma_{\mu}, \tag{4}$$

where the terms with $\partial \mu(x)$ in the numerator are first partial derivatives, the terms with $\partial^2 \mu(x)$ in the numerator are second partial derivatives, and Σ_{β} and Σ_{α} are the heteroscedasticity consistent covariance matrices for β and α , respectively. The White heteroscedasticity consistent estimator can be found in Greene (2000, p. 463) and can be computed by most regression packages. The third term of Eq. (4), Σ_{μ} , is the sample variance of $\partial \mu(x_i) / \partial x^k$ divided by N . This term is needed because we use the sample average to estimate the average slope over the parent distribution of the regressors.

Before describing how to estimate these in practice, we first note several features of these variances. Each variance is a scalar. Each variance is the sum of two (or three) parts; the first part corresponds to the retransformation of the predicted value of $\ln(y)$, and the second part corresponds to the retransformation of the error term with estimated heteroscedasticity. Each of the first two parts is a sandwich estimator.

In summary, the computation of the estimated variance for the normal linear model requires four steps. Estimate the model on the log scale, estimate the form of heteroscedasticity, compute several derivatives (the hardest step), and plug the MLE estimates for the variance-covariance matrices into Eqs. (1)–(4).

Step I: Apply ordinary least squares to the logarithm of y_i on x_i . Compute the scalar predicted value $x'\hat{\beta}$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix as $\hat{\Sigma}_\beta$. Save the residuals $\hat{\varepsilon}_i$, and square them.

Step II: Apply ordinary least squares to $\hat{\varepsilon}_i^2$ on x_i . Compute the scalar predicted value $x'\hat{\alpha}$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix as $\hat{\Sigma}_\alpha$.

Step III: Compute $\hat{\mu}(x)$, and the first derivatives of $\hat{\mu}(x)$ with respect to β , α , and x , and their cross partial derivatives either numerically or analytically (see Appendix A for formulas). To estimate Σ_μ , compute the sample variance of $\alpha\hat{\mu}(x_i)/\partial x^k$.

Step IV: Plug the estimated values from Step III into Eqs. (1)–(4). The equation for $\hat{\omega}_1(x)$ simplifies to $\hat{\omega}_1(x) = (x'\hat{\Sigma}_\beta x + 0.25x'\hat{\Sigma}_\alpha x)\hat{\mu}^2(x)$, although the other equations do not simplify.

2.3. Normal nonlinear model

Next we extend the linear results to nonlinear models but maintain the normality assumption. There are three ways the nonlinear model is a generalization of the linear model. First, the nonlinear model allows nonlinearity in regressors. In our notation, the linear specification excludes squared and interaction terms which are commonly used in the empirical economic literature. Although squared and interaction terms can be estimated using ordinary least squares, such a specification has more complex derivatives than the simple linear model, so is included in this section. Second, the nonlinear model allows for arbitrary nonlinear functions of the dependent variable to be estimated by least squares. Third, the linear specification of the variance of the error ($s(x, \hat{\alpha}) = x'\hat{\alpha}$) could allow a negative estimate, which is clearly impossible for the variance itself. The nonlinear specification such as $e^{x'\alpha}$ guarantees that the estimated variance will always be positive. In addition one may wish to include higher-order and interaction terms in the variance function.

The parameter estimates $\hat{\beta}$ and $\hat{\alpha}$ now can be obtained by nonlinear least squares regressions, or by ordinary least squares if $h(x, \beta)$ and $s(x, \alpha)$ are linear in the parameters. The parameter estimates $\hat{\beta}$ and $\hat{\alpha}$ are now asymptotically independent and normally distributed (see Appendix A for a formal proof). Hence, Eqs. (1)–(4) are still valid. Despite these changes, the four steps for computing the three estimands and their respective standard errors are nearly the same as before except that the derivatives are more complicated.

Step I: Apply nonlinear least squares, or ordinary least squares if $h(x, \beta)$ is linear in the parameters, to the logarithm of y_i on $h(x_i, \beta)$. Compute the scalar predicted value $h(x, \hat{\beta})$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix as $\hat{\Sigma}_\beta$. Save the residuals $\hat{\varepsilon}_i$ and square them.

Step II: Apply nonlinear least squares, or ordinary least squares if $s(x, \alpha)$ is linear in the parameters, to $\hat{\varepsilon}_i^2$ on $s(x_i, \alpha)$. Compute the scalar predicted value

$s(x, \hat{\alpha})$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix as $\hat{\Sigma}_{\alpha}$.

Step III: Compute $\hat{\mu}(x)$, and the first derivatives of $\hat{\mu}(x)$ of with respect to β , α , and x , and their cross partial derivatives either numerically or analytically (see Appendix A for formulas). To estimate Σ_{μ} , compute the sample average of $\partial \hat{\mu}(x_i) / \partial x^k$.

Step IV: Plug the estimated values from Step III into Eqs. (1)–(4)

2.4. Nonparametric model with homoscedasticity

The normal linear and nonlinear models both assume that the distribution of the error term is known and normal. In practice, the distribution of the error term is often not known and not normal. If the distribution of the error term is not normal, then the normal linear and nonlinear models will give biased parameter estimates of $E[y|x]$, although the models are still consistent for $E[\ln(y|x)]$. This section extends the results of Section 2.2 to the case where the distribution of v is not known but the error is homoscedastic, and Section 2.5 has the most general model that is both heteroscedastic and nonparametric.

When the error is not normal, then the conditional mean of y on the natural scale is the exponentiated predicted value of y multiplied by a smearing factor adjusted for heteroscedasticity

$$\mu(x) = E[y|x] = e^{h(x, \beta)} D(x, \alpha).$$

where $D(x, \alpha) = E[e^{v\sqrt{s(x, \alpha)}} | x]$ is the smearing factor. The smearing factor was originally proposed by Duan (1983) to adjust for the retransformation of an error term with unknown distribution in the case of homoscedasticity, and extended by Manning (1998) to the case of heteroscedasticity by groups.

Before analyzing the general case with heteroscedasticity, we analyze the case of an error term with unknown distribution under homoscedasticity. If $s(x, \alpha) = \alpha$ (i.e. $s(\cdot)$ does not depend on x so the error term is homoscedastic) then the smearing factor simplifies to the one proposed by Duan (1983). In the homoscedasticity case considered by Duan, the smearing factor is estimated by $N^{-1} \sum e^{\hat{\epsilon}_i}$, where $\hat{\epsilon}_i$ is the least squares regression residual. The three estimands are simpler under homoscedasticity than under heteroscedasticity because the derivative of the heteroscedasticity term disappears. The estimated first two estimands for the nonparametric model with homoscedasticity are

$$\hat{\mu}(x) = e^{h(x, \hat{\beta})} \left(\frac{1}{N} \sum_{i=1}^N e^{\hat{\epsilon}_i} \right),$$

$$\frac{\partial \hat{\mu}(x)}{\partial x^k} = \hat{\mu}(x) \frac{\partial h(x, \hat{\beta})}{\partial x^k}.$$

The first version of the third estimand is estimated by

$$\hat{\theta}_1 = \frac{1}{N} \sum_{j=1}^N \hat{\mu}(x_j) \frac{\partial h(x_j, \hat{\beta})}{\partial x^k}.$$

The second version of the third estimand can be estimated in the same way as the first version. However, due to the presence of the estimated smearing factor, the computation of the variance is complicated by the randomness of the regressors. To simplify the computation, we present an equivalent version of θ_2 . By manipulating expectations, we have

$$\theta_2 = E \left[y \left(\frac{\partial h(x, \beta)}{\partial x^k} + v \frac{\partial \sqrt{s(x, \alpha)}}{\partial x^k} \right) \right].$$

When the errors are homoscedastic, $s(x, \alpha)$ does not depend on x , so we estimate θ_2 by

$$\hat{\theta}_2 = \frac{1}{N} \sum_{i=1}^N y_i \frac{\partial h(x_i, \hat{\beta})}{\partial x^k}$$

Removing the normality assumption creates three complications in the computation of the variances of the three estimands. First, the smearing factor cannot be computed analytically. Instead, it must be estimated by the sample average. The sample average introduces a variance term of its own which must be accounted for when computing the variances of the three estimands.

The other two complications are best understood by examining the Taylor-series expansion of the smearing factor to the first-order term

$$\frac{1}{N} \sum_{i=1}^N e^{\hat{\varepsilon}_i} \cong \frac{1}{N} \sum_{i=1}^N e^{\varepsilon_i} - \frac{1}{N} \sum_{i=1}^N \left[e^{\varepsilon_i} \left(\frac{\partial h(x_i, \beta)}{\partial \beta'} \right) (\hat{\beta} - \beta) \right]$$

The second complication comes in the second term of the expansion, which obviously depends on the parameter estimates $\hat{\beta}$. This dependency must be accounted for when computing the derivatives with respect to β . The third complication is that the smearing factor is a sample average, so the estimated smearing factor (the first term of the expansion) is a function of the residuals which may be correlated with the parameter estimates. The covariance must also be accounted for when computing the variances of the three estimands.

Under homoscedasticity, the variances of the three estimands are:

$$\omega_1(x) = \left(\frac{\partial \mu(x)}{\partial \beta'} \Sigma_{\beta} \frac{\partial \mu(x)}{\partial \beta} \right) + 2 \left(\frac{\partial \mu(x)}{\partial \beta'} \Sigma_{1D\beta} \right) + \Sigma_{1DD}, \quad (5)$$

$$\omega_{2k}(x) = \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \Sigma_{\beta} \frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right) + 2 \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \Sigma_{2D\beta} \right) + \Sigma_{2DD}, \quad (6)$$

$$\omega_{3k} = \left(\left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \beta'} \right) \Sigma_{\beta} \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial \beta \partial x^k} \right) \right) + 2 \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \beta'} \right) \Sigma_{3D\beta} + \Sigma_{3DD}, \tag{7}$$

$$\omega_{4k} = \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \right] \Sigma_{\beta} E \left[\frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right] \right) + 2 \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \right] \Sigma_{\mu\beta} \right) + \Sigma_{\mu}, \tag{8}$$

where $\Sigma_{1D\beta}$ is the covariance between the sample average of $\exp(h(x, \beta) + \varepsilon_i)$ and $\hat{\beta}$, Σ_{1DD} is the sample variance of $\exp(h(x, \beta) + \varepsilon_i)$, $\Sigma_{2D\beta}$ is the covariance between the sample average of and $\exp(h(x, \beta) + \varepsilon_i)(\partial h(x, \beta)/\partial x^k)$ and $\hat{\beta}$, Σ_{2DD} is the sample variance of $\exp(h(x, \beta) + \varepsilon_i)(\partial h(x, \beta)/\partial x^k)$, $\Sigma_{3D\beta}$ is the covariance between the sample average of $N^{-1} \sum_j \exp(h(x_j, \beta) + \varepsilon_j)(\partial h(x_j, \beta)/\partial x^k)$ and $\hat{\beta}$, Σ_{3DD} is the sample variance of $N^{-1} \sum_j \exp(h(x_j, \beta) + \varepsilon_j)(\partial h(x_j, \beta)/\partial x^k)$, $\Sigma_{\mu\beta}$ is the covariance between the sample average $y_i(\partial h(x_i, \beta)/\partial x^k)$ and $\hat{\beta}$, and Σ_{μ} is the sample variance of $y_i(\partial h(x_i, \beta)/\partial x^k)$.

To estimate the variances, simply replace those derivatives by their estimates. The proposed procedure is described as follows.

Step I: Apply nonlinear least squares, or ordinary least squares if $h(x, \beta)$ is linear in the parameters, to the logarithm of y_i on $h(x_i, \hat{\beta})$. Compute the scalar predicted value $h(x, \hat{\beta})$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix as $\hat{\Sigma}_{\beta}$. Save the residuals $\hat{\varepsilon}_i$, but do not square them. (The only difference with Step I of Section 2.3 is that the residuals are not squared.)

Step II: This step is not needed because $\hat{\Sigma}_{\alpha} = 0$.

Step III: Compute $\hat{\mu}(x)$, and the first derivatives of $\hat{\mu}(x)$ with respect to β and x , and their cross partial derivatives either numerically or analytically. To estimate the vectors $\Sigma_{1D\beta}$, $\Sigma_{2D\beta}$, $\Sigma_{3D\beta}$, and $\Sigma_{\mu\beta}$, run simple regressions and save the regression coefficients. To estimate the variances Σ_{1DD} , Σ_{2DD} , Σ_{3DD} , and Σ_{μ} , compute the sample variances and divide by N (see Appendix A for all formulas).

Step IV: Plug the estimated values from Step III into Eqs. (5)–(8).

2.5. Nonparametric model with heteroscedasticity

In the more general case where the variance depends on regressors, the smearing factor is estimated by the simple sample average $N^{-1} \sum_{i=1}^N e^{\hat{v}_i \sqrt{s(x, \hat{\alpha})}}$, where $\hat{v}_i = \hat{\varepsilon}_i / \sqrt{s(x_i, \hat{\alpha})}$ is the estimate of the standardized residuals. The exponent does not simplify to $\hat{\varepsilon}_i$ because $s(x, \hat{\alpha})$ is a scalar evaluated at the conditional x , while $s(x_i, \hat{\alpha})$ takes on a different value for each observation i .

The estimated smearing factor now depends on both parameter estimates $\hat{\beta}$ and $\hat{\alpha}$. These dependencies must be accounted for when taking derivatives with

respect to β and α . To simplify later notation, define $g_i(x, \beta, \alpha)$ to be the unbiased predicted value of $\mu(x)$ based on x_i and $q_i(\beta, \alpha)$ to be the unbiased estimate of the average slope

$$g_i(x, \beta, \alpha) = e^{h(x, \beta) + [(\ln(y_i) - h(x_i, \beta)) / \sqrt{s(x_i, \alpha)}] \sqrt{s(x, \alpha)}},$$

$$q_i(\beta, \alpha) = y_i \left(\frac{\partial h(x_i, \beta)}{\partial x^k} + 0.5 \frac{\partial \ln s(x_i, \alpha)}{\partial x^k} (\ln y_i - h(x_i, \beta)) \right).$$

The three estimands are

$$\hat{\mu}(x) = \frac{1}{N} \sum_{i=1}^N g_i(x, \hat{\beta}, \hat{\alpha}),$$

$$\frac{\partial \hat{\mu}(x)}{\partial x^k} = \frac{1}{N} \sum_{i=1}^N \frac{\partial g_i(x, \hat{\beta}, \hat{\alpha})}{\partial x^k},$$

$$\hat{\theta}_1 = \frac{1}{N} \sum_{j=1}^N \left(\frac{1}{N} \sum_{i=1}^N \frac{\partial g_i(x_j, \hat{\beta}, \hat{\alpha})}{\partial x^k} \right),$$

$$\hat{\theta}_2 = \frac{1}{N} \sum_{i=1}^N q_i(\hat{\beta}, \hat{\alpha}).$$

The heteroscedastic nonparametric model has two further complications. First, in the normal model the covariance between $\hat{\beta}$ and $\hat{\alpha}$ is zero. However, in the nonparametric model, $\hat{\beta}$ and $\hat{\alpha}$ are neither independent nor uncorrelated. Second, the covariances between the smearing factor and the parameter estimates $\hat{\beta}$ and $\hat{\alpha}$ are not zero. Thus, the formulas for the variances include additional terms for the covariances between $\hat{\beta}$ and $\hat{\alpha}$, the functions g and q and their derivatives.

$$\omega_1(x) = \left(\frac{\partial \mu(x)}{\partial \beta'} \Sigma_{\beta} \frac{\partial \mu(x)}{\partial \beta} \right) + \left(\frac{\partial \mu(x)}{\partial \alpha'} \Sigma_{\alpha} \frac{\partial \mu(x)}{\partial \alpha} \right)$$

$$+ 2 \left(\frac{\partial \mu(x)}{\partial \beta'} \Sigma_{\beta\alpha} \frac{\partial \mu(x)}{\partial \alpha} \right) + 2 \frac{\partial \mu(x)}{\partial \beta'} \Sigma_{1D\beta} + 2 \frac{\partial \mu(x)}{\partial \alpha'}$$

$$\times \Sigma_{1D\alpha} + \Sigma_{1DD}, \tag{9}$$

$$\omega_{2k}(x) = \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \Sigma_{\beta} \frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right) + \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha'} \Sigma_{\alpha} \frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right)$$

$$+ 2 \left(\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \Sigma_{\beta\alpha} \frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right) + 2 \frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \Sigma_{2D\beta} + 2 \frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha'}$$

$$\times \Sigma_{2D\alpha} + \Sigma_{2DD}, \tag{10}$$

$$\begin{aligned}
 \omega_{3k} = & \left[\left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \beta'} \right) \Sigma_{\beta} \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial \beta \partial x^k} \right) \right] \\
 & + \left[\left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \alpha'} \right) \Sigma_{\alpha} \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial \alpha \partial x^k} \right) \right] \\
 & + 2 \left[\left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \beta'} \right) \Sigma_{\beta \alpha} \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial \alpha \partial x^k} \right) \right] \\
 & + 2 \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \beta'} \right) \Sigma_{3D\beta} + 2 \left(\frac{1}{N} \sum_{j=1}^N \frac{\partial^2 \mu(x_j)}{\partial x^k \partial \alpha'} \right) \Sigma_{3D\alpha} + \Sigma_{3DD},
 \end{aligned} \tag{11}$$

$$\begin{aligned}
 \omega_{4k} = & \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \right] \Sigma_{\beta} E \left[\frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right] \right) + \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha'} \right] \Sigma_{\alpha} E \left[\frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right] \right) \\
 & + 2 \left(E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \right] \Sigma_{\beta \alpha} E \left[\frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right] \right) + 2 E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta'} \right] \Sigma_{q\beta} \\
 & + 2 E \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha'} \right] \Sigma_{q\alpha} + \Sigma_{\mu}.
 \end{aligned} \tag{12}$$

where $\Sigma_{1D\beta}$ is the covariance between the sample average of $g_i(x, \beta, \alpha)$ and $\hat{\beta}$, $\Sigma_{1D\alpha}$ is the covariance between the sample average of $g_i(x, \beta, \alpha)$ and $\hat{\alpha}$, Σ_{1DD} is the sample variance of $g_i(x, \beta, \alpha)$, $\Sigma_{2D\beta}$ is the covariance between the sample average of $\partial g_i(x, \beta, \alpha) / \partial x^k$ and $\hat{\beta}$, $\Sigma_{2D\alpha}$ is the covariance between the sample average of $\partial g_i(x, \beta, \alpha) / \partial x^k$ and $\hat{\alpha}$, Σ_{2DD} is the sample variance of $\partial g_i(x, \beta, \alpha) / \partial x^k$, $\Sigma_{3D\beta}$ and $\Sigma_{3D\alpha}$ are the covariances between the sample average of $N^{-1} \sum_j \partial g_i(x_j, \beta, \alpha) / \partial x^k$ and $\hat{\beta}$ and $\hat{\alpha}$, Σ_{3DD} is the sample variance of $N^{-1} \sum_j \partial g_i(x_j, \beta, \alpha) / \partial x^k$, $\Sigma_{q\beta}$ is the covariance between the sample average of $q_i(\beta, \alpha)$ and $\hat{\beta}$, $\Sigma_{q\alpha}$ is the covariance between the sample average of $q_i(\beta, \alpha)$ and $\hat{\alpha}$, and Σ_{μ} is the sample variance of $q_i(\beta, \alpha)$.

Step I: Apply nonlinear least squares, or ordinary least squares if $h(x, \beta)$ is linear in the parameters, to the logarithm of y_i on $h(x_i, \beta)$. Compute the scalar predicted value $h(x, \hat{\beta})$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix as $\hat{\Sigma}_{\beta}$. Save the residuals $\hat{\varepsilon}_i$, but do not square them (same as Section 2.4.).

Step II: Apply nonlinear least squares, or ordinary least squares if $s(x, \alpha)$ is linear in the parameters, to $\hat{\varepsilon}_i^2$ on $s(x_i, \alpha)$. Compute the scalar predicted value $s(x, \hat{\alpha})$ for some fixed x . Save the heteroscedasticity-consistent covariance matrix

as $\hat{\Sigma}_\alpha$. Save the residuals as $\hat{\eta}_i$ (the only difference with Step II in Section 2.3 is the addition of the last step).

Step III: Compute $\hat{\mu}(x)$ and its derivatives using the function $\hat{g}(x, \hat{\beta}, \hat{\alpha})$. To estimate the vectors $\Sigma_{1D\beta}$, $\Sigma_{1D\alpha}$, $\Sigma_{2D\beta}$, $\Sigma_{2D\alpha}$, $\Sigma_{3D\beta}$, $\Sigma_{3D\alpha}$, $\Sigma_{q\beta}$, and $\Sigma_{q\alpha}$, run simple regressions and save the regression coefficients. To estimate the variances Σ_{1DD} , Σ_{2DD} , Σ_{3DD} , and Σ_μ , compute the sample variances and divide by N (see Appendix A for all formulas). Estimate $\Sigma_{\beta\alpha}$ using the formula in Appendix A.

Step IV: Plug the estimated values from Step III into Eqs. (9)–(12).

2.6. Bootstrapping

Bootstrapping is an alternative method of computing variances. The bootstrapping method has two advantages compared to the delta methods — it gives a better approximation to the finite sample distribution in small samples and it does not require computing derivatives. So far, we have relied on the large sample approximation to the finite sample distribution of three estimands to compute confidence intervals. It is well documented that the asymptotic distribution may be a poor approximation to the finite sample distribution when the sample size is small. Even if the asymptotic distribution is a good approximation, the analytical method requires computing many first and second partial derivatives, which can be more complicated and error-prone than bootstrapping.

The disadvantage of bootstrapping is that it is very costly in terms of computer time when nonlinear least squares estimation is involved. A rule of thumb is that it takes about 500 iterations to estimate standard errors, which may be extremely time-consuming and even more if nonnormality is suspected. The choice between bootstrap and formula to compute the standard errors depends on the data set and the model. The important thing is to compute the standard errors.

3. Empirical example

3.1. Data

We illustrate the methods by analyzing Massachusetts Medicaid data, which were collected by the Massachusetts Division of Medical Assistance. The Medicaid data set includes information on all Massachusetts Medicaid claims from fiscal years 1991 and 1992. We limited the sample to the 9863 persons who were diagnosed with a severe mental illness and had at least one Medicaid claim in both years.

The dependent variable is the logarithm of total expenditures, which includes inpatient, outpatient, and pharmacy. The mean expenditure was just over

Table 1
Summary statistics for data on mental health care ($N = 9863$)

Variable	Minimum	Mean	Maximum	Standard deviation	Skewness
<i>Dependent variables</i>					
Expenditures	5	16,031	273,949	24,440	3.45
ln(Expenditures)	1.55	8.80	12.52	1.42	-0.19
<i>Prior health care use</i>					
Lagged expenditures	2	15,364	334,725	24,686	4.13
ln(Lagged expenditures)	0.78	8.75	12.72	1.41	-0.14
<i>Demographics</i>					
Age	18.0	41.1	64.0	11.7	
Male	0	0.415	1	0.493	
African-American	0	0.077	1	0.267	
Hispanic	0	0.020	1	0.139	
White	0	0.903	1	0.295	
<i>Health status</i>					
Schizophrenia	0	0.537	1	0.499	
Major affective disorder	0	0.412	1	0.492	
Other psychoses	0	0.050	1	0.218	
Substance abuse	0	0.100	1	0.301	

US\$16,000, but ranged from essentially zero to US\$274,000 (see Table 1). One of the explanatory variables of interest was lagged expenditures, which had a similar distribution. Taking the logarithm of the dependent variable removed much of the skewness, although the skewness is negative and significantly different than zero.

We also controlled for the standard demographic characteristics and health status. The patient population was 41.5% male, 89.5% white, 7.7% black, and 2% other race. The mean age was 41 years, and ranged from 18 to 64. All of the patients in the sample had one of the following diagnoses during the year: schizophrenia (53.7%), major affective disorder (41.3%), or other psychoses (5.0%). Schizophrenia is considered the most serious of these conditions, followed by major affective disorder and other psychoses. Substance abuse is a comorbidity strongly associated with mental health problems. Ten percent of our sample have a substance use comorbidity.

3.2. Results

The regression results are for illustrative purposes only. We treat the independent variables x as random because the experiment is not controlled. Also, because the sample is relatively large, we believe that the normal distribution is a

good approximation to the finite sample distribution of the three estimands, and therefore do not report bootstrapping results. If the sample were relatively small, then the assumption of the normal distribution would be questionable, and bootstrapping would be preferred.

The results show, not surprisingly for this population with chronic disease, that logged annual health care expenditures are higher when lagged expenditures are higher (see Table 2). Expenditures are also higher for persons who are younger, female, white, schizophrenic, and have substance abuse comorbidity, results that are consistent with the literature. The heteroscedasticity regression, which predicts the squared residuals, has a low adjusted R^2 , but the estimated parameters of most of the explanatory variables are statistically significant. The LM form of the White (1980) heteroscedasticity test, which regresses the squared estimated error term on the independent variables and their squares and interaction terms, is 106.5, leading to rejection of the null hypothesis of homoscedasticity (see second column of Table 2). For this example, we do not include squared terms or interaction terms, but it would be straightforward to do so. The results are shown for specific values of x , namely a 40-year-old white female with schizophrenia and no substance abuse comorbidity. The derivatives with respect to x^k are for the continuous variable age.

Table 2

Regression results for linear model to predict logged health care expenditures

Variable	OLS: $\ln(\text{Expenditure})$	OLS: $\hat{\varepsilon}^2$
Constant	3.39 (0.11) * *	2.23 (0.26) * *
<i>Prior health care use</i>		
$\ln(\text{Lagged expenditures})$	0.6421 (0.0097) * *	-0.077 (0.026) * *
<i>Demographics</i>		
Age	-0.00374 (0.00096) * *	-0.0123 (0.0020) * *
Male	-0.105 (0.023) * *	0.129 (0.049) * *
African-American	-0.055 (0.045)	0.333 (0.095) * *
Hispanic	-0.133 (0.064) *	-0.315 (0.098) * *
<i>Health status</i>		
Major affective disorder	-0.126 (0.022) * *	-0.056 (0.047)
Other psychoses	-0.132 (0.056) *	0.25 (0.13)
Substance abuse	0.542 (0.035) * *	-0.108 (0.068)
N	9863	9863
Adjusted R^2	0.45	0.01

The reference category is a white female with schizophrenia and no substance abuse comorbidity. Robust standard errors are corrected for heteroscedasticity using Huber–White robust standard errors.

* Statistically significant at the 5% level.

* * Statistically significant at the 1% level.

Table 3

Computation of the conditional mean of y , the slope of y , the average slope of y , and their standard errors for the normal linear model

	Assume hetero- scedasticity	Estimate	Standard error
$f(x) = E[y x]$	No	11,884	264
	Yes	11,667	314
$\partial f(x) / \partial x^k = \partial E[y x] / \partial x^k$	No	-44.5	11.7
	Yes	-115.4	17.1
$E[\partial f(x) / \partial x^k] = E[\partial E[y x] / \partial x^k]$	No	-66.6	24.6
	Yes	-170.6	24.5

Calculations are for a 40-year-old white female with schizophrenia and no substance abuse comorbidity. Derivatives are with respect to age.

The three estimands are calculated both with and without the assumption of heteroscedasticity to show the importance of controlling for it in the normal linear model (see Table 3). The results show standard errors that are relatively small when compared to the magnitude of the three estimates. The estimated expenditures are slightly less than US\$12,000, with a standard error of US\$264 under homoscedasticity. The slope and the average slope in particular are quite different in their point estimates. For example, under homoscedasticity an increase in age from 40 to 41 for the base-case person would decrease expected expenditures by US\$44.50, but under heteroscedasticity it would decrease by US\$115.40. In our example, the estimates of the slope and average slope are biased towards zero

Table 4

Computation of the conditional mean of y , the slope of y , the average slope of y , and their standard errors for the nonparametric model

	Assume hetero- scedasticity	Estimate	Standard error
$f(x) = E[y x]$	No	12,296	327
	Yes	12,044	300
$\partial f(x) / \partial x^k = \partial E[y x] / \partial x^k$	No	-46.0	12.0
	Yes	-131.1	16.9
$E[\partial f(x) / \partial x^k] = E[\partial E[y x] / \partial x^k]$	No	-60.6	15.5
	Yes	-144.9	20.2

Calculations are for a 40-year-old white female with schizophrenia and no substance abuse comorbidity. Derivatives are with respect to age.

when heteroscedasticity is ignored, and the bias is large enough that the estimate without controlling for heteroscedasticity falls outside of the 95% confidence interval. This example shows the importance of controlling for heteroscedasticity both in the point estimate and in the standard error.

The three estimands are also calculated for the nonparametric model (see Table 4). The pattern of results is similar. Controlling for heteroscedasticity is clearly important for both the point estimates and the standard errors. The point estimates for the slope and average slope are much larger in absolute value after controlling for heteroscedasticity. The standard errors for all three statistics are quite different after controlling for heteroscedasticity, and in two of the cases are larger.

In this example, however, the difference between the results for the normal and nonparametric models is small. Comparisons between Tables 3 and 4 show qualitatively similar results. In our example, the residuals are approximately normal, so the nonparametric model is not different from the normal model. In other data sets, however, the nonparametric model may be more appropriate and could give quite different results.

4. Conclusion

The use of the log-transformed dependent variable in applied economics creates a potential bias when computing estimates of $E[y|x]$ on the original scale, if the error term either does not have a normal distribution or is heteroscedastic. Estimates on the original scale should be reported with standard errors, like all estimated values. One reason that the calculation of standard errors is not common is that the equations are not commonly known. This paper provides equations for the general case of error terms that have any distribution and are heteroscedastic, as well as for simpler cases. Another reason is that computing the standard errors is not automatically done in software. The authors will provide sample programs in Stata to compute the standard errors upon request.

Acknowledgements

Chunrong Ai received financial support through a summer grant from the Warrington College of Business Administration. We would like to thank Willard Manning, John Mullahy, and two anonymous referees for their comments on an earlier draft. An earlier version of this paper was presented at the Second World Conference of the International Health Economics Association, Rotterdam, the Netherlands, June 6–9, 1999.

Appendix A

A.1. Formulas for the normal linear model

For the normal linear model the formulas for Step III are:

$\hat{\mu}(x) = e^{x' \hat{\beta} + 0.5x' \hat{\alpha}}$	scalar
$\hat{\beta}_k, \hat{\alpha}_k$	scalars, estimated parameters for x_k
e_k	vector of 0s, with 1 in the k th position
$\frac{\partial \hat{\mu}(x)}{\partial x^k} = \hat{\mu}(x)(\hat{\beta}_k + 0.5\hat{\alpha}_k)$	scalar
$\frac{\partial \hat{\mu}(x)}{\partial \beta} = \hat{\mu}(x)x$	vector
$\frac{\partial \hat{\mu}(x)}{\partial \alpha} = 0.5\hat{\mu}(x)x$	vector
$\frac{\partial^2 \hat{\mu}(x)}{\partial \beta \partial x^k} = \frac{\partial \hat{\mu}(x)}{\partial x^k} x + \hat{\mu}(x)e_k$	vector
$\frac{\partial^2 \hat{\mu}(x)}{\partial \alpha \partial x^k} = 0.5 \left(\frac{\partial \hat{\mu}(x)}{\partial x^k} x + \hat{\mu}(x)e_k \right)$	vector
$\hat{E} \left[\frac{\partial^2 \mu(x)}{\partial \beta \partial x^k} \right] = \frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \hat{\mu}(x_i)}{\partial \beta \partial x^k}$	vector, simple average
$\hat{E} \left[\frac{\partial^2 \mu(x)}{\partial \alpha \partial x^k} \right] = \frac{1}{N} \sum_{i=1}^N \frac{\partial^2 \hat{\mu}(x_i)}{\partial \alpha \partial x^k}$	vector, simple average

A.2. Formulas for the normal nonlinear model

For the normal nonlinear model the formulas for Step III are:

$\hat{\mu}(x) = e^{(h(x, \hat{\beta}) + 0.5s(x, \hat{\alpha}))}$	scalar
$\frac{\partial \hat{\mu}(x)}{\partial \beta} = \hat{\mu}(x) \left[\frac{\partial h(x, \hat{\beta})}{\partial \beta} \right]$	vector
$\frac{\partial \hat{\mu}(x)}{\partial \alpha} = \hat{\mu}(x) \left[0.5 \frac{\partial s(x, \hat{\alpha})}{\partial \alpha} \right]$	vector
$\frac{\partial \hat{\mu}(x)}{\partial x^k} = \hat{\mu}(x) \left[\frac{\partial h(x, \hat{\beta})}{\partial x^k} + 0.5 \frac{\partial s(x, \hat{\alpha})}{\partial x^k} \right]$	scalar

$$\frac{\partial^2 \hat{\mu}(x)}{\partial x^k \partial \beta'} = \frac{\partial \hat{\mu}(x)}{\partial \beta} \left[\frac{\partial h(x, \hat{\beta})}{\partial x^k} + 0.5 \frac{\partial s(x, \hat{\alpha})}{\partial x^k} \right] + \hat{\mu}(x) \frac{\partial^2 h(x, \hat{\beta})}{\partial x^k \partial \beta}$$

vector

$$\frac{\partial^2 \hat{\mu}(x)}{\partial x^k \partial \alpha'} = \frac{\partial \hat{\mu}(x)}{\partial \alpha} \left[\frac{\partial h(x, \hat{\beta})}{\partial x^k} + 0.5 \frac{\partial s(x, \hat{\alpha})}{\partial x^k} \right] + 0.5 \hat{\mu}(x) \frac{\partial^2 s(x, \hat{\alpha})}{\partial x^k \partial \alpha}$$

vector

A.3. Formulas for the nonparametric model with homoscedasticity

For the nonparametric model with homoscedasticity the formulas for Step III are:

$$\frac{\partial^2 \hat{\mu}(x)}{\partial \beta} = \hat{\mu}(x) \frac{\partial h(x, \hat{\beta})}{\partial \beta} - e^{h(x, \hat{\beta})} \frac{1}{N} \sum_{i=1}^N \left(e^{\hat{\epsilon}_i} \frac{\partial h(x_i, \hat{\beta})}{\partial \beta} \right) \quad \text{vector}$$

The second term in the first derivative above arises from the fact that the residuals $\hat{\epsilon}_i$ are estimated and hence depend on the parameter estimate $\hat{\beta}$.

$$\frac{\partial \hat{\mu}(x)}{\partial x^k \partial \beta} = \hat{\mu}(x) \frac{\partial^2 h(x, \hat{\beta})}{\partial x^k \partial \beta} + \frac{\partial \hat{\mu}(x)}{\partial \beta} \frac{\partial h(x, \hat{\beta})}{\partial x^k} \quad \text{vector}$$

$$\hat{E} \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta} \right] = \frac{1}{N} \sum_{i=1}^N y_i \frac{\partial^2 h(x_i, \hat{\beta})}{\partial x^k \partial \beta} \quad \text{vector}$$

Four of the Σ 's are estimated by running a simple regression, saving the vector of regression coefficients, and dividing them by N . The part of the dependent variable that is not $\hat{\epsilon}_i$ or $\hat{\eta}_i$ must have its mean subtracted before multiplying by the estimated error.

Estimated coefficients	Dependent variable	Independent variable
$\hat{\Sigma}_{1D\beta}$	$\exp(h(x, \hat{\beta}) + \hat{\epsilon}_i) \times \hat{\epsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{2D\beta}$	$\exp(h(x, \hat{\beta}) + \hat{\epsilon}_i) \times \partial h(x, \hat{\beta}) / \partial x^k \times \hat{\epsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{3D\beta}$	$[N^{-1} \sum_j \exp(h(x_j, \hat{\beta}) + \hat{\epsilon}_i) \times \partial h(x_j, \hat{\beta}) / \partial x^k] \times \hat{\epsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{\mu\beta}$	$y_i \times [\partial h(x_i, \hat{\beta}) / \partial x^k] \times \hat{\epsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$

Four of the Σ 's are estimated by computing the sample variance divided by N (scalar).

Estimated variance	Sample variance of
$\hat{\Sigma}_{1DD}$	$\exp(h(x, \hat{\beta}) + \varepsilon_i)$
$\hat{\Sigma}_{2DD}$	$\exp(h(x, \hat{\beta}) + \varepsilon_i)(\partial h(x, \hat{\beta})/\partial x^k)$
$\hat{\Sigma}_{3DD}$	$N^{-1}\sum_j \exp(h(x_j, \hat{\beta}) + \varepsilon_i) \times \partial h(x_j, \hat{\beta})/\partial x^k$
$\hat{\Sigma}_\mu$	$y_i(\partial h(x_i, \hat{\beta})/\partial x^k)$

A.4. Formulas for the nonparametric model with heteroscedasticity

For the nonparametric model with heteroscedasticity the formulas for Step III are:

$$\frac{\partial \hat{\mu}(x)}{\partial \beta} = \frac{1}{N} \sum_{i=1}^N \frac{\partial g_i(x, \hat{\beta}, \hat{\alpha})}{\partial \beta} \quad \text{vector}$$

$$\frac{\partial \hat{\mu}(x)}{\partial \alpha} = \frac{1}{N} \sum_{i=1}^N \frac{\partial g_i(x, \hat{\beta}, \hat{\alpha})}{\partial \alpha} \quad \text{vector}$$

$$\frac{\partial^2 \hat{\mu}(x)}{\partial x^k \partial \beta} = \frac{1}{N} \sum_{i=1}^N \frac{\partial^2 g_i(x, \hat{\beta}, \hat{\alpha})}{\partial x^k \partial \beta} \quad \text{vector}$$

$$\frac{\partial^2 \hat{\mu}(x)}{\partial x^k \partial \alpha} = \frac{1}{N} \sum_{i=1}^N \frac{\partial^2 g_i(x, \hat{\beta}, \hat{\alpha})}{\partial x^k \partial \alpha} \quad \text{vector}$$

$$\hat{E} \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \beta} \right] = \frac{1}{N} \sum_{i=1}^N \left(\frac{\partial q_i(\hat{\beta}, \hat{\alpha})}{\partial \beta} \right) \quad \text{vector}$$

$$\hat{E} \left[\frac{\partial^2 \mu(x)}{\partial x^k \partial \alpha} \right] = \frac{1}{N} \sum_{i=1}^N \left(\frac{\partial q_i(\hat{\beta}, \hat{\alpha})}{\partial \alpha} \right) \quad \text{vector.}$$

$$\hat{\Sigma}_{\beta\alpha} = \left(\sum_{i=1}^N \frac{\partial h(x_i, \hat{\beta})}{\partial \beta} \frac{\partial h(x_i, \hat{\beta})}{\partial \beta'} \right)^{-1} \left(\sum_{i=1}^N \frac{\partial h(x_i, \hat{\beta})}{\partial \beta} \hat{\varepsilon}_i \hat{\eta}_i \frac{\partial s(x_i, \hat{\alpha})}{\partial \alpha'} \right) \\ \times \left(\sum_{i=1}^N \frac{\partial s(x_i, \hat{\alpha})}{\partial \alpha} \frac{\partial s(x_i, \hat{\alpha})}{\partial \alpha'} \right)^{-1} \quad \text{matrix}$$

Eight of the Σ 's are estimated by running a simple regression, saving the vector of regression coefficients, and dividing them by N . The part of the

dependent variable that is not $\hat{\varepsilon}_i$ or $\hat{\eta}_i$ must have its mean subtracted before multiplying by the estimated error.

Estimated coefficient	Dependent variable	Independent variable
$\hat{\Sigma}_{1D\beta}$	$g_i(x, \hat{\beta}, \hat{\alpha}) \times \hat{\varepsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{1D\alpha}$	$g_i(x, \hat{\beta}, \hat{\alpha}) \times \hat{\eta}_i$	$\partial s(x_i, \hat{\alpha}) / \partial \alpha$
$\hat{\Sigma}_{2D\beta}$	$(\partial g_i(x, \hat{\beta}, \hat{\alpha}) / \partial x^k) \times \hat{\varepsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{2D\alpha}$	$(\partial g_i(x, \hat{\beta}, \hat{\alpha}) / \partial x^k) \times \hat{\eta}_i$	$\partial s(x_i, \hat{\alpha}) / \partial \alpha$
$\hat{\Sigma}_{3D\beta}$	$(N^{-1} \sum_j \partial g_i(x_j, \hat{\beta}, \hat{\alpha}) / \partial x^k) \times \hat{\varepsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{3D\alpha}$	$(N^{-1} \sum_j \partial g_i(x_j, \hat{\beta}, \hat{\alpha}) / \partial x^k) \times \hat{\eta}_i$	$\partial s(x_i, \hat{\alpha}) / \partial \alpha$
$\hat{\Sigma}_{q\beta}$	$q_i(\hat{\beta}, \hat{\alpha}) \times \hat{\varepsilon}_i$	$\partial h(x_i, \hat{\beta}) / \partial \beta$
$\hat{\Sigma}_{q\alpha}$	$q_i(\hat{\beta}, \hat{\alpha}) \times \hat{\eta}_i$	$\partial s(x_i, \hat{\alpha}) / \partial \alpha$

Four of the Σ 's are estimated by computing the sample variance divided by N (scalar).

Estimated variance	Sample variance of
$\hat{\Sigma}_{1DD}$	$g_i(x, \hat{\beta}, \hat{\alpha})$
$\hat{\Sigma}_{2DD}$	$\partial g_i(x, \hat{\beta}, \hat{\alpha}) / \partial x^k$
$\hat{\Sigma}_{3DD}$	$N^{-1} \sum_j \partial g_i(x_j, \hat{\beta}, \hat{\alpha}) / \partial x^k$
$\hat{\Sigma}_{\mu}$	$q_i(\hat{\beta}, \hat{\alpha})$

A.5. Proof of asymptotic independence

Let $\hat{\beta}$ denote the least squares estimator of the regression

$$\ln(y_i) = h(x_i, \beta) + \varepsilon_i \quad i = 1, 2, \dots, N.$$

Then applying a Taylor expansion, we obtain

$$\hat{\beta} - \beta = \left(\sum_{i=1}^N \frac{\partial h(x_i, \beta)}{\partial \beta} \frac{\partial h(x_i, \beta)}{\partial \beta'} \right)^{-1} \sum_{i=1}^N \frac{\partial h(x_i, \beta)}{\partial \beta} \varepsilon_i + o_p(N^{-0.5}).$$

See Amemiya (1985) for a derivation. Let $\hat{\varepsilon}_i$ denote the least squares residuals. Note that

$$E[\varepsilon_i^2 | x_i] = s(x_i, \alpha).$$

We can write

$$\varepsilon_i^2 = s(x_i, \alpha) + \eta_i$$

where η_i is the error term. Let $\hat{\alpha}$ denote the least squares estimator from

$$\hat{\varepsilon}_i^2 = s(x_i, \alpha) + \eta_i^*$$

Applying the first-order Taylor series approximation, we obtain

$$\hat{\alpha} - \alpha = \left(\sum_{i=1}^N \frac{\partial s(x_i, \alpha)}{\partial \alpha} \frac{\partial s(x_i, \alpha)}{\partial \alpha'} \right)^{-1} \sum_{i=1}^N \frac{\partial s(x_i, \alpha)}{\partial \alpha} \eta_i + o_p(N^{-0.5}).$$

Note that the covariance between $\hat{\beta}$ and $\hat{\alpha}$ depends on $E[\varepsilon_i \eta_i | x_i] = E[\varepsilon_i^3 \eta_i | x_i]$. For normal ε_i , $E[\varepsilon_i^3 | x_i] = 0$, implying that $\hat{\beta}$ and $\hat{\alpha}$ are asymptotically independent.

References

- Amemiya, T., 1985. *Advanced Econometrics*. Harvard Univ. Press, Cambridge, MA.
- Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. *J. Am. Stat. Assoc.* 78, 605–610.
- Duan, N. et al., 1983. A comparison of alternative models for the demand for medical care. *J. Bus. Econ. Stat.* 1, 115–126.
- Duan, N. et al., 1984. Choosing between the sample selection model and the multi-part model. *J. Bus. Econ. Stat.* 2, 283–289.
- Greene, W.H., 2000. *Econometric Analysis*. Prentice-Hall, Upper Saddle River, NJ.
- Jones, A.M., Yen, S.T., 1999. A Box–Cox Double-Hurdle Model. Manchester School, in press.
- Manning, W.G., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *J. Health Econ.* 17, 283–296.
- Manning, W.G., Mullahy, J., 1999. Estimating Log Models: To Transform or not to Transform. University of Chicago working paper.
- Mullahy, J., 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J. Health Econ.* 17, 247–282.
- White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–830.