AMERICAN UNIVERSITY
WASHINGTON, DC

Department of Economics
Working Paper Series

Heteroskedasticity-Robust Elasticities in Logarithmic and Two-Part Models

By:

Tom Hertz
Department of Economics, American University

No. 2007-19

http://www.american.edu/academic.depts/cas/econ/workingpapers/workpap.htm

Heteroskedasticity-Robust Elasticities in Logarithmic and Two-Part Models

Abstract

Logarithmic models are widely used to study highly skewed positive outcomes, either alone or in combination with an equation that first distinguishes between zero and non-zero values (the two part model). A well-known drawback of such models is that to obtain marginal effects that pertain to the arithmetic mean, rather than the mean of logs, we must exponentiate, and this retransformation is complicated in the presence of heteroskedasticity. This paper presents a simple method for correcting estimated elasticities for the effects of heteroskedasticity, in both log-linear and log-log (constant elasticity) equations. An example, drawing on Bulgarian farm survey data, demonstrates that this correction leads to significantly different estimates of the elasticity of expenditures on agricultural inputs with respect to land area and the age of the household head.

Length, excluding title page and abstract, but including all else: 1996 words

Please address correspondence to:
Prof. Tom Hertz
Dept. of Economics, American University,
4400 Mass. Ave. NW, Washington, DC 20016, USA.
Email: hertz@american.edu
Tel. (202) 885-2756
Fax (202) 885-3790
Word count: 7573 (text and references); 1655 (tables and figures).

1. Introduction

Few functional forms are as widely used in the social sciences to study highly skewed positive outcomes as is the logarithmic specification, which reduces the leverage exerted by extreme values, producing more robust and often more efficient estimates (Deb, Manning, and Norton 2005). However, many outcomes have distributions that are characterized not only by a long right tail, but also by a large spike at zero, for which the log is undefined. Examples include expenditures on consumer durables and investment goods, number of cigarettes smoked, and countless others. Two-part models, the first part consisting of a probit, logit, or linear equation to distinguish between zero and positive values, and the second part using OLS on the logs of the positives, are well-suited to such outcomes. Unlike the Tobit, the two-part model allows the use of logs, and also allows different parameters to determine the two parts of the data-generating process. Two-part models are also arguably more appropriate than Heckman-style selection-bias models when the zero values represent outcomes of interest, rather than censored values of a latent variable, or data that have been made missing by taking the log of zero (Duan*, et al.* 1984).

Yet the log form has the drawback of generating estimates of marginal effects on the conditional mean of the log of the outcome, not on the conditional mean of the outcome itself, which is usually of more interest. Although many economists ignore this difference, some, primarily health economists, have taken these concerns seriously, and have developed retransformation techniques that produce consistent estimates of the desired expected values, and marginal effects. This is straightforward if the error terms in the logged equation are homoskedastically normal, but is more complicated in the presence of non-normality, and, especially, heteroskedasticity (Duan 1983; Mullahy 1998).

This paper describes an approach to this problem that is applicable when we wish to express our marginal effects as elasticities. I use a procedure suggested, but not pursued explicitly, by Mullahy (1998), to test the proposition that the retransformation problem is ignorable, and to correct the estimated elasticities if it is not. An empirical example, based on a model of expenditures on variable inputs by Bulgarian family farmers, demonstrates that these corrections may be non-trivial.

2. The Retransformation Problem

Consider a semi-log equation of the following form, defined only for $y > 0$:

[1] $$\ln y = X\beta + u \quad \text{with} \quad E(u \mid X) = 0 .$$

To find the conditional mean of y, as opposed to $\ln y$, we first take antilogs, then expectations:

[2] $$E(y \mid y > 0, X) = e^{X\beta} E(e^{u} \mid X)$$

If u is normal and homoskedastic, with variance σ^2, then $E(e^{u} \mid X) = \sigma^{2}/2$, but Duan (1983) notes that violations of this assumption render this approximation inadequate. Duan's solution is to estimate $E(e^{u} \mid X)$ using $n^{-1}\sum e^{\hat{u}}$, the average of the antilogged residuals from [1]. This estimator makes no assumptions about the distribution of u, performs well for non-normal errors, and can accommodate heteroskedasticity, provided it is not related to the covariates of interest.

To illustrate why this last proviso is important, take the partial derivative of [2] with respect to some x_k; the matrix X now contains the remaining covariates.

[3] $$\frac{\partial E(y \mid y > 0, X, x_k)}{\partial x_k} = \beta_k e^{X\beta} E(e^{u} \mid X, x_k) + e^{X\beta} \frac{\partial E(e^{u} \mid X, x_k)}{\partial x_k} .$$

If the second term of this expression is zero, then Duan's estimator of the expectation in the first term suffices. But if u is heteroskedastic in x_k, the latter term will generally not be zero.

The algebra is simpler if we are interested in the elasticity of y with respect to x_k $(\varepsilon_{x_k}^{y})$:

$$\text{[4]} \qquad \varepsilon_{x_k}^{y} = \beta_k x_k + \frac{\partial E(e^u \mid X, x_k)}{\partial x_k} \frac{x_k}{E(e^u \mid X, x_k)}$$

Note that if $\frac{\partial E(e^u \mid X, x_k)}{\partial x_k} = 0$, then Duan's estimator is not needed: in calculating elasticities, under the assumption of homoskedasticity (or the stronger assumption of the independence of x_k and u), there is no retransformation problem, a statement which does not apply to the marginal effects of equation [3]. However, if u is heteroskedastic in x_k, then the second term in [4] again cannot be ignored.

If the covariate of interest enters in log form ($\ln x_k$) then [4] becomes:

$$\text{[4']} \qquad \bar{\varepsilon}_{x_k}^{y} = \beta_k + \frac{\partial E(e^u \mid X, \ln x_k)}{\partial x_k} \frac{x_k}{E(e^u \mid X, \ln x_k)}$$

Now x_k drops out of the first term, and under the model for $E(e^u \mid X, x_k)$ that we will soon describe, the second term is constant with respect to $[X, x_k]$ as well. The log-log model thus produces a constant elasticity with respect to a logged covariate ($\bar{\varepsilon}_{x_k}^{y}$) even after correcting for heteroskedasticity. By contrast, both parts of [4] depend on $[X, x_k]$; in estimating them the mean elasticity across observations is reported.

Next consider a two-part model, whose first equation models the probability of a positive outcome, using a probit. Two versions are presented, first with x_k in levels and then again with x_k in logs:

[5] $$\Pr(y > 0 \mid X, x_k) = \Phi(X\gamma + \gamma_k x_k),$$

[5’] $$\Pr(y > 0 \mid X, x_k) = \Phi(X\gamma + \gamma_k \ln x_k),$$ where Φ is the cumulative standard normal. For the positive values, we have:

[6] $$\ln y = X\beta + \beta_k x_k + u \qquad \text{and}$$

[6’] $$\ln y = X\beta + \beta_k \ln x_k + u \qquad \text{with} \qquad E(u \mid X, x_k) = 0.$$

The expected value of *y* is then:

[7] $$E(y \mid X, x_k) = \Pr(y > 0 \mid X, x_k) * E(y \mid y > 0, X, x_k),$$

and differentiating this with respect to x_k allows us to compute the elasticity of interest $(\Lambda^y_{x_k})$. For a covariate entering in levels [8], or logs [8’], we get:

[8] $$\Lambda^y_{x_k} \equiv \frac{\partial E(y \mid X)}{\partial x_k} \frac{x_k}{E(y \mid X, x_k)} = \gamma_k x_k \frac{\phi(X\gamma + \gamma_k x_k)}{\Phi(X\gamma + \gamma_k x_k)} + \varepsilon^y_{x_k}$$

[8’] $$\overline{\Lambda}^y_{x_k} \equiv \frac{\partial E(y \mid X)}{\partial x_k} \frac{x_k}{E(y \mid X, x_k)} = \gamma_k \frac{\phi(X\gamma + \gamma_k \ln x_k)}{\Phi(X\gamma + \gamma_k \ln x_k)} + \bar{\varepsilon}^y_{x_k}$$

The first term of the rightmost expression is the elasticity of the probability of *y*’s being positive, with respect to x_k,[1] and the second terms are as in [4] and [4’].

There are thousands of examples of equations like [6] or [6’] in the economics literature whose authors interpret the coefficients β_k, or $\beta_k \bar{x}_k$ in log-linear settings, as elasticities. However, both are elasticities of the conditional mean of the log of *y*, not of *y* itself, with respect to x_k. The difference is that between $\frac{\partial \ln E(y \mid X)}{\partial \ln x_k}$, which is equivalent to [8], and $\frac{\partial E(\ln y \mid X)}{\partial \ln x_k}$, which is not. Wooldridge (2002, p. 17) notes that “For the most part, little is lost by treating [the two] as the same when $y > 0$.” Yet to do so is to ignore the retransformation problem. Moreover, the first term in [8] pertains to a conditional mean probability, not a conditional mean log probability. Thus even if we were content to speak of elasticities at the mean of logs for the positive values, to avoid the retransformation problem, this would still be inconsistent with the way the elasticity of the probability of positive outcomes is defined.

The last step is to estimate $\frac{\partial E(e^u \mid X, x_k)}{\partial x_k}$. Mullahy (1998) suggests (pp. 15-16) that since e^u is necessarily positive, it makes sense to model its conditional expectation exponentially:

[9] $$E(e^u \mid X, x_k) = e^{(X\lambda + \lambda_k x_k)}$$

[9'] $$E(e^u \mid X, \ln x_k) = e^{(X\lambda + \lambda_k \ln x_k)}$$

Estimates of λ and λ_k may be generated by non-linear least-squares regression of the antilogged residuals $(e^{\hat{u}})$ against all covariates in the model, including a constant. Given this setup, λ_k is the final term in [8] or [8']:

[10] $$\Lambda^y_{x_k} = \hat{\gamma}_k x_k \frac{\phi(X\hat{\gamma} + \hat{\gamma}_k x_k)}{\Phi(X\hat{\gamma} + \hat{\gamma}_k x_k)} + \hat{\beta}_k x_k + \hat{\lambda}_k x_k \,,$$

[10'] $$\overline{\Lambda}^y_{x_k} = \hat{\gamma}_k \frac{\phi(X\hat{\gamma} + \hat{\gamma}_k \ln x_k)}{\Phi(X\hat{\gamma} + \hat{\gamma}_k \ln x_k)} + \hat{\beta}_k + \hat{\lambda}_k \,.$$

Mullahy does not calculate this estimator explicitly, but does present several related alternatives, and shows how estimates of the term $\hat{\lambda}_k$ effectively reconcile the differences between his estimators and the homoskedastic two-part model that uses Duan's scalar smearing adjustment. I prefer the approach outlined here because, unlike Mullahy's alternatives, equation [10] incorporates a direct estimate of $\hat{\lambda}_k$, and, in principle, a means for testing whether this parameter is small enough to ignore. That test rests on heteroskedasticity- and cluster-robust standard errors; however, Mullahy cautions that these have not been shown to be consistent for this application, and must be interpreted with care. As an alternative, I also present bootstrapped standard errors, for each component of [10] and [10'], and for their sum. Note that the bootstrap re-samples survey clusters, not households.

3. An Empirical Example

The data for this example come from a recent survey from Bulgaria, described in more detail in ([Self] 2007). A two-part model is used to predict total expenditures on variable inputs such as feed and herbicides, for a sample of 1206 family farms. Seventy-five percent (n=907) had positive expenditures, with a mean of $US 623, but with a highly skewed distribution (minimum $5, maximum $18,333). For this exercise, the covariates are the log of non-farm income, the log of land under cultivation, the number of farm implements owned, an indicator for households owning livestock, and the household head's age.

The first column of Table 1 reports the probit estimates of equation [5], expressed as marginal effects on $\Pr(y > 0)$, not elasticities. The next column reports estimates of β, from equation [6], for households with positive expenditures. All coefficients are of the same sign as the probit results: for example, older heads are less likely to purchase variable inputs, and spend less when they do. The final column reports the non-linear least squares estimates of λ, from equation [9]. We see significant effects of heteroskedasticity with respect to the log of land area, and age. For non-farm income, the variable of interest to [Self] (2007), the effect is nearly significant (p=0.145).

Table 2 presents the full and component elasticities for those three variables for which the effects of heteroskedasticity were significant, or nearly so. The first column translates the probit marginal effects into elasticities. The second does the same for the OLS equation (which affects only the non-logged covariate, head's age). Note that the bootstrapped standard errors for the logged variables are comparable to the conventional results in the previous table. Next come the estimates of $\hat{\lambda}_k$ for the logged variables (also the same as in the previous table) and of $\hat{\lambda}_k \bar{x}_k$ for head's age. All three are large in a practical sense: the heteroskedasticity correction knocks nine points off the elasticity of variable inputs with respect to non-farm income, reducing it to 0.179; for land area, the correction adds nearly seven points, raising the elasticity to 0.433; and for head's age, the correction lowers the elasticity by 41 points, to -0.987. Using the bootstrapped standard errors, these adjustments are statistically significant at the ten percent level for land area (p=0.080) and head's age (p=0.078) and nearly so for non-farm income (p=0.104).

This method offers a means for determining when the retransformation problem makes a material difference to the elasticities generated by single-equation models in logs, as well as for the two-part model. Subject to the concerns regarding the correct computation of standard errors, a large and significant result for $\hat{\lambda}_k$, or for $\hat{\lambda}_k \bar{x}_k$ when covariates appear in levels, is a sign that conventionally-derived elasticities may be significantly biased. The recommended correction is then easily implemented in most econometric software packages.

References

[Self]. 2007. "The Effect of Non-Farm Income on Investment in Bulgarian Family Farming." Unpublished paper. June.

Deb, Partha, Willard Manning, and Edward Norton. 2005. "Modeling Health Care Costs and Counts." Paper presented at the International Health Economics Association 2005 World Congress, Barcelona.

Duan, Naihua. 1983. "Smearing Estimate: A Nonparametric Retransformation Method." *Journal of the American Statistical Association*, 78(383):605-610.

Duan, Naihua, Jr. Willard G. Manning, Carl M. Morris, and Joseph P. Newhouse. 1984. "Choosing between the Sample Selection Model and the Multi-Part Model." *Journal of Business & Economic Statistics*, 2(3):283-289.

Mullahy, John. 1998. "Much Ado About Two: Reconsidering Retransformation and the Two Part Model in Health Econometrics " NBER, Technical Working Papers 228.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Table 1: Regression Results

Outcome:	Pr(Y>0)	$\ln y$	$e^{\hat{u}}$	
Estimate of:	dPr(Y>0)/dx	β	λ	
Covariates:				
Log Non-Farm Income	0.034	0.208	-0.086	ns
	(0.017)	(0.054)	(0.059)	
Log of Area of Land Planted	0.031	0.317	0.065	
	(0.011)	(0.035)	(0.037)	
Household Breeds Livestock	0.201	0.902	-0.233	ns
	(0.036)	(0.109)	(0.161)	
Number of Agricultural Implements	0.178	0.237	0.001	ns
	(0.052)	(0.047)	(0.048)	
FTE's of Farm Labour	0.147	0.340	0.012	ns
	(0.026)	(0.066)	(0.074)	
Head's Age	-0.002	-0.006	-0.007	
	(0.001)	(0.003)	(0.004)	
Constant		2.340	1.788	
		(0.507)	(0.654)	
Sample Size	1206	907	907	
Pseudo R-squared	0.21			
R-squared		0.39	0.37	

Note: Heteroskedasticity-robust and clustered standard errors in parentheses.

ns: Not significant at 10% or better.

Table 2: Elasticities for Selected Covariates

Elasticities of: **With respect to:**	Pr(Y>0)		$E(y\|X, y>0)$		$E(e^{\hat{u}}\|X, y>0)$		$\hat{E}^{y}_{x_k}$
Log Non-Farm Income	0.056	+	0.208	+	-0.086	=	0.179
	(0.031)		(0.053)		(0.053)		(0.078)
Log of Area of Land Planted	0.051	+	0.317	+	0.065	=	0.433
	(0.019)		(0.034)		(0.037)		(0.048)
Head's Age	-0.216	+	-0.361	+	-0.410	=	-0.987
	(0.103)		(0.200)		(0.233)		(0.334)

Note: Bootstrapped standard errors in parentheses, based on 500 repetitions.

Notes

[1] This is the elasticity of the expected number of (say) households that will have positive expenditures, with respect to x_k.