View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

Tests of Inference for Dummy Variables in Regressions with Logarithmic Transformed Dependent Variables

Jenny N. Lye and Joseph G. Hirschberg[1]

Department of Economics, University of Melbourne, Melbourne, Victoria, 3010, Australia.

May 02

Abstract:

The interpretation of dummy variables in regressions where the dependent variable is subject to a log transformation has been of continuing interest in economics. However, in the main, these earlier papers do not deal with the inferential aspects of the parameters estimated. In this paper we compare the inference implied by the hypotheses tested on the linear parameter estimated in the model and the tests applied to the proportional change that this parameter implies. An important element in this analysis is the asymmetry introduced by the log transformation. Suggestions are made for the appropriate test procedure in this case. Examples are presented from some common econometric applications of this model in the estimation of hedonic price models and wage equations.

Key Words: Hypothesis tests; lognormal distribution; measures of proportional change; wage equation; hedonic price model

J E L Codes: C12; C21; R15; J31

[1] Correspondence to: Joseph Hirschberg, Department of Economics, University of Melbourne, Melbourne, 3010, Australia, email: j.hirschberg@unimelb.edu.au.

1. Introduction

There are many examples of semilog models in which discrete variables are used as regressors (e.g. wage equations, hedonic price models). However, it is well known that the traditional interpretation of these dummy variables does not follow in the case of the log of the dependent variable. The percentage change in the level of the dependent variable is not equal to the coefficient of the dummy variable multiplied by 100 as it is in the case of continuous variables. Writing the semilogarithmic regression equation as

$$\log(y_t) = a + \sum_{i=1}^{k} b_i z_{it} + cD_t + u_t \qquad (1)$$

where z_i represent continuous variables with corresponding coefficient b_i, D is a dummy variable with coefficient c and u_t is distributed identically and independently with mean of zero and variance σ_u^2. Alternatively this model can be written as

$$y_t = \exp(a + \sum_{i=1}^{k} b_i z_{it} + cD_t)\exp(u_t) \qquad (1a)$$

Halvorsen and Palmquist (1980) and Kennedy (1981) note that the appropriate measure for the proportional effect on Y given D=1 is defined by g thus $y_{D=1} = (1+g)y_{D=0}$ and taking the logs of both sides we get $\log(y_{D=1}) = \log(1+g) + \log(y_{D=0})$ [2] thus the difference $\log(y_{D=1}) - \log(y_{D=0}) = \log(1+g) = c$ and we can solve for g as

$$g = \exp(c) - 1 \qquad (2)$$

And we can define c as

[2] As long as $g > 0$. Since the sign of g can be changed by the redefinition of the dummy variable this condition should pose no problem. In general, the estimated parameter for (1-D) is equal and of opposite sign of the estimate for c although the estimate of the constant will vary and one needs to be careful when using interaction of dummy variables since a redefinition will change the t-statistic on the non-interacted term.

$$c = \log(1+g) \tag{3}$$

2. Measures of the central tendency of *g*

By making the same assumption as Kennedy (1981), that u_t is normally distributed and $\hat{c}$ is the OLS estimate of *c*, we then have that $\hat{c}$ is distributed normally with an expected value of *c* and a variance of $\sigma_c^2 = \sigma_u^2 m^{cc}$ where m^{cc} is the diagonal element of the $(X'X)^{-1}$ matrix that corresponds to the parameter *c* and *X* is the full matrix of regressors. We define

$$\hat{g} = \exp(\hat{c}) - 1 \tag{4}$$

consequently $\hat{g}$ is lognormally distributed with the mean given by

$$\mathrm{E}(\hat{g}) = \exp(c + \tfrac{1}{2}\sigma_c^2) - 1 \tag{5}$$

and the median of $\hat{g}$ given by

$$\mathrm{Med}(\hat{g}) = \exp(c) - 1 \tag{6}$$

and the mode is given by

$$\mathrm{Mode}(\hat{g}) = \exp(c - \sigma_c^2) - 1 \tag{7}$$

Due to the lognormal nature of $\hat{g}$ the distribution of $\hat{g}$ is positively skewed and the greater the variance of $\hat{c}$ the greater the skewness. The relationship between these three measures of central tendency is given by

$$\mathrm{Mode}(\hat{g}) < \mathrm{Med}(\hat{g}) < \mathrm{E}(\hat{g})$$

There has been some debate in the literature as to the appropriate point estimate of $\hat{g}$. Kennedy (1981) suggests that the appropriate estimator is given by

$$g_K^* = \exp(\hat{c} - \tfrac{1}{2}\hat{\sigma}_c^2) - 1 \tag{8}$$

where the estimate of the variance of $\hat{c}$ is used. This estimator is equivalent to the maximum likelihood estimator for the expected value of g. Giles(1982) proposes an alternative estimate for g defined as

$$g_G^* = \left[\exp(\hat{c})\right] \sum_{j=0}^{\infty} \left(\frac{\left(\nu/2\right)^j \Gamma\left(\nu/2\right) \left(-\frac{1}{2}\hat{\sigma}_c^2\right)^j}{\Gamma\left(j + \nu/2\right) j!} \right) - 1 \tag{9}$$

Where v is the degrees of freedom of the estimate of c. This estimator is the minimum variance unbiased (MVU)estimator for the median of g as defined by Goldberger(1968). Both Shimizu and Iwase (1981) and Goldberger (1968) show that the MVU estimate for the expected value is given by

$$g_M^* = \left[\exp(\hat{c})\right] \sum_{j=0}^{\infty} \left(\frac{\left(\nu/2\right)^j \Gamma\left(\nu/2\right) \left(-\frac{1}{2}(\hat{\sigma}_c^2 - \hat{\sigma}_u^2)\right)^j}{\Gamma\left(j + \nu/2\right) j!} \right) - 1 \tag{10}$$

Simulations by Derrick (1984) show that the two estimators g_K^* and g_G^* are very similar when $\hat{c}$ is significant even in small samples. However this may be more a sign of the symmetry of the distribution of the samples used in the simulation than a general statement about these estimators. This symmetry is a function of the variance of the estimator for c.

While much attention has been paid to the estimation of g little has been given to the tests on g. In the next section we will show how it is also possible to specify g^* in

terms of $\hat{c}$ and its t-statistic and then use this relationship to express an approximate t-statistic for g* in terms of $\hat{c}$ and its t-statistic.

3. Tests for g_K^*

It is common practice when estimating semilog equations with dummy variables to make inferences concerning *g* from tests based on *c* alone thus using a test of the form:

$$H_0 : c = 0$$
$$H_1 : c \neq 0$$

which is equivalent to the test that:

$$H_0 : \exp(c) = 1, \text{ or } \exp(c) - 1 = 0$$
$$H_1 : \exp(c) \neq 1, \text{ or } \exp(c) - 1 \neq 0$$

However, since exp(*c*) - 1 is the median of $\hat{g}$, the usual *t*-test on the regression parameter is actually a test of the median of *g* rather than a test of the expected value of *g* which is the usual form of the tests. In addition, since the median is always less than the expected value we would anticipate that this test might always result in more rejections of H_0 than a test based on the expected value for the same case.

Instead of testing $\hat{g}_K^*$ directly we perform tests on the log of $\hat{g}_K^*$. If we wish to test

$$H_0 : \exp(c + \tfrac{1}{2}\sigma_c^2) - 1 = 0$$
$$H_1 : \exp(c + \tfrac{1}{2}\sigma_c^2) - 1 \neq 0$$

we can add one to both sides and take logs to get the equivalent test for the expected value of *g* to be;

$$H_0 : c + \tfrac{1}{2}\sigma_c^2 = 0$$
$$H_1 : c + \tfrac{1}{2}\sigma_c^2 \neq 0$$

The estimates of the log of $\hat{g}_K^*$ (here referred to as $\widehat{l\_g_K^*}$) are then given by

$$\widehat{l\_g_K^*} = \log(\hat{g}_K^* + 1) = \hat{c} + \frac{\hat{\sigma}_{\hat{c}}^2}{2} \tag{11}$$

Defining the usual *t*-statistic reported for the regression coefficient ($\hat{c}$) as *t*, the estimated variance of $\hat{c}$ is then given by $\hat{\sigma}_c^2 = \hat{c}^2/t^2$ and the log of $\hat{g}_K^*$ can be written as

$$\widehat{l\_g_K^*} = \hat{c} + \frac{\hat{c}^2}{2t^2} \quad . \tag{12}$$

Figure 1 plots the difference between $\widehat{l\_g_K^*}$ and $\hat{c}$ when $t = 2$. When $t < 2$ the difference would be above the curve and when $t > 2$ it would be below this curve. However the value of $\widehat{l\_g_K^*}$ cannot be interpreted alone thus the need to derive an equivalent *t*-statistic for $\widehat{l\_g_K^*}$.

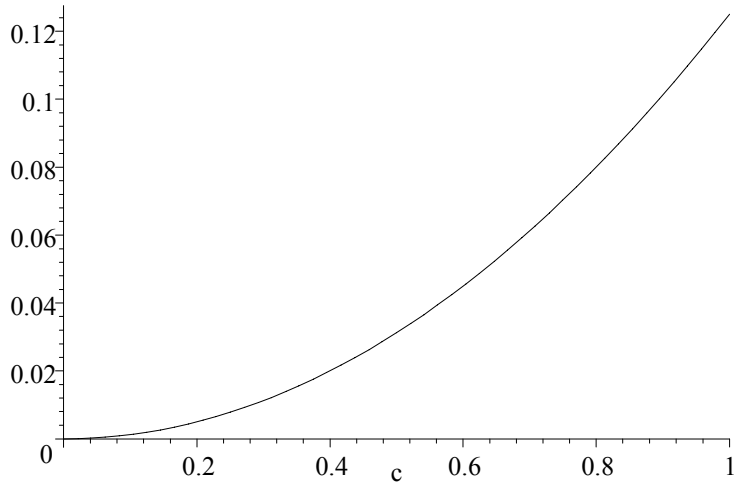

Figure 1 The value of $\widehat{l\_g_K^*} - \hat{c}$ versus $\hat{c}$ when $t = 2$.

First, an estimate of the variance of $\widehat{l\_g_K^*}$ is needed. Land (1972) refers to an approximation for the variance deviation of the expected value of $\widehat{l\_g_K^*}$ as Cox's Direct

approximation. He demonstrates that this approximation has quite good properties. This approximation is derived by using the MVU estimator for $\sigma_{\hat{c}}^2$ defined as $\hat{\sigma}_{\hat{c}}^2$.

$$\hat{\sigma}^2_{l\_g_M^*} = \hat{\sigma}^2_{\hat{c}} + \tfrac{1}{2}(\hat{\sigma}^4_{\hat{c}}/v) \tag{13}$$

where *v* is the degrees of freedom for the estimate of σ_u^2 (*n*-*k* for a regression). Thus as $v \to \infty$ the estimates of the variances approach each other $\hat{\sigma}^2_{l\_g_M^*} \to \hat{\sigma}^2_{\hat{c}}$. Substituting the value of *t* in (13) we get:

$$\hat{\sigma}^2_{l\_g_M^*} = \left(\hat{c}^2/t^2\right)\left\{1 + \tfrac{1}{2}\left[\left(\hat{c}^2/t^2\right)/v\right]\right\} \tag{14}$$

Now substituting (14) into the equation for the t-statistic of the log of g_K^* and defining this as t^* we obtain:

$$t^* = \frac{\widehat{l\_g_K^*}}{\hat{\sigma}_{l\_\hat{g}_K^*}} = \left(2t^2 + \hat{c}\right)\sqrt{\frac{v}{4tv + 2\hat{c}^2}} \tag{15}$$

as a function of *v*, $\hat{c}$ and t. Two levels of degrees of freedom are illustrated in Figures 2 and 3. In Figure 2 the values of *t**- *t* are plotted as a function of both $\hat{c}$ and *t* for the case when the number of degrees of freedom (df or *v*) of 25. We can see that for small values of *t* the test on the expected value of *g* results in much larger t-statistics than the *t* statistic for the estimated parameter value. In Figure 3 we change the df to 1000 which is closer in magnitude to many common econometric applications. From figures 2 and 3 we note that large discrepancies occur between the *t* and *t** when the value of *t* is less than .5 in small samples and .2 in larger ones while at the same time the parameter estimate is quite large.

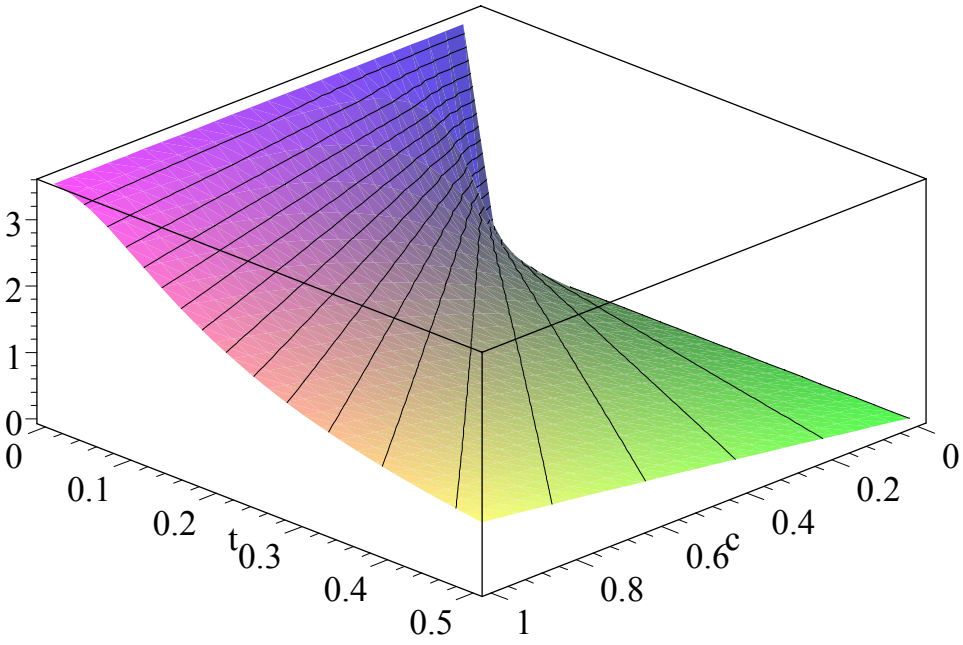

Figure 2 *t**- *t* as a function of $\hat{c}$ and t when the df (*v*) of 25.

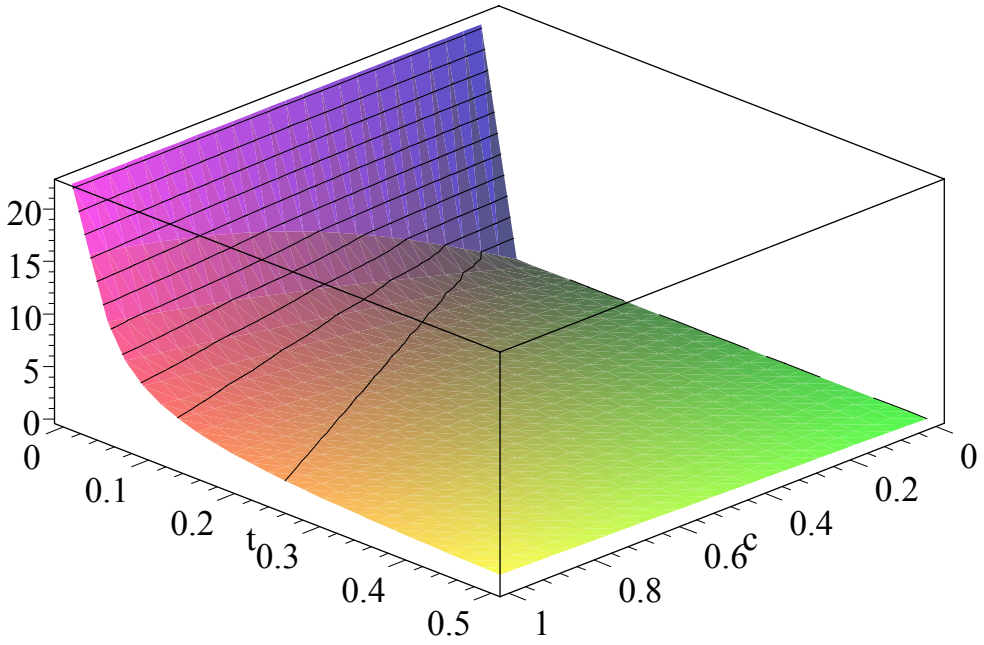

Figure3 *t**- *t* as a function of $\hat{c}$ and t when the df (*v*) of 1000.

4. Log dummies for more than two categories

Generally a discrete variable may refer to more than two categories and hence the equation to be estimated includes a number of related dummy variables. Two equivalent ways of estimating involve either omitting one of the categories or alternatively imposing a constraint on the values of the dummy variable coefficients. The equivalence of these two approaches is shown in Lye and Hirschberg (1999).

Suppose the regression equation to be estimated is:

$$y = \alpha + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \mathbf{X}\beta + \varepsilon \qquad (16)$$

where y is a Tx1 vector of observations on the dependent variable, $\mathbf{X}$ is a Txk vector of observations on k quantitative explanatory variables, α is the estimated intercept term, β is a kx1 vector of parameters, ε is a Tx1 vector of disturbances and D_i are dummy variables defined as whether the observation is in one of three separate groups. For example the observation in a wage equation may be for individuals and the dummy variables may indicate different occupations.

$$D_j = 1 \text{ for occupation } j = 1, 2, 3$$
$$= 0 \text{ otherwise}$$

Two equivalent methods of estimating (16) are to either impose the restriction that the one group has a zero value and acts as the reference group whose effect is included in the intercept (for example if $\gamma_3 = 0$) which would result in the following model to be estimated

$$y = \alpha + \gamma_1 D_1 + \gamma_2 D_2 + \mathbf{X}\beta + \varepsilon \qquad (17)$$

Or to impose the restriction that the sum of the dummy variable coefficients is equal to zero ($\gamma_1 + \gamma_2 + \gamma_3 = 0$) which implies that one coefficient can be set to minus the sum of the rest. In this case we can estimate the following model

$$\begin{aligned} y &= \varphi + \pi_1(D_1 - D_3) + \pi_2(D_2 - D_3) + \mathbf{X}\beta + \varepsilon \\ \pi_3 &= -(\pi_1 + \pi_2); \end{aligned} \tag{18}$$

Note that the parameters estimated for the occupations are not the same ($\gamma_j \neq \pi_j$) nor is the intercept the same in models (17) and (18). Define g_1 to be equal to the percentage effect on y given D_1=1 then

$$g_1 = \exp(\gamma_1) - 1 = \frac{y_{D_1=1} - y_{D_3=1}}{y_{D_3=1}} \tag{19}$$

then in terms of equation (18) we have

$$g_1 = \frac{\exp(\pi_1) - \exp(-\pi_1 - \pi_2)}{\exp(-\pi_1 - \pi_2)} \tag{20}$$

The following relationships can be found to related these parameters.

(18)	(17)
$\varphi = \alpha + \frac{\gamma_1}{3} + \frac{\gamma_2}{3} + \frac{\lambda_1}{2}$	$\alpha = \varphi - \pi_1 - \pi_2 - \rho_1$
$\pi_1 = \frac{2\gamma_1 - \gamma_2}{3}$	$\gamma_1 = 2\pi_1 + \pi_2$
$\pi_2 = \frac{2\gamma_2 - \gamma_1}{3}$	$\gamma_2 = 2\pi_2 + \pi_1$

And we can show that

$$\frac{\exp(\pi_1) - \exp(-\pi_1 - \pi_2)}{\exp(-\pi_1 - \pi_2)} = \exp(\gamma_1) - 1 \tag{21}$$

If (17) is the form estimated then an estimate of g_1 using

Kennedy's (1981) approach is given by

$$g^*_{K1} = \exp\left(2\hat{\pi}_1 + \hat{\pi}_2 - \tfrac{1}{2}\left\{4\hat{\sigma}^2_{\hat{\pi}_1} + \hat{\sigma}^2_{\hat{\pi}_2} + 4\hat{\sigma}_{\hat{\pi}_1,\hat{\pi}_2}\right\}\right) - 1 \qquad (22)$$

Using a similar argument it can be shown that

$$g_2 = \exp(\gamma_2) - 1 = \frac{\exp(\pi_1) - \exp(-\pi_1 - \pi_2)}{\exp(-\pi_1 - \pi_2)} \qquad (23)$$

and

$$g^*_{K2} = \exp\left(\hat{\pi}_1 + 2\hat{\pi}_2 - \tfrac{1}{2}\left\{\hat{\sigma}^2_{\hat{\pi}_1} + 4\hat{\sigma}^2_{\hat{\pi}_2} + 4\hat{\sigma}_{\hat{\pi}_1,\hat{\pi}_2}\right\}\right) - 1 \qquad (24)$$

Following the discussion above we can also form the appropriate approximate test statistic for the parameter implied by these estimates as well by substituting the variance for the combination of parameters so that the formulas developed above can be used directly.

5. Typical econometric applications

5.1 Hedonic Price Model

In this section we consider two econometric applications that use dummy variables in log models. The first is the estimation of hedonic price models and the second is in the estimation of wage equations. Both applications are characterized by typical size parameter estimates and sample sizes.

Rosen first coined the term Hedonic price model to describe models which were designed to explain the variation in market prices for different goods and services (Rosen(1974)). However regressions that predict the market price as a function of product and service characteristics have a much longer history. Berndt (chapter 4, 1991) details a number of earlier examples of these models. Hedonic price models are used in a variety of

different markets and have been applied to the computation of price indexes (Chow (1967)) and for the imputation of the value of public amenities (see Bartik and Smith (1987)) among other uses. Most commonly used models apply a lognormal transformation of the dependent variable (price or value) and include dummy variables as indicators for certain characteristics of the goods or services sold.

A typical model would be a regression to explain the sale price of a house. The characteristics of the house such as the presence of a swimming pool or the location on a cul de sac are usually represented by dummy variables. The application presented below is for a set of 288 houses sold in Dallas Texas in July of 1986 where a model was fit to the log of the house price in thousands of dollars. Note that even though the dependent variable has a mean of 4.6 the values of the parameters estimated for the dummy variables are less than one and in most cases much less than .5. In this example from Hayes et al (1999) the variables DISTANCE, DISTANCE^2 are the distance to the center of the city and distance squared, FAC1_D, FAC1_H, FAC2_D, FAC2_H, FAC3_D, and FAC3_H are factors based on demographic and housing characteristics of the neighborhood measured in continuous values, SQFTLA is the square footage of the house and YRBLT is the year the house was built. The dummy variables are FIREPL for the presence of a fire place, POOL for the presence of a swimming pool, and ND for whether the house is located north of the central business district in Dallas. The dependent variable is the log of the house price and from an earlier analysis it was found that the model in the linear price resulted in a skewed distribution of the residuals.

From the results listed in Table 1 it can be seen that values of the dummy variables fall in the range of 0 to .44. The last column of the table lists the estimated values of the *t*-

statistic based on the approximation given by t^*. Due to the high values of the *t*-statistics in this case we find no cases where there is a large difference between *t* and t^* as predicted by the relationships plotted in figures 2 and 3.

Dependent Variable: LPRICE
Included observations: 288
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	t^*
C	3.577865	0.151789	23.57130	
DISTANCE	-0.162585	0.114482	-1.420185	
DISTANCE^2	0.007287	0.022592	0.322566	
FAC1_D	-0.174645	0.051026	-3.422666	
FAC1_H	-0.007344	0.025801	-0.284656	
FAC2_D	-0.161454	0.035083	-4.601993	
FAC2_H	0.014516	0.029969	0.484380	
FAC3_D	-0.042816	0.022131	-1.934643	
FAC3_H	-0.003601	0.042254	-0.085226	
FIREPL	**0.149494**	**0.036014**	**4.151022**	**4.168999**
POOL	**0.004048**	**0.049260**	**0.082171**	**.106806**
SQFTLA	0.000535	2.92E-05	18.28822	
YRBLT	0.004238	0.001628	2.603561	
ND	**0.057615**	**0.051883**	**1.110476**	**1.136418**
R-squared	0.889689	Mean dependent var		4.645586
Adjusted R-squared	0.884455	S.D. dependent var		0.765940
S.E. of regression	0.260357	Akaike info criterion		0.193868
Sum squared resid	18.57337	Schwarz criterion		0.371928
Log likelihood	-13.91694	F-statistic		169.9909
Durbin-Watson stat	1.968638	Prob(F-statistic)		0.000000

Table 1. An example from an Hedonic Price Regression.

5.2 Wage Equation

Another common application of log models in econometrics are found in the estimation of wage equations (see Heckman and Polachek (1974)). These models are typically used to determine the effects of particular characteristics of individuals on the level of there earnings. They are often conducted using survey data and thus the number of observations tends to be at least 1000 or more. Table 2 lists a typical set of regression

results from an application of this method to a set of household survey data from Australia collected in 1989 as described in (Hirschberg and Lye (1999)). In this case age and age squared along with the level of alcohol consumed are included as independent variables. In addition we have included dummy variables for gender, marital status and if they reside in a state capital (most major cities in Australia are the state capital). Again note that even though the mean of the dependent variable (the log of income in dollars) is listed as 10.07 we find that the majority of the dummy variables have coefficients with absolute values less than .5 (recall that the sign of these coefficients can always be reversed by a redefinition of the dummy variable).

In this case we find that the dummy variable for gender indicates that male workers are paid significantly more than female workers and that married workers are paid less than non-married workers. The values for the proportional change in this case are also significant for all the dummy variables in this model due to the very high *t*-statistics for the regression parameters and the large number of degrees of freedom – over 10,000 in this case.

Dependent Variable: LINC
Included observations: 11515
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	t^*
FEMALE	**-0.476528**	**0.011042**	**-43.15588**	**-43.16147**
AGE	0.228030	0.006586	34.62471	
CAPCITY	**-0.120499**	**0.009168**	**-13.14370**	**-13.14802**
AGE^2	-0.001001	3.06E-05	-32.74577	
ALCOHOL	2.79E-05	1.83E-05	1.520986	
MARRIED	**-0.010228**	**0.002937**	**-3.481808**	**-3.483934**
C	9.529160	0.047897	198.9519	

R-squared	0.295094	Mean dependent var	10.07627
Adjusted R-squared	0.294726	S.D. dependent var	0.548213
S.E. of regression	0.460392	Akaike info criterion	1.287132
Sum squared resid	2439.249	Schwarz criterion	1.291601
Log likelihood	-7403.665	F-statistic	802.9283
Durbin-Watson stat	1.910517	Prob(F-statistic)	0.000000

Table 2. An example from a Wage Equation Regression.

6. Conclusions

This paper shows that inferences drawn from the standard *t*-statistic for dummy variables in regressions with log transformed dependent variables are equivalent to tests of the median of the proportional change. We propose an approximation for the equivalent t-statistic for the expected value of the proportional change and examine how these two test statistics vary with the parameters of the model under consideration. The potential for large deviations in the inferences drawn between the median and the expected value occurs where parameters that have estimated values that are greater than .5 and *t*-statistics of less than .2. We note that this difference is greatest with small sample sizes. In examining two representative econometric applications of the estimation of simple dummy variables with a log transformed dependent variable, we find that they do not have result in parameter estimates that would indicate any changes in inference.

In addition to considering the case of a dummy variables for a single category we also considered the multiple category case as well where we show that the approximation for the multiple category case can be defined in the same manner as for the single category situation.

A concern with the approximation proposed here is the employment of *t*-statistics to form confidence bounds due to the symmetric nature of the distribution implied. One future direction for this research would be the investigation of bootstrap based confidence intervals based on the pivot statistic that can be formed using the approximate *t*-statistic defined as t^*.

References

Asher, C. (1992), 'Hedonic Analysis of Reliability and Safety for New Automobiles', *Journal of Consumer Affairs*, 26 377-96.

Australian Bureau of Statistics (1993a*)*, Training and Education Experience Australia 1993, catalogue no.6278.0.

Australian Bureau of Statistics (1993b), 'Notes for Interviewers for the Training and Education Experience Survey, 1993', mimeo.

Bartik, Timothy J., and V. Kerry Smith (1987), "Urban Amenities and Public Policy," in *Handbook of Regional and Urban Economics*, ed. Edwin S. Mills, (Amsterdam: North Holland Press).

Berndt, E. R. (1991). *The Practice of Econometrics: Classic and Contemporary,* Addison-Wesley Publishing Company.

Borland, J., J. Hirschberg, and J. Lye (1997), 'Computer Knowledge and Earnings: Evidence for Australia', RP 571 Department of Economics, University of Melbourne.

Chow, G. C. (1967), 'Technological Change and the Demand for Computers', American Economic Review, 57, 1117-1130.

Derrick, F. (1984), 'Interpretation of Dummy Variables in Semilogarithmic Equations: Small Sample Implications', *Southern Economic Journal*, 50, 1185-88.

Giles, D. (1982), 'The Interpretation of Dummy Variables in Semilogarithmic Equations', *Economics Letters*, 10, 77-79.

Goldberger, A. S. (1968), "The Interpretation and Estimation of Cobb-Douglas Functions", *Econometrica*, 53, 464-472.

Halvorsen, R. and R. Palmquist (1980), 'The Interpretation of Dummy Variables in Semilogarithmic Equations', *American Economic Review*, 70, 474-75.

Lye, J. and J. Hirschberg (1999), "Dealing with the dummies", mimeo, Economics Department, University of Melbourne.

Hayes, K., Hirschberg, J., Lye J. and Taylor, L. (1999), "Multivariate Generated Regressors and Heteroskedasticity in a Cross-section: An application tot the value of neighborhood schools", University of Melbourne Department of Economics research paper #692.

Heckman, J. and S. Polachek (1974), “Empirical Evidence on the Functional Form of Earnings-Schooling Relationship”, *Journal of the American Statistical Association*, 69, 350-354.

Hirschberg. J. and J. Lye, (1999), “Wages and Alcohol Consumption, Smoking, Weight gain and Exercising: Evidence on Australian Men and Women”, Research Paper Number 684, Department of Economics, University of Melbourne, March 1999.

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

Land. C. E. (1972) “An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means”, *Technometrics*, 14, 145-158.

Rosen, Sherwin (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy* 82 (January/February): 34-55.

Sweeney, R. and E. Ulveling (1972), 'A Transformation for Simplifying the Interpretation of Coefficients of Binary Variables in Regression Analysis', *The American Statistician*, 26, 30-33.

Suits, D. (1984), 'Dummy Variables: Mechanics V. Interpretation', *The Review of Economics and Statistics*, 51, 177-180.