Linear Regression Models with Logarithmic Transformations

Kenneth Benoit*
Methodology Institute
London School of Economics
kbenoit@lse.ac.uk

March 17, 2011

1 Logarithmic transformations of variables

Considering the simple bivariate linear model $Y_i = \alpha + \beta X_i + \epsilon_i$,[1] there are four possible combinations of transformations involving logarithms: the linear case with no transformations, the linear-log model, the log-linear model[2], and the log-log model.

	X	
Y	X	log$X$
Y	*linear*	*linear-log*
	$\hat{Y}_i = \alpha + \beta X_i$	$\hat{Y}_i = \alpha + \beta \log X_i$
log$Y$	*log-linear*	*log-log*
	$\log\hat{Y}_i = \alpha + \beta X_i$	$\log\hat{Y}_i = \alpha + \beta \log X_i$

Table 1: Four varieties of logarithmic transformations

Remember that we are using *natural* logarithms, where the base is $e \approx 2.71828$. Logarithms may have other bases, for instance the decimal logarithm of base 10. (The base 10 logarithm is used in the definition of the Richter scale, for instance, measuring the intensity of earthquakes as Richter = log(intensity). This is why an earthquake of magnitude 9 is 100 times more powerful than an earthquake of magnitude 7: because $10^9/10^7 = 10^2$ and $\log_{10}(10^2) = 2$.)

Some properties of logarithms and exponential functions that you may find useful include:

1. $\log(e) = 1$
2. $\log(1) = 0$
3. $\log(x^r) = r\ \log(x)$
4. $\log e^A = A$

*With valuable input and edits from Jouni Kuha.

[1] The bivariate case is used here for simplicity only, as the results generalize directly to models involving more than one X variable, although we would need to add the caveat that all other variables are held constant.

[2] Note that the term "log-linear model" is also used in other contexts, to refer to some types of models for other kinds of response variables Y. These are different from the log-linear models discussed here.

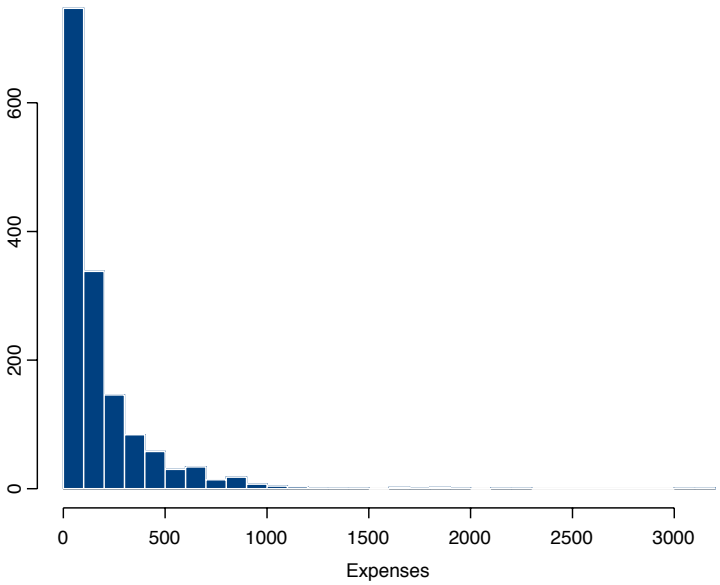
5. $e^{\log A} = A$
6. $\log(AB) = \log A + \log B$
7. $\log(A/B) = \log A - \log B$
8. $e^{AB} = \left(e^{A}\right)^{B}$
9. $e^{A+B} = e^{A}e^{B}$
10. $e^{A-B} = e^{A}/e^{B}$

2 Why use logarithmic transformations of variables

Logarithmically transforming variables in a regression model is a very common way to handle situations where a non-linear relationship exists between the independent and dependent variables.[3] Using the logarithm of one or more variables instead of the un-logged form makes the effective relationship non-linear, while still preserving the linear model.

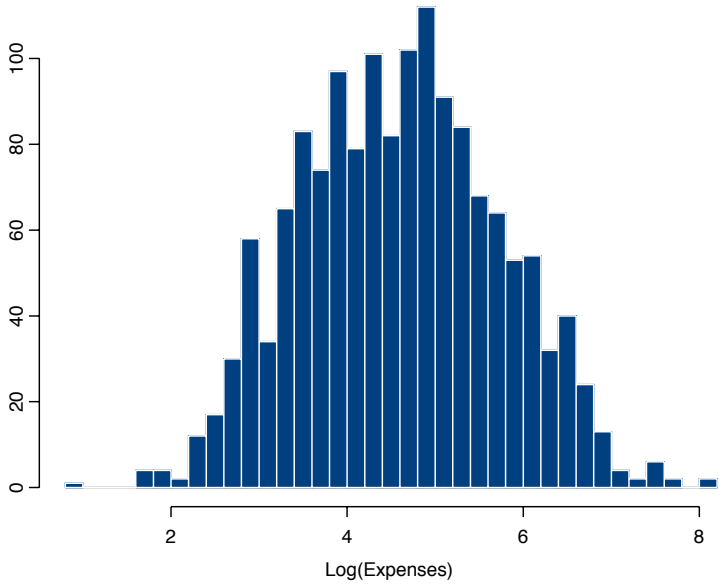
Logarithmic transformations are also a convenient means of transforming a highly skewed variable into one that is more approximately normal. (In fact, there is a distribution called the *log-normal* distribution defined as a distribution whose logarithm is normally distributed – but whose untransformed scale is skewed.)

For instance, if we plot the histogram of expenses (from the MI452 course pack example), we see a significant right skew in this data, meaning the mass of cases are bunched at lower values:



If we plot the histogram of the logarithm of expenses, however, we see a distribution that looks much more like a normal distribution:

[3]The other transformation we have learned is the *quadratic* form involving adding the term X^2 to the model. This produces curvature that unlike the logarithmic transformation that can reverse the direction of the relationship, something that the logarithmic transformation cannot do. The logarithmic transformation is what as known as a monotone transformation: it preserves the ordering between x and $f(x)$.



3 Interpreting coefficients in logarithmically models with logarithmic transformations

3.1 Linear model: $Y_i = \alpha + \beta X_i + \epsilon_i$

Recall that in the linear regression model, $\log Y_i = \alpha + \beta X_i + \epsilon_i$, the coefficient β gives us directly the change in Y for a one-unit change in X. No additional interpretation is required beyond the estimate $\hat{\beta}$ of the coefficient itself.

This literal interpretation will still hold when variables have been logarithmically transformed, but it usually makes sense to interpret the changes not in log-units but rather in percentage changes.

Each logarithmically transformed model is discussed in turn below.

3.2 Linear-log model: $Y_i = \alpha + \beta \log X_i + \epsilon_i$

In the linear-log model, the literal interpretation of the estimated coefficient $\hat{\beta}$ is that a one-unit increase in $\log X$ will produce an expected increase in Y of $\hat{\beta}$ units. To see what this means in terms of changes in X, we can use the result that

$$\log X + 1 = \log X + \log e = \log(eX)$$

which is obtained using properties 1 and 6 of logarithms and exponential functions listed on page 1. In other words, *adding* 1 to $\log X$ means *multiplying* X itself by $e \approx 2.72$.

A proportional change like this can be converted to a percentage change by subtracting 1 and multiplying by 100. So another way of stating "multiplying X by 2.72" is to say that X increases by 172% (since $100 \times (2.72 - 1) = 172$).

So in terms of a change in X (unlogged):

- $\hat{\beta}$ is the expected change in Y when X is multiplied by e.
- $\hat{\beta}$ is the expected change in Y when X increases by 172%
- For other percentage changes in X we can use the following result: The expected change in Y associated with a p% increase in X can be calculated as $\hat{\beta} \cdot \log([100+p]/100)$. So to work out the expected change associated with a 10% increase in X, therefore, multiply $\hat{\beta}$ by $\log(110/100) = \log(1.1) = .095$. In other words, $0.095\hat{\beta}$ is the expected change in Y when X is multiplied by 1.1, i.e. increases by 10%.
- For small p, approximately $\log([100+p]/100) \approx p/100$. For $p = 1$, this means that $\hat{\beta}/100$ can be interpreted approximately as the expected increase in Y from a 1% increase in X

3.3 Log-linear model: $\log Y_i = \alpha + \beta X_i + \epsilon_i$

In the log-linear model, the literal interpretation of the estimated coefficient $\hat{\beta}$ is that a one-unit increase in X will produce an expected increase in $\log Y$ of $\hat{\beta}$ units. In terms of Y itself, this means that the expected value of Y is multiplied by $e^{\hat{\beta}}$. So in terms of effects of changes in X on Y (unlogged):

- Each 1-unit increase in X multiplies the expected value of Y by $e^{\hat{\beta}}$.
- To compute the effects on Y of another change in X than an increase of one unit, call this change c, we need to include c in the exponent. The effect of a c-unit increase in X is to multiply the expected value of Y by $e^{c\hat{\beta}}$. So the effect for a 5-unit increase in X would be $e^{5\hat{\beta}}$.
- For small values of $\hat{\beta}$, approximately $e^{\hat{\beta}} \approx 1+\hat{\beta}$. We can use this for the following approximation for a quick interpretation of the coefficients: $100 \cdot \hat{\beta}$ is the expected percentage change in Y for a unit increase in X. For instance for $\hat{\beta} = .06$, $e^{.06} \approx 1.06$, so a 1-unit change in X corresponds to (approximately) an expected increase in Y of 6%.

3.4 Log-log model: $\log Y_i = \alpha + \beta \log X_i + \epsilon_i$

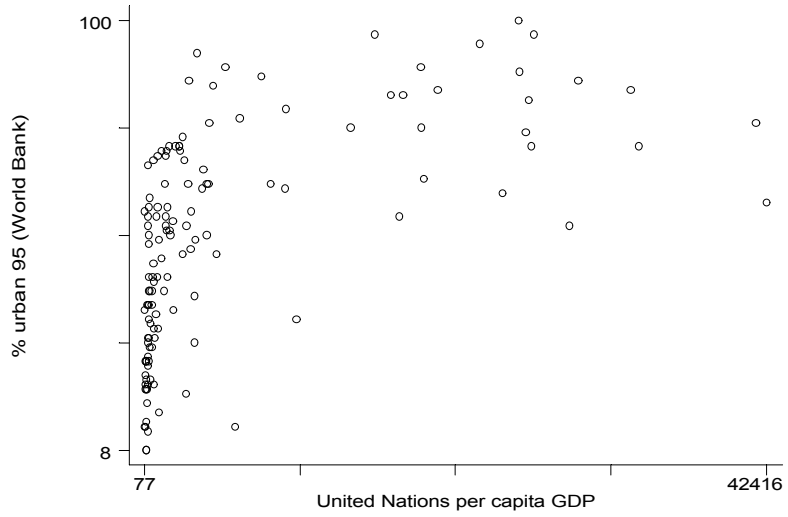
In instances where both the dependent variable and independent variable(s) are log-transformed variables, the interpretation is a combination of the linear-log and log-linear cases above. In other words, the interpretation is given as an expected percentage change in Y when X increases by some percentage. Such relationships, where both Y and X are log-transformed, are commonly referred to as elastic in econometrics, and the coefficient of $\log X$ is referred to as an elasticity.

So in terms of effects of changes in X on Y (both unlogged):

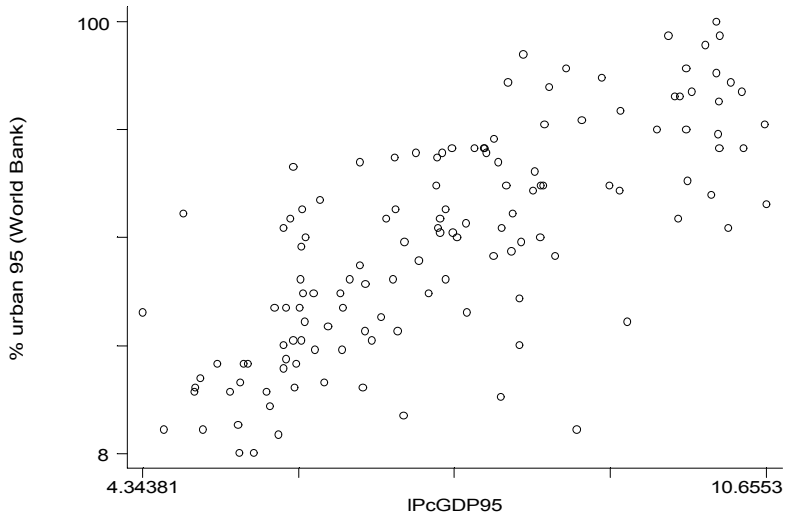
- multiplying X by e will multiply expected value of Y by $e^{\hat{\beta}}$
- To get the proportional change in Y associated with a p percent increase in X, calculate $a = \log([100+p]/100)$ and take $e^{a\hat{\beta}}$

4 Examples

Linear-log. Consider the regression of % urban population (1995) on per capita GNP:



The distribution of per capita GDP is badly skewed, creating a non-linear relationship between X and Y. To control the skew and counter problems in heteroskedasticity, we transform GNP/capita by taking its logarithm. This produces the following plot:



and the regression with the following results:

```
. regress urb95 lPcGDP95
  Source |       SS       df       MS                  Number of obs =     132
---------+------------------------------                F(  1,   130) =  158.73
   Model |  38856.2103     1  38856.2103                Prob > F      =  0.0000
Residual |  31822.7215   130  244.790165                R-squared     =  0.5498
---------+------------------------------                Adj R-squared =  0.5463
   Total |  70678.9318   131  539.533831                Root MSE      =  15.646
------------------------------------------------------------------------------
   urb95 |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
lPcGDP95 |   10.43004   .8278521     12.599   0.000       8.792235    12.06785
   _cons |  -24.42095   6.295892     -3.879   0.000      -36.87662   -11.96528
------------------------------------------------------------------------------
```

To interpret the coefficient of 10.43004 on the log of the GNP/capita variable, we can make the following statements:

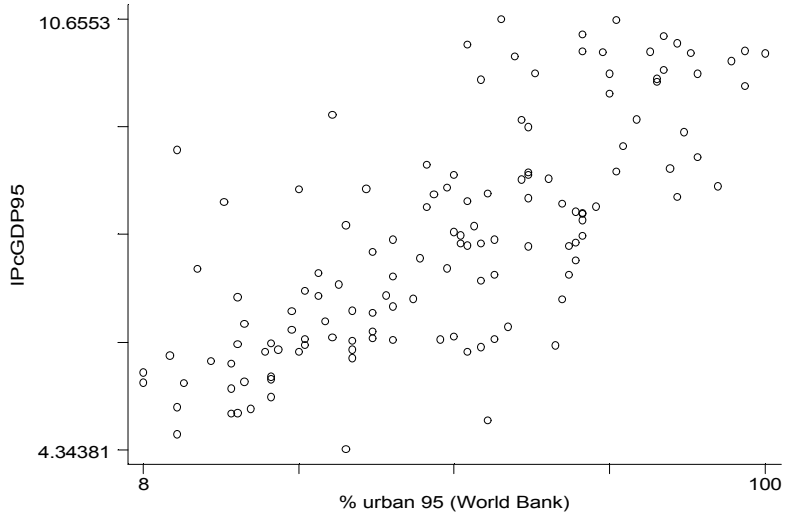
Directly from the coefficient: An increase of 1 in the log of GNP/capita will increase Y by 10.43004. (This is not extremely interesting, however, since few people are sure how to interpret the natural logarithms of GDP/capita.)

Multiplicative changes in e: Multiplying GNP/cap by e will increase Y by 10.43004.

A 1% increase in X: A 1% increase in GNP/cap will increase Y by $10.43004/100 = .1043$

A 10% increase in X: A 10% increase in GNP/cap will increase Y by $10.43004 * \log(1.10) = 10.43004 * .09531 \approx 0.994$.

Log-linear. What if we reverse X and Y from the above example, so that we regress the log of GNP/capita on the %urban? In this case, the logarithmically transformed variable is the Y variable. This leads to the following plot (which is just the transpose of the previous one — this is only an example!):



with the following regression results:

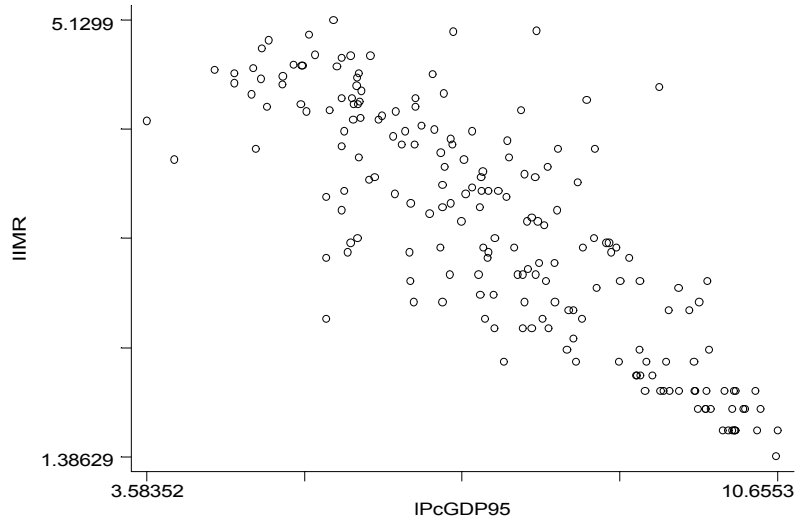
```
. regress lPcGDP95 urb95
  Source |       SS       df       MS                  Number of obs =     132
---------+------------------------------                F(  1,   130) =  158.73
   Model |  196.362646     1  196.362646                Prob > F      =  0.0000
Residual |  160.818406   130  1.23706466                R-squared     =  0.5498
---------+------------------------------                Adj R-squared =  0.5463
   Total |  357.181052   131  2.72657291                Root MSE      =  1.1122
------------------------------------------------------------------------------
lPcGDP95 |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   urb95 |    .052709   .0041836     12.599   0.000       .0444322    .0609857
   _cons |   4.630287   .2420303     19.131   0.000       4.151459    5.109115
------------------------------------------------------------------------------
```

To interpret the coefficient of .052709 on the `urb95` variable, we can make the following statements:

Directly from the coefficient, transformed Y: Each one unit increase `urb95` in increases `lPcGDP95` by .052709. (Once again, this is not particularly useful as we still have trouble thinking in terms of the natural logarithm of GDP/capita.)

Directly from the coefficient, *un*transformed Y: Each one unit increase of `urb95` increases the untransformed GDP/capita by a *multiple* of $e^{0.52709} = 1.054$ – or a 5.4% increase. (This is very close to the 5.3% increase that we get using our quick approximate rule described above for interpreting the .053 as yielding a 5.3% increase for a one-unit change in X.)

Log-log. Here we consider a regression of the logarithm of the infant mortality rate on the log of GDP/capita. The plot and the regression results look like this:




```
. regress lIMR lPcGDP95
  Source |       SS       df       MS                  Number of obs =     194
---------+------------------------------                F(  1,   192) =  404.52
   Model |  131.035233     1  131.035233                Prob > F      =  0.0000
Residual |  62.1945021   192  .323929698                R-squared     =  0.6781
---------+------------------------------                Adj R-squared =  0.6765
   Total |  193.229735   193  1.00119034                Root MSE      =  .56915
------------------------------------------------------------------------------
    lIMR |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
lPcGDP95 |  -.4984531   .0247831    -20.113   0.000      -.5473352    -.449571
   _cons |   7.088676   .1908519     37.142   0.000       6.71224    7.465111
------------------------------------------------------------------------------
```

To interpret the coefficient of -.4984531 on the `lPcGDP95` variable, we can make the following statements:

Directly from the coefficient, transformed Y: Each one unit increase `lPcGDP95` in increases `lIMR` by -.4984531. (Since we cannot think directly in natural log units, then once again, this is not particularly useful.)

Multiplicative changes in both X and Y: Multiplying X (GNP/cap) by $e \approx 2.72$ multiplies Y by $e^{-.4984531} = 0.607$, i.e. reduces the expected IMR by about 39.3%.

A 1% increase in X: A 1% increase in GNP/cap multiplies IMR by $e^{-.4984531 * \log(1.01)} = .0.9950525$. So a 1% increase in GNP/cap reduces IMR by 0.5%.

A 10% increase in X: A 10% increase in GNP/cap multiplies IMR by $e^{-.4984531 * \log(1.1)} \approx .954$. So a 10% increase in GNP/cap reduces IMR by 4.6%.