ECON 3040 - Log models

Ryan T. Godwin

University of Manitoba

Logarithms

Another way to approximate the non-linear relationship between Y and X is by using logarithms.

- Logarithms can be used to approximate a percentage change.
- If the relationship between two variables can be expressed in terms of proportional or percentage changes, then it is a type of non-linear effect.
- To see this, consider a 1% increase in 100 (which is 1), and a 1% increase in 200 (which is 2). The same 1% increase can be generated by different changes in the variable (e.g. a change of 1 or of 2).

For example, consider an increase in hourly wage of \$1.

- That is not a big increase for someone making \$50 per hour (an increase of only 2%).
- This change in wage is unlikely to have much effect on the behaviour of the individual.
- However, imagine an individual whose hourly wage is only \$1 per hour. An increase of \$1 doubles the wage (100% increase)!
- This is likely to have a big impact on behaviour.
- It is desirable to measure thinks like wage in terms of proportional or percentage changes (regardless of whether it is included in a model as the dependent variable or as a regressor).
- This can be accomplished by using the log of the variable in the regression model, instead of the variable itself.

Percentage change

Let's be explicit about what is meant by a percentage change. A percentage change in X is:

$$\frac{\Delta X}{X} \times 100 = \frac{X_2 - X_1}{X_1} \times 100$$

where X_1 is the starting value of X, and X_2 is the final value.

Logarithm approximation to percentage change

The approximation to percentage changes using logarithms is:

$$\log(X + \Delta X) - \log(X) \times 100 \approx \frac{\Delta X}{X} \times 100$$

or

$$\log(X_2 - X_1) \times 100 \approx \frac{X_2 - X_1}{X_1} \times 100$$

- So, when X changes, the change in $\log(X)$ is approximately equal to a percentage change in X.
- The approximation is more accurate the smaller the change in X.
- The approximation does not work well for changes above 10%.

Table: Percentage change, and approximate percentage change using the log function.

Change in X		% change:	Approx. % change:
X_1	X_2	$\frac{X_2 - X_1}{X_1} \times 100$	$(\log X_2 - \log X_1) \times 100$
1	2	100%	69.32%
1	1.1	10%	9.53%
1	1.01	1%	0.995%
5	6	20%	18.23%
11	12	9.09%	8.70%
11	11.1	0.91%	0.91%

Logs in the population model

The log function can be used in our population model so that the βs have various *percentage changes* interpretations. There are three ways we can introduce the log function into our models. The three different possibilities arise from taking logs of the left-hand-side variable, one or more of the right-hand-side variables, or both.

Table: Three population models using the log function.

Population model	Population regression function
I. linear-log	$Y = \beta_0 + \beta_1 \log X + \epsilon$
II. log-linear	$\log Y = \beta_0 + \beta_1 X + \epsilon$
III. log-log	$\log Y = \beta_0 + \beta_1 \log X + \epsilon$

For each of the three different population models above, β_1 has a different percentage change interpretation. We don't derive the interpretations of β_1, but instead list them for the three different cases in table 2:

- linear-log: a 1% change in X is associated with a $0.01\beta_1$ change in Y.
- log-linear: a change in X of 1 is associated with a $100 \times \beta_1\%$ change in Y.
- log-log: a 1% change in X is associated with a $\beta_1\%$ change in Y. β_1 can be interpreted as an *elasticity*.

A note on R^2

R^2 and $\bar{R}^2$ measure the proportion of variation in the dependent variable (Y) that can be explained using the X variables.

- When we take the log of Y in the log-linear or log-log model, the variance of Y changes.
- That is, $\text{Var}[\log Y] \neq \text{Var}[Y]$.
- We cannot use R^2 or $\bar{R}^2$ to compare models with different dependent variables.
- That is, we should not use R^2 to decide between two models, where the dependent variable is Y in one, and $\log Y$ in the other.

Log-linear model for the CPS data

It is common to use the log of *wage* as the dependent variable, instead of just *wage*. This allows for the factors that determine differences in wages be associated with approximate percentage changes in *wage*. In the following, we'll see an example of a log-linear model estimated using the CPS data. Start by loading the data:

```r
install.packages("AER")
library(AER)
data("CPS1985")
```

and estimate a log-linear model:

$$\log(wage) = \beta_0 + \beta_1 education + \beta_2 gender + \beta_3 age + \beta_4 experience + \epsilon$$

```r
summary(lm(log(wage) ~ education + gender + age + experience
    , data = CPS1985))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.15357    0.69387   1.663    0.097 .
education     0.17746    0.11371   1.561    0.119
genderfemale -0.25736    0.03948  -6.519 1.66e-10 ***
age          -0.07961    0.11365  -0.700    0.484
experience    0.09234    0.11375   0.812    0.417
```

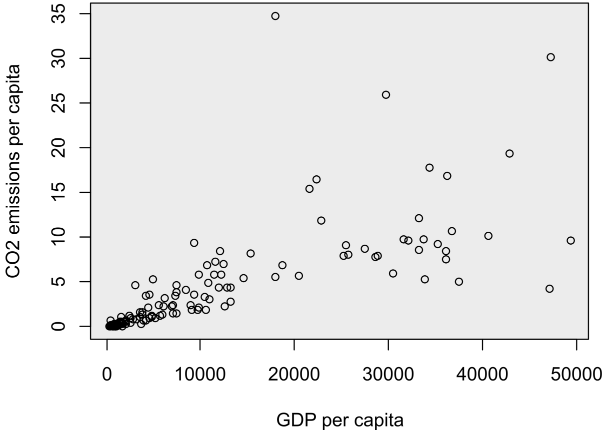
- The interpretation of the estimated coefficient on `education`, for example, is that a 1 year increase in *education* is associated with a 17.8% increase in *wage*.
- The interpretation of the coefficient on the dummy variable `genderfemale` is a bit more tricky.
- It is estimated that women make ($100 \times (\exp(-0.257) - 1) = -22.7\%$) 22.7% less than men.
- For simplicity, however, we can say that women make approximately 25.7% less than men, but you should know that this interpretation is actually wrong.
- The advantage of using log *wage* as the dependent variable is that it allows the estimated model to capture non-linear effects.
- The 25.7% decrease in wages for women means that the dollar difference in wages between women and men in high-paying jobs (such as medicine) is larger than the dollar difference in wages between women and men in lower-paying jobs.

Log-log model for CO_2 emissions

In this section, we use data on per capita CO_2 emissions, and GDP per capita (data is from 2007). We will suppose that CO_2 emissions is the *dependent* variable. Load the data, and create the plot:

```
co2 <- read.csv("http://rtgodwin.com/data/co2.csv")
plot(co2$gdp.per.cap, co2$co2,
     ylab = "CO2 emissions per capita",
     xlab = "GDP per capita")
```

Figure: Per capita CO_2 emissions and GDP.



Consider this (possibly wrong) population model:

$$CO_2 = \beta_0 + \beta_1 GDP + \epsilon \tag{1}$$

- As GDP gets larger, CO_2 emissions are all over the place.
- The problem with model 1 is that GDP has the same effect on CO_2 everywhere (for all levels of GDP).
- Since energy consumption (which produces CO_2 emissions) is a relatively inelastic good, it may be reasonable to think that an increase in GDP per capita of say \$1000 has a much bigger impact on CO_2 emissions when GDP per capita is low.
- That is, their may be a non-linear relationship.

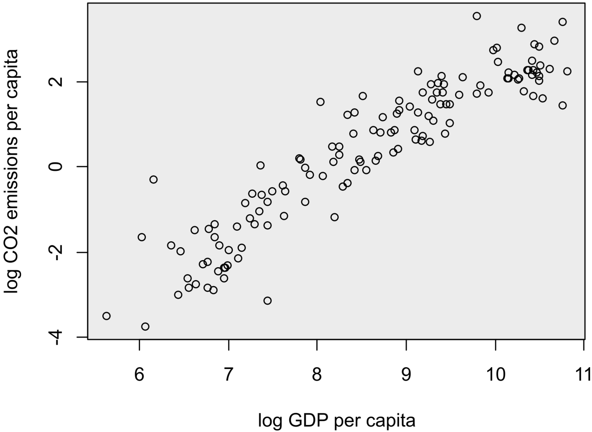
If we take the *logs* of CO_2 and GDP per capita, then we are saying that percentage changes in per-capita GDP lead to percentage changes in CO_2:

$$\log(CO_2) = \beta_0 + \beta_1 \log(GDP) + \epsilon \tag{2}$$

Plot the data:

```
plot(log(co2$gdp.per.cap), log(co2$co2),
     ylab = "log CO2 emissions per capita", xlab = "log GDP
    per capita")
```

Figure: Log per capita CO_2 emissions and log GDP.



Now, let's estimate model 2:

```r
co2mod <- lm(log(co2) ~ log(gdp.per.cap), data = co2)
summary(co2mod)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -9.94045    0.36806  -27.01   <2e-16 ***
log(gdp.per.cap)  1.20212    0.04234   28.39   <2e-16 ***
---
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6642 on 132 degrees of freedom
Multiple R-squared:  0.8593,  Adjusted R-squared:  0.8582
F-statistic: 806.1 on 1 and 132 DF,  p-value: < 2.2e-16
```

The interpretation of the results is that for every 1% increase in GDP per capita, it is estimated that CO_2 emissions increase by 1.2%.