Working with logarithmically transformed variables

Johan A. Elkink

October 12, 2016

During this class we often make use of the logarithmic transformation of a variable. The reason is that in political science, it is generally difficult to find relevant continuous variables—most variables we use are categorical, e.g. party voted for, democracy or not, war or not, etc. For regression analysis, however, which is at the core of this class, continuous variables are typically assumed. The most prominent continuous variables in political science are probably those related to the economy, i.e. money related variables. The distribution of financial data is typically very skewed, however—some people are very rich, but most people are not. This leads to problems in regression analysis.

There are a number of motivations, some more reasonable than others, to transform variables in this manner:

- In many cases (but not by definition!) it transformes a skewed variable into a much more symmetric one. The intuition behind this can be derived from Figure 1: on the left, there is a very steep curve, and you can see how a small distance on the x-axis translates into a long distance on the y-axis; on the right, there is a very shallow curve, and a large distance on the x-axis translates into a short distance on the y-axis. So, the distribution of a variable is stretched on the left and squeezed on the right, such that the high density of a skewed variable on the left gets spread over a wider range and the low density of a skewed variable on the right squeezed into a smaller range, thus making the variable more symmetric after the transformation. Figure 2 shows this for an artificially generated example with GDP—a typically skewed variable.
- In economic theory we will often have equations that are based on multipliers rather than being additive. For example, we might not have a function that looks like

  $$y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x2i + \varepsilon_i,$$

  but instead have something like

  $$y_i = \beta_1 \cdot p^{\beta_2} \cdot r^{\beta_3} \cdot \varepsilon_i.$$

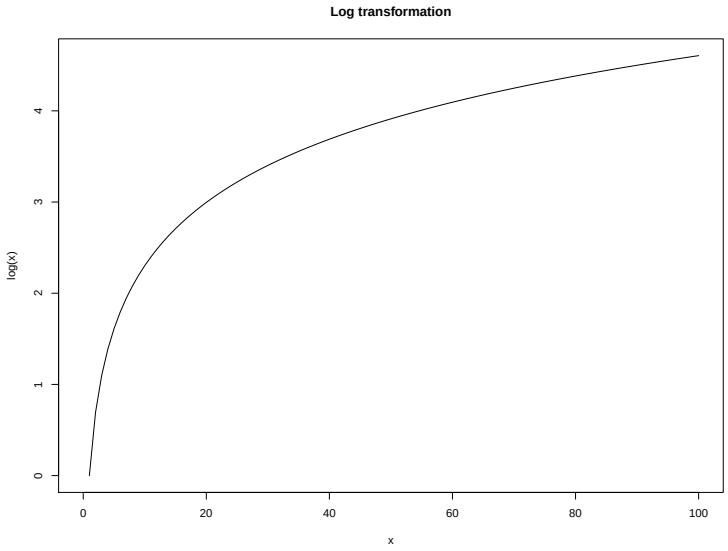

Figure 1: Shape of the log function

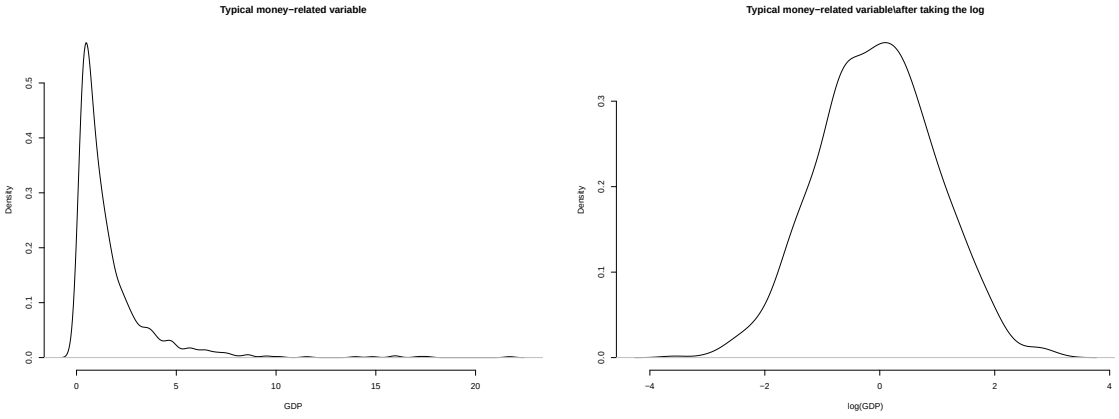

Figure 2: Distribution of a typical money-related variable (left) and the same distribution, after the log transformation.

This cannot be estimated with a linear regression, but we can do:

$$\log(y_i) = \log(\beta_1^* \cdot p^{\beta_2} \cdot r^{\beta_3} \cdot \varepsilon_i^*) = \beta_1 + \beta_2 \log(p) + \beta_3 \log(r) + \varepsilon_i,$$

where $\beta_1 = \log(\beta_1^*)$ and $\varepsilon_i = \log(\varepsilon_i^*)$. We can then run a standard linear regression to estimate the parameters. Money is of course generated in such a manner (e.g. r could represent interest rates), which is the reason it is skewed.

- If the underlying model is as described and the logarithmic transformation is reasonable, it also leads to a better fit of the model—R^2 will be better after the transformation.
- We will get to this much later in the class, but when we perform statistical inference based on regression analysis, we make the assumption that ε is normally distributed. If the dependent variable is highly skewed, this is unlikely to be the case—a log transformation of the dependent variable can potentially address this.

Taking this transformed variable makes interpretation harder, however. Normally, one would like to see an interpretation along the lines of: "If GDP per capita increases by 1,000 US dollar, we expect the level of corruption to decrease by 1.05 on our scale from 0 (no corruption) to 10 (high corruption)," but with a log transformed variable this is going to sound odd: "If GDP per capita increases by 1 log US dollar, we expect the level of corruption to decrease by 1.35 on our scale from 0 (no corruption) to 10 (high corruption)." What can we do in this scenario?

- After the logarithmic transformation, the relationship between x and y is not going to be linear anymore. This means that an increase in x from, say, 1 to 2, is not going to have the same effect as an increase from 100 to 101. This is often reasonable—going from poor to less poor should have a much bigger impact than going from very rich to even richer. One way to deal with this in terms of interpretation is to use numerical examples. For example, if we estimate the following model:

  $$corruption_i = \hat{\beta}_1 + \hat{\beta}_2 \log(GDPPC_i) = 3.04 + 1.05 \log(GDPPC_i),$$

  with GDPPC measured in 1,000 US dollars, we could have the interpretation above, but we could also calculate an example, like "if GDP per capita increases from 1,000 to 2,000 US dollar, we expect the level of corruption to increase from 3.0 to 3.8 on our scale from 0 (no corruption) to 10 (high corruption)." This would be based on the following calculations: $3.04 + 1.05 \cdot \log(1) = 3.04$ and $3.04 + 1.05 \cdot \log(2) = 3.767805$.
- When using logarithmically transformed variables, it is possible to interpret coefficients approximately as the effects of a percentage change. We get the following:[1]

  $y_i = \beta_1 + \beta_2 x_i$: one unit change in x corresponds to a β_2 unit change in y.

  $y_i = \beta_1 + \beta_2 \log(x_i)$: one percentage change in x corresponds to a β_2 unit change in y.

[1] See also `https://www.cscu.cornell.edu/news/statnews/stnews83.pdf`.

$\log(y_i) = \beta_1 + \beta_2 \log(x_i)$: one percentage change in x corresponds to a β_2 percentage change in y.

$\log(y_i) = \beta_1 + \beta_2 x_i$: one unit change in x corresponds to a β_2 percentage change in y.

So, in our previous example: "If GDP per capita increases by 1%, we expect the level of corruption to increase by 1.05 on our scale from 0 (no corruption) to 10 (high corruption)." (Note that this does not work if you standardize the variable after the log transform!)