PLSC 502: "Statistical Methods for Political Research"

Nonlinearity and Data Transformations
November 29, 2016

Why Transform?

It is often the case that we have to deal with data (either or both Ys or $\mathbf{X}$s) that are *skewed*; that is, data that have significant outliers in either direction. In most cases, its a good idea to transform the data into something more symmetrical, for at least three reasons:

1. Skewed distributions are difficult to examine graphically, since the outliers often mask interesting information.

2. In skewed distributions, the mean $\bar{X}$ is a lousy measure of the central tendency of X. As we noted at the outset, OLS relies heavily on means (in fact, OLS with a constant is just an estimate of the mean of Y).

3. Relatedly, skewed values of Y and/or $\mathbf{X}$ can lead to non-normality in the error terms, which in turn can render inferences suspect.

Moreover, there are sometimes theoretical reasons for transforming variables. Consider the following model:

$$Y_i = \beta_0 X_i^{\beta_1} u_i \tag{1}$$

This model says that Y has some baseline level (β_0), but that X has a multiplicative effect on that baseline, where the size of the multiplicative effect is scaled by a second coefficient (β_1). This sort of model might arise in a number of settings (e.g., as a model of "accelerated failure times"); however, it is not a model we can estimate via OLS. Note, however that if we take the natural logarithm of both sides, we get:

$$\ln(Y_i) = \ln(\beta_0) + \beta_1 X_i + \ln(u_i) \tag{2}$$

This model, by contrast, is easily estimated using standard linear-model technology; moreover, it yields the same estimates for β_0 and β_1 that would a nonlinear model. Likewise, a model like:

$$\exp(Y_i) = \beta_0 + \beta_1 X_i + u_i \tag{3}$$

can be transformed to yield

$$Y_i = \ln(\beta_0) + \beta_1 \ln(X_i) + \ln(u_i) \tag{4}$$

which can also be estimated and interpreted in a linear fashion.

Monotonic Transformations

Fox discusses the "ladder of powers" for transforming a variable X. He notes that these are best written as $X^{(p)} = \frac{X^p - 1}{p}$, though, in practice, most people just use the direct transformation. Both are straightforward:

Transformation	p	$f(X)$	Fox's $f(X)$
Cube	3	X^3	$\frac{X^3-1}{3}$
Square	2	X^2	$\frac{X^2-1}{2}$
(None/Identity)	(1)	(X)	(X)
Square Root	$\frac{1}{2}$	$\sqrt{X}$	$2(\sqrt{X}-1)$
Cube Root	$\frac{1}{3}$	$\sqrt[3]{X}$	$3(\sqrt[3]{X}-1)$
Log	0 (sort of)	$\ln(X)$	$\ln(X)$
Inverse Cube Root	-$\frac{1}{3}$	$\frac{1}{\sqrt[3]{X}}$	$\frac{\left(\frac{1}{\sqrt[3]{X}}-1\right)}{-\frac{1}{3}}$
Inverse Square Root	-$\frac{1}{2}$	$\frac{1}{\sqrt{X}}$	$\frac{\left(\frac{1}{\sqrt{X}}-1\right)}{-\frac{1}{2}}$
Inverse	-1	$\frac{1}{X}$	$\frac{\left(\frac{1}{X}-1\right)}{-1}$
Inverse Square	-2	$\frac{1}{X^2}$	$\frac{\left(\frac{1}{X^2}-1\right)}{-2}$
Inverse Cube	-3	$\frac{1}{X^3}$	$\frac{\left(\frac{1}{X^3}-1\right)}{-3}$

Fox also suggests using base-2 or base-10 logs, which, frankly, is never, ever done in political science.

The general rule to remember is that:

> **Using higher-order power transformations (e.g. squares, cubes, etc.) "inflates" large values and "compresses" small ones; conversely, using lower-order power transformations (logs, etc.) "compresses" large values and "inflates" (or "expands") smaller ones.**

Two other things:

1. In order for power transformations to "work", all the values of X need to be positive; this is easily accomplished by adding the absolute value of the lowest value of X (call that X_l), plus some small additional value ϵ, to X itself:

$$X^* = X + (|X_l| + \epsilon) \tag{5}$$

   - Note, however, that the value one adds can make a difference in what the transformation "looks like." One practical rule of thumb, then, is to choose ϵ such that it is equal to half of the difference between the lowest value of X and its next-lowest value.

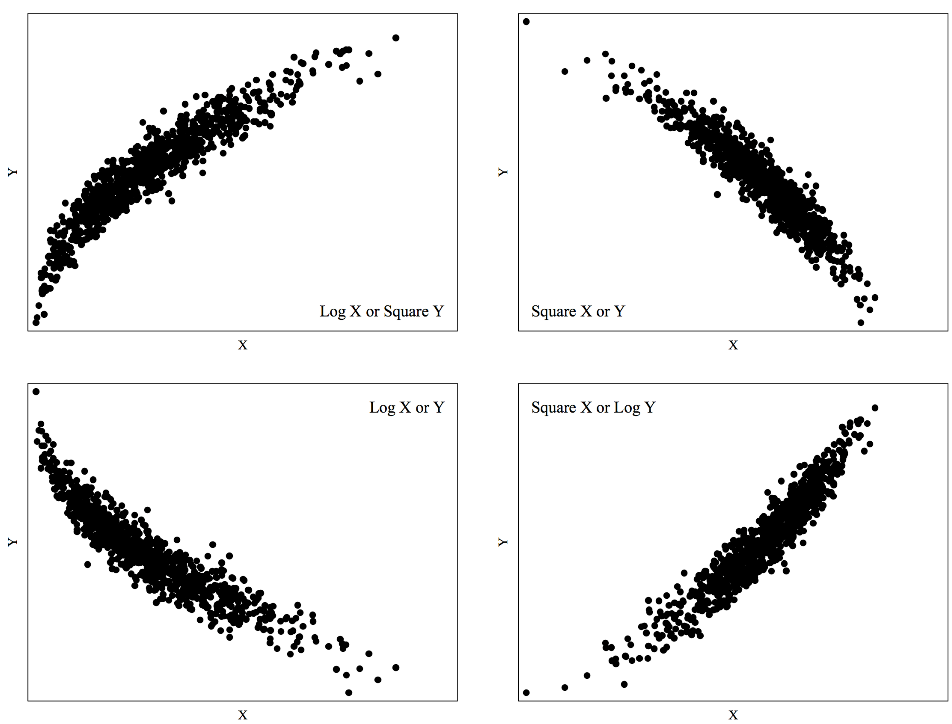
- So, if X is (say) annual growth rates for OECD nations, and the lowest value in the data is -3.3, while the second-lowest value is -3.0, then one could add $|-3.3| + [-3.0 - (-3.3)]/2 = 3.3 + 0.15 = 3.45$ to the data before transforming it.

2. In order for these transformations to be useful, the difference between the largest and smallest values in the data has to be significant (that is, the ratio of them has to be on the order of five or greater). Again, this can often be accomplished by adding a suitable "start value" to the data – so, for example, if we're talking about years, use $\text{Year}_i - 1900$ rather than Year_i in the transformation.

Which Transformation to Use?

That depends on the situation. The most common reason to use transformations is to induce linearity between Y and one or more Xs. Tukey and Mosteller (two gurus of exploratory data analysis) suggest a "bulging rule" for power transformations to make things more linear:

Figure 1: Tukey and Mosteller's "Bulging Rule" for Monotone Transformations to Linearity



Nonmonotonicity

What happens when we have a curvilinear relationship between Y and X?...

- Question is whether the function "turns over" or not...
- If not, can use the power transformations we just discussed... (e.g. $Y_i = \beta_0 + \beta_1\sqrt{X}_i + u_i$)
- If so, we can use *higher-order polynomial terms* in the regression to model this...

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i,$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i,$$

and so forth. Fox points out that X-Y and simple residual plots aren't necessarily the most effective way of adducing whether or not nonmonotonicity is an issue or not. Using theory is also a good idea.

Of course, there's a lot more to curvilinearity than this. We'll talk more about some robust methods for dealing with complicated curvilinear relationships a bit later in the course.

Interpretation of Transformed Variables

In general, interpretation of results in a regression with transformed (X or Y) variables can be done in the same way as with untransformed variables, so long as one is careful to be aware of the transformation(s). One-sided log transformations are among the more intuitive of these, once you understand that a log transformation converts the linearity in the model from an additive factor to a multiplicative one.

So, for:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i, \tag{6}$$

it's clear that

$$\mathrm{E}(Y) = \exp(\beta_0 + \beta_1 X_i).$$

This, in turn, means that the effect of a one-unit increase in X on the expected value of Y is

$$\frac{\partial \mathrm{E}(Y)}{\partial X} = \exp(\beta_1).$$

A change in X thus impacts Y multiplicatively; thus, if (say) $\beta_1 = 0.69$, then $\exp(\beta_1) = 2$; in other words, a one-unit change in X leads to a doubling of $\mathrm{E}(Y)$. Conversely, if $\beta_1 = -0.69$, then $\exp(\beta_1) = 0.5$, meaning that a one-unit change in X decreases $\mathrm{E}(Y)$ by about 50 percent (that is, cuts it in half).

A similar dynamic holds for the regression where X is transformed:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i. \tag{7}$$

Here, the expected value of Y varies linearly in the log of X; substantively, that means that the effect of X is also multiplicative, but in the "opposite direction." That is, now a multiplicative increase in X has a linear increase on Y of β_1 magnitude.

So suppose we double the value of X (say, from X_ℓ to $2X_\ell$). Then the associated change in $\mathrm{E}(Y)$ is

$$\begin{aligned}
\Delta \mathrm{E}(Y) &= \mathrm{E}(Y|X = 2X_\ell) - \mathrm{E}(Y|X = X_\ell) \\
&= [\beta_0 + \beta_1 \ln(2X_\ell)] - [\beta_0 + \beta_1 \ln(X_\ell)] \\
&= \beta_1[\ln(2X_\ell) - \ln(X_\ell)] \\
&= \beta_1 \ln(2)
\end{aligned}$$

A Special Case: Log-Log Regressions

One special case is the instance where both Y and X have been transformed by taking (natural) logarithms. For the regression

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + ... + u_i \tag{8}$$

the slope coefficient β_1 can be interpreted as the *elasticity* of Y with respect to X. An elasticity is probably best thought of as the ratio of the percentage change in one variable to the percentage change in the other; here, think of it as:

$$\text{Elasticity}_{YX} = \frac{\%\Delta Y}{\%\Delta X}.$$

This turns out to be a very useful thing in many cases. For example:

- it leads to a direct means of interpreting a nonlinear effect: a one-percent change in X leads to a $\hat{\beta}_1$-percent change in Y.
- It also can be (equivalently) thought of as a doubly multiplicative relationship: an increase in X of some multiplicative factor δ leads to a multiplicative increase in Y equal to δ^{β_1}.
- So, for example, if we have an estimate of $\hat{\beta}_1 = 0.5$ in a log-log regression, and we want to know the effect on $\mathrm{E}(Y)$ of tripling X, that effect is equal to $3^{0.5} \approx 1.73$. That is, tripling X will increase $\mathrm{E}(Y)$ by a factor of 1.73 (i.e., increase it by 73 percent).

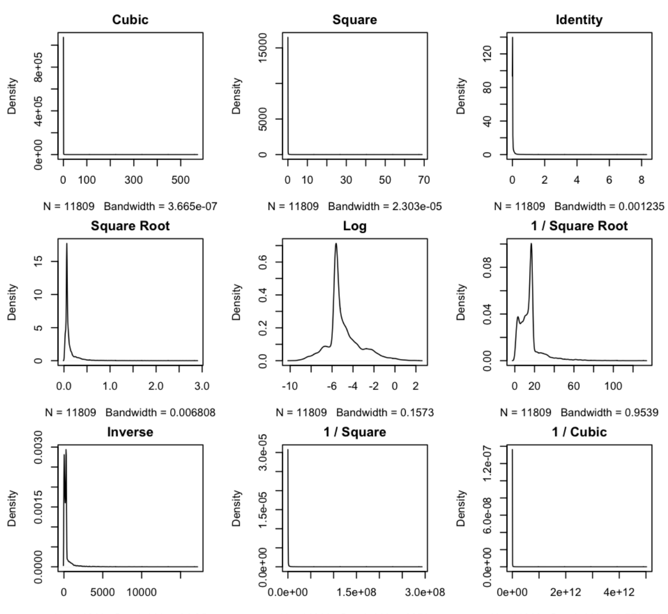
A Brief Example

We'll go back to the Fordham and Walker example on military spending, and consider a simple, two-variable relationship: That between military spending (as a percentage of GDP) and GDP itself. The GDP variable would seem to be a particularly good candidate for transformation, since cross-national GDP data is inevitably massively right-skewed. And, these data are no exception:

```
> library(TSA)
> skewness(gdp,na.rm=TRUE)
[1] 15.14007
> skewness(milgdp,na.rm=TRUE)
[1] 8.450093
```

That is, it is clear that there is skew in both variables, and that some sort of transformation is probably a good idea; the Tukey/Mosteller rule suggests we might consider logging GDP, spending, or both. The ladder-of-powers plots for the two variables are here:

Figure 2: Ladder-of-powers Plot, GDP Variable



The ladder plot suggest that both military spending and GDP are highly right-skewed, and that a log transformation is probably going to bring them closest to being more normally distributed. Turning to the relationship between the two, a scatterplot of the two variables looks like this:

Figure 3: Ladder-of-powers Plot, Military Spending Variable

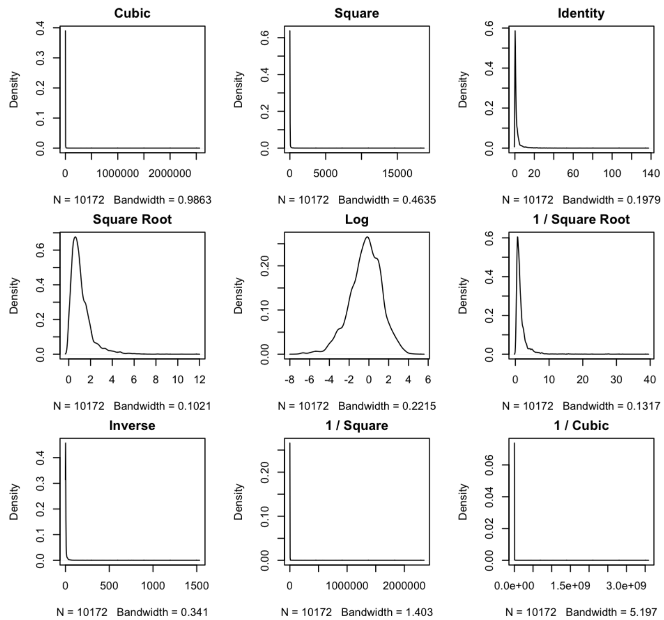
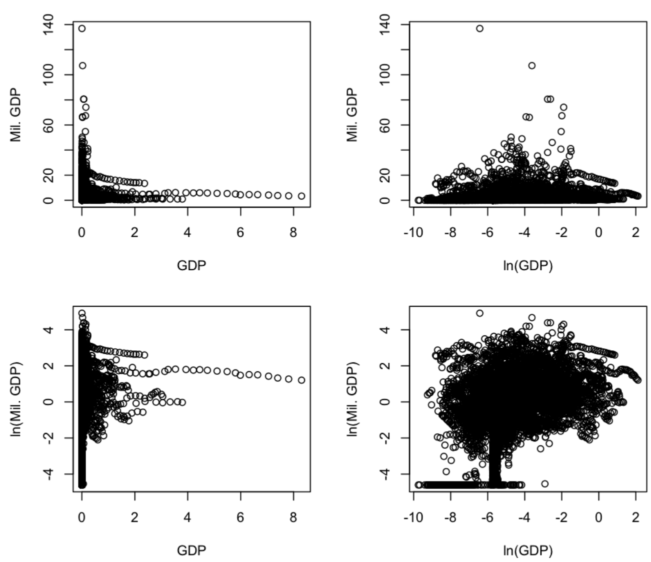


Figure 4: Scatterplots of Military Spending and GDP



Note a few things:

- Because the variables are so badly right-skewed, logging each of them fixes the scale for one, but not the other. As a result, spending vs. logged GDP looks a bit like logged spending vs. GDP if the latter was "turned on its side."

- Probably the most "linear" of the three is the log-log plot.
- Also, notice that the horizontal line of values on the log-log plots is a function of the (largely arbitrary) value that we set zero values of military spending to before we logged that variable.

Of course, the important thing is to assess how these transformations affect the regression results, particularly the values and distribution of the residuals. The regressions are:

```
> summary(lm(milgdp~gdp))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0538     0.0481  42.696  < 2e-16 ***
gdp           1.0038     0.1540   6.518 7.45e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.757 on 10170 degrees of freedom
  (4327 observations deleted due to missingness)
Multiple R-squared: 0.00416,Adjusted R-squared: 0.004062
F-statistic: 42.49 on 1 and 10170 DF,  p-value: 7.454e-11


> summary(lm(milgdp~log(gdp)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.60137    0.13969   32.94   <2e-16 ***
log(gdp)     0.52196    0.02766   18.87   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.686 on 10170 degrees of freedom
  (4327 observations deleted due to missingness)
Multiple R-squared: 0.03384,Adjusted R-squared: 0.03374
F-statistic: 356.2 on 1 and 10170 DF,  p-value: < 2.2e-16


> summary(lm(log(milgdp+0.01)~gdp))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.45918    0.01669  -27.51   <2e-16 ***
gdp          0.75794    0.05343   14.18   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.651 on 10170 degrees of freedom
  (4327 observations deleted due to missingness)
Multiple R-squared: 0.0194,Adjusted R-squared: 0.0193
F-statistic: 201.2 on 1 and 10170 DF,  p-value: < 2.2e-16


> summary(lm(log(milgdp+0.01)~log(gdp)))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.644270   0.044736   36.76   <2e-16 ***
log(gdp)    0.431875   0.008858   48.76   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.501 on 10170 degrees of freedom
  (4327 observations deleted due to missingness)
Multiple R-squared: 0.1895,Adjusted R-squared: 0.1894
F-statistic:  2377 on 1 and 10170 DF,  p-value: < 2.2e-16
```

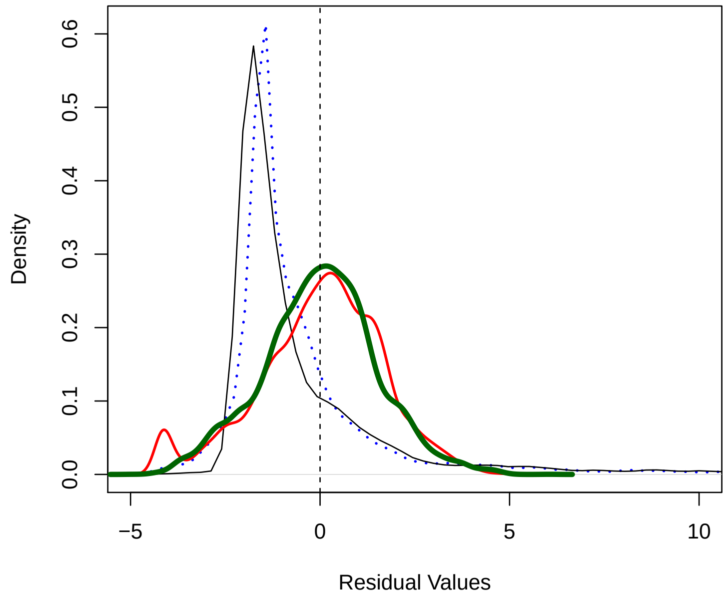
Of these, note that:

- The unlogged regression is clearly the weakest fit; that one suggests that there is a more or-less one-to-one relationship between the variables (i.e., that a $1 billion increase in GDP corresponds to a one percent increase in spending).
- Using the log of GDP provides a stronger fit; that interpretation would be that a one-unit change in the log of GDP (that is, essentially, a doubling of GDP) yields a 0.5 percent increase in military spending.
- As we suspected from the scatterplot, the best fit is that provided by the log-log model.

We can examine density plots of the residuals to get a sense of how normal they are:

The residuals from the model with the transformed Y are substantially more normal than those with an untransformed Y. Given these results, along with the models above, then, we'd almost certainly favor the log-log model.

Figure 5: Kernel Density Plots of Residuals



Final Tips

- As always, **theory is valuable**. For example, expectations of diminishing returns to scale are a good way to motivate logging a variable.
- **Try different things**, and **look at plots**. Your eyes are usually the best gauges of whether or not something is linear, normal, etc.
- **It takes practice**. This is the sort of thing that is a bit closer to "art" than most statistics. You get better at it over time.