Simple Linear Regression

Transformations in Linear Regression

Transformation of the data can sometimes improve situations where a linear model is not appropriate or when the conditions needed to carry out inference on the slope don't appear to hold.

Example 1

For the data of life expectancy and GDP per capita for most countries in the world, a straight line is not the best model. GDP is a measure of the standard of living based on material possessions. It seems reasonable to expect that wealthier countries, with higher GDPs, will have larger life expectancies. From the scatterplot (Figure 1), it does seem that countries with high GDP tend to also be the counties with higher life expectancies. But the relationship is definitely not linear.

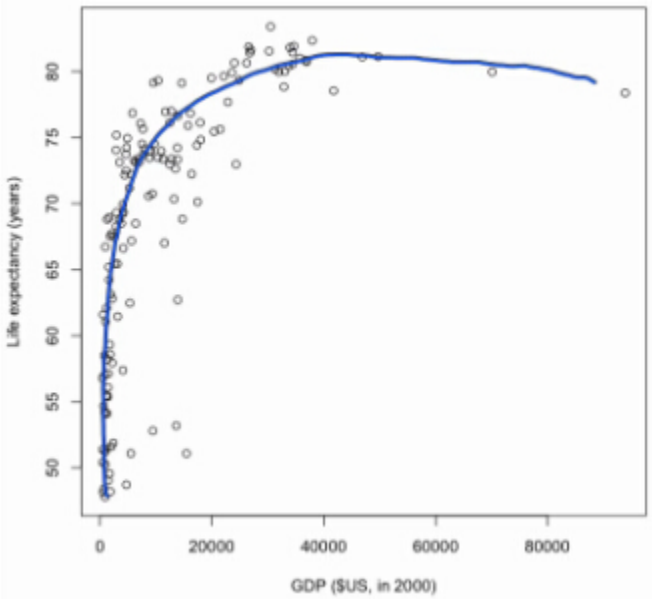

Figure 1: Life expectancy and GDP: An example of a non-linear relationship.

Mathematically transforming data in statistics can serve many purposes, including turning relationships that are curved into a straight line relationship so that linear regression methodology can be applied. The particular transformation that we'll consider is the logarithm, which is the most commonly used transformation. It's the most commonly used because it is effective in many situations in producing transformed data that are easier to work with than the original data, and because it gives results that can be easily interpreted.

Typically, statisticians use the natural logarithm, but the base 10 logarithm works equally well and for simplicity, that's what we'll work with here.

Some useful facts about the base 10 logarithm:

$$\log_{10}(a) = b \text{ means } 10^b = a$$

$$\log_{10}(10) = 1,\ \log_{10}(100) = 2,\ \log_{10}(1000) = 3,...$$

$$\log_{10}(a \times b) = \log_{10}(a) + \log_{10}(b)$$

$$\log_{10}(x) + 1 = \log_{10}(x) + \log_{10}(10) = \log_{10}(10x)$$

Here is why the logarithm transformation will be useful for us in looking at the relationship between life expectancy and GDP. The distribution of GDP is very right-skewed. Many countries in the world are not very wealthy, with low GDPs, and there are fewer of the wealthiest countries. with high GDPs. The logarithmic function changes quickly for small values and changes slowly for larger values. As a result, the log transformation of a variable will spread out the small values, and bring the large values closer in to the other values. So taking the logarithm of a variable that has a right-skewed distribution results in a transformed variable with a distribution that is often closer to symmetric.

Figure 2 shows a histogram of GDP both before and after a base 10 log transformation, and the scatterplot of life expectancy versus GDP before and after the transformation. The histogram of the log transformation of GDP is much closer to symmetric than the untransformed GDP and in the scatterplot the values of log transformed GDP are more spread out and we can see a linear relationship.

Using the method of least squares to find the regression line, we get a line with intercept = 12.26, slope = 14.93. The P-value for the test with null hypothesis that the slope is 0 is < 0.0001, so there is very strong evidence that this slope is different from 0. Interpreting the slope requires a little more care when we've transformed one or more of our variables. In general, the slope tells us how y changes, on average, when x increases by 1 unit. So when log of GDP increases by 1, life expectancy increases by an average of 14.9 years. But what does it mean for log of GDP to increase by 1? We need to use this property of logarithms:

$$\log_{10}(x) + 1 = \log_{10}(10x),$$

So our slope of 14.9 years is the average amount life expectancy increases when GDP is multiplied by 10. Interpretation of log-transformed variables typically involves talking about changing variables by multiplicative factors, such as multiplying GDP by 10. Many variables tend to grow in a multiplicative rather than additive way and the log-transformed version of these variables often work well in statistical models.

There are a few points that don't fit the regression line well. These can be seen in the scatterplot and the plot of the residuals in Figure 3. The corresponding residuals are large

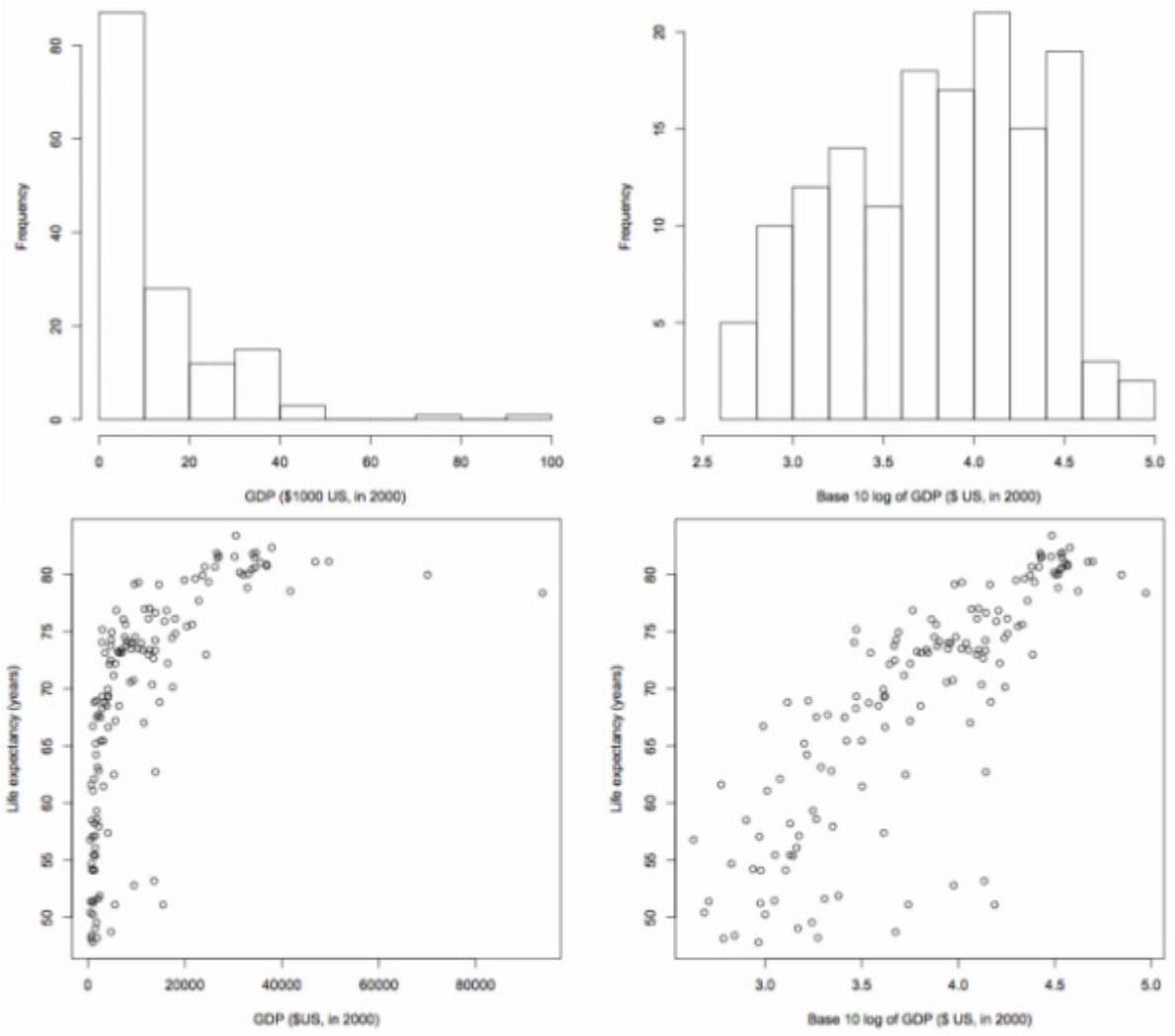

Figure 2: Histograms of GDP before and after base 10 log transformation and scatterplots of life expectancy versus GDP before and after base 10 log transformation.

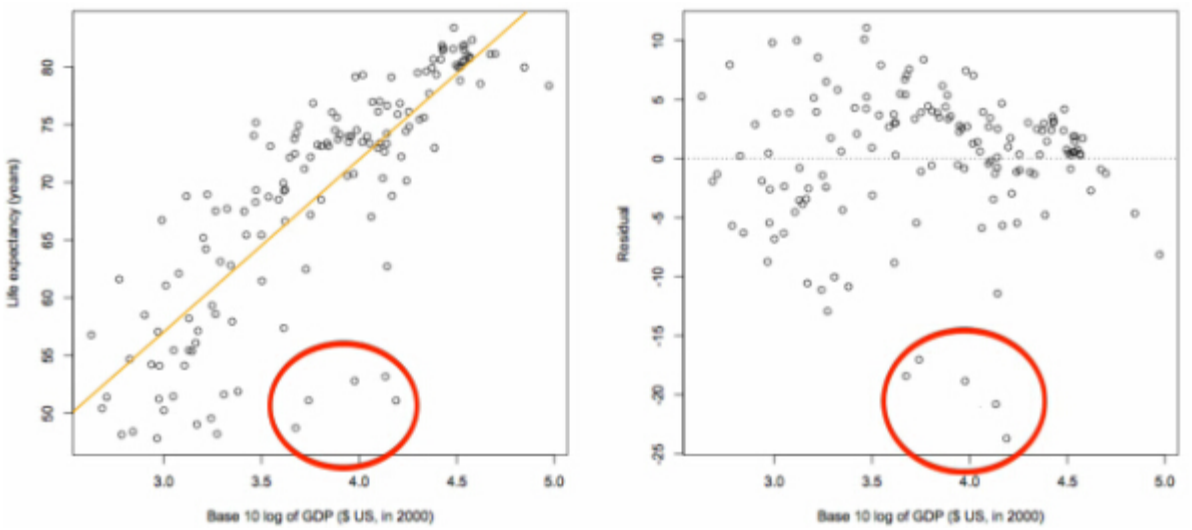

Figure 3: Points that aren't well-explained by the regression line.

in absolute value. The linear relationship between life expectancy and log of GDP does not explain these countries well. These countries are Swaziland, Angola, South Africa, Botswana, and Equatorial Guinea. These are countries with lower life expectancies than their GDP predicts, and there is typically a good reason why. For example, Equatorial Guinea is a large oil producer which has created wealth, but it is not distributed equally. The country ranks quite low on the United Nations Human Development Index; less than half of the population has access to clean drinking water and 20% of children die before age 5. In Botswana, mining of gemstones and precious metals contributes to its wealth, but it has a relatively low life expectancy because about one-quarter of its population is infected with HIV. As this example illustrates, while the linear relationship explains much of what is going on in the data, sometimes the points that don't fit the model tell an interesting story.

Log transforming the response variable can sometimes help meet the conditions of linear regression.

Example 2

Figure 4 shows a scatterplot of the amount of carbohydrates in food items sold at a coffee shop on the vertical axis, and the number of calories in the food items on the horizontal axis, and a residual plot from the linear regression fit to these data. From the scatterplot, we can see a positive relationship, but it also shows increasing variability in the amount of carbohydrates with greater calories. This increasing variability is also seen in the residual plot. This pattern of increasing variance means that it would not be appropriate to carry out inference on the slope of the line for these data. The lower two plots are the corresponding scatterplot and residual plot after taking a base 10 log transformation of the response variable. There's still lots of scatter, but the variability in the y's seems roughly the same for all values of x.

To interpret the slope when the response variable has been log transformed:

$$\log_{10}(y) = b_0 + b_1 x \implies y = 10^{b_0} 10^{b_1 x}$$

and

$$10^{b_1(x+1)} = 10^{b_1 x} 10^{b_1}$$

So increasing x by 1 is equivalent to change y by multiplying it by 10^{b_1}. For this particular regression, the estimate for the slope is

$$b_1 = 0.00122 \text{ and } 10^{b_1} \doteq 1.0028$$

We can interpret this as follows: An increase in calories of 1 is associated with multiplying amount of carbohydrates by approximately 1.03. That is, an increase in calories of 1 is associated with an increase in amount of carbohydrates of about 0.3%. Again, interpretation of log-transformed variables typically involves talking about changing variables by multiplicative factors, in this case, multiplying our y by 10 raised to the power of the estimated slope.

As with the life expectancy-GDP example, and as is often the case with real data, there is a little more to the story. From the residual plot (Figure 4) using the log transformation of calories we see some large negative residuals. These come from points that don't fit the line particularly well. They are all below the line, so the line is underestimating their amount of carbohydrates relative to their calories. To investigate this further, we re-plot the data with different symbols for the types of foods (Figure 5). The points that the model doesn't explain well are bistro boxes, which are higher in protein than most of the other food items.

This illustrates a common error made in carrying out regression. The common mistake is to treat data from different groups as if they all came from one big group. Type of food itme is a **lurking variable** here. That is, it is an important variable that helps explain the relationship, but hasn't been accounted for in the analysis.

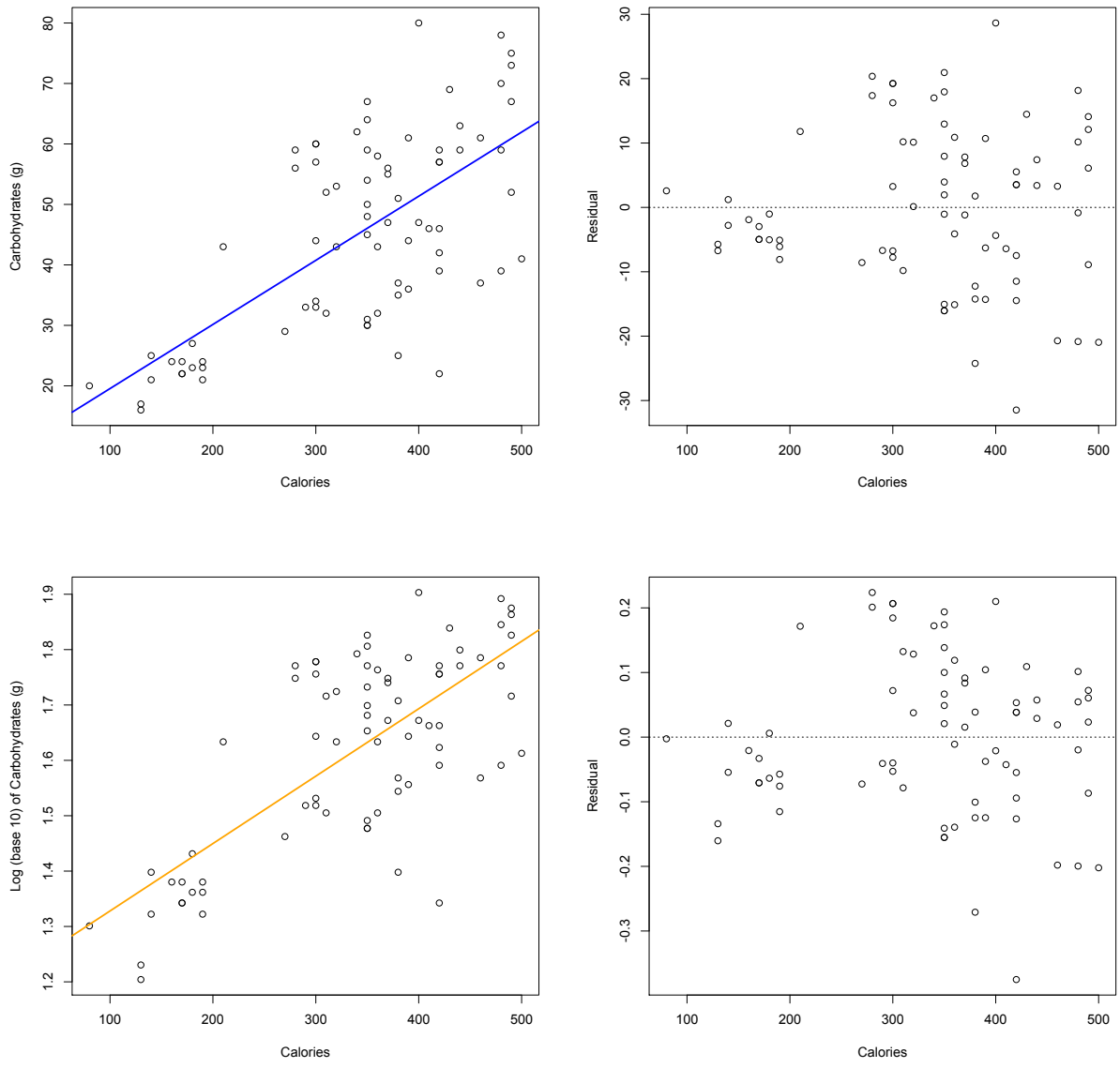

Figure 4: Scatterplot and residual plot for regression of amount of carbohydrates on calories before and after log transformation of the response variable.

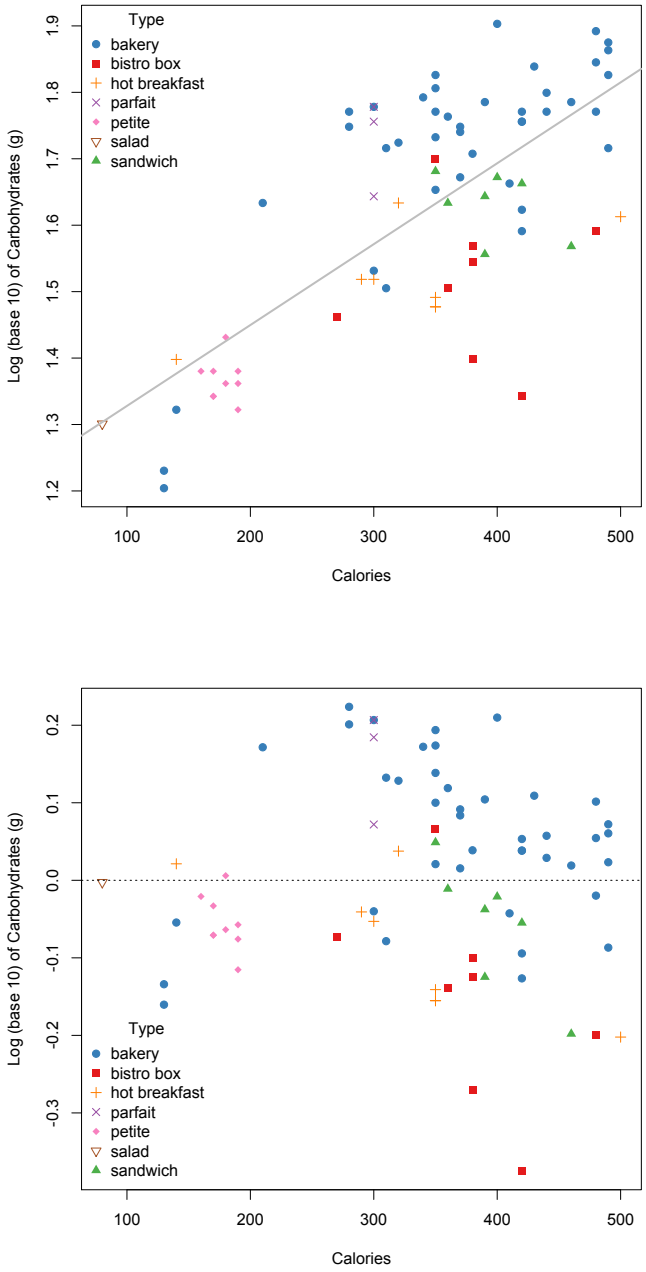

Figure 5: Scatter and residual plots for regression of amount of carbohydrates on calories, coded by type of food item.