Assignment

For this assignment, you’ll use the Quality of Governance (QOG) dataset from the University of Goetheburg, Sweden. The codebook on the analysis homepage describes it. Basically, the data are a set of global, country level indicators measured in 2002, on a wide range of topics. Some of the topics are theoretical, involving measures of various qualities of democratic governance, while others are measures of Gross Domestic Product (GDP) or carbon dioxide emissions. The dataset is stored in a workspace you’ll open, the data frame being labelled ‘q’.

This week’s assignment will be about transforming data — country level indicators from roughly 2002 — and comparing the transformed and untransformed measures in scatterplots.

In this assignment sheet I explain the purpose of the transformations, provide an example, and point you toward a couple of refresher resources if you need them. The script file for the assignment, ‘Lab 5.R’ provides the code for doing a couple of examples, including loading the data. You’ll need to review this assignment sheet, the QOG codebook and

Before getting to the assignment, let’s think about an underlying problem in data analysis – highly skewed variables. Lots of variables are skewed. Because of the sheer diversity of global social and economic conditions around the world, most country level measures contain skewness and need to be re-scaled to be useable.

Transformations

In some statistical analyses, data may need to be transformed prior to analysis. This transformation is often necessary in order to meet the underlying assumptions of the analytical model. Transformation means we apply specific mathematical operators on each observation within a variable, and then use that observation within an analysis. Transformations are often useful for inspecting bi-variate relationships in scatterplots, particularly when the measures are highly skewed. An example will illustrate this idea.

Again, because there is so much variation around the world in country conditions, measures at the nation-state (country) level are often plagued with skewness: Let’s take a basic fact that infant mortality varies widely around the world. One measure of this concept is ‘neonatal mortality rate’, as the “Probability of death from birth to age 1 month, expressed as deaths per 1,000.” (QOG codebook, 49). The histogram in Figure 1 displays the skewness. As you’ll see in Figure 1, most countries have low rates, but others have high rates.

By far the most common transformation to reduce skewness is a logarithmic transformation. For any variable that is strictly positive (greater than 0) variable, by far the most important transformation to be familiar with is the log transformation. This is simply taking the log of the individual observation. This is simply taking the log of the individual observation. Do you remember doing this in high school or another course, such as STA 215? No? No problem! To refresh your memory, take a look at the website listed in the Resources section at the bottom of the page. For now, just think of logarithms as a mathematical operation used to reduce skewness in data. We’ll get more complicated later in the semester.

Most commonly, researchers use the natural log (‘ln’, to the base e 2.718 or so), but any base can be used (such as common logs to the base 10), as the choice simply means a difference in score by a constant factor. Let’s look at the log transformation, which in R is simply accomplished with
Neonatal death rates are highly skewed. Many countries have low death rates, while some have very high rates. A log function.\(^1\) The next histogram, figure 2, shows the log transformation, with a much more symmetric distribution. This is what log transformations do – they spread out the small values and bring in (compress) the large outlier values.

If you’ve taken any course on international relations or comparative politics, you’ve probably seen a measure of a country’s economy (Gross Domestic Product — GDP) expressed as ‘logged GDP’, or GDP in log scale. This log scaling is performed for the same reason — skew. Look at the density plots in figures 3-4, showing two scalings of GDP around the world: one is “GDP converted to constant 2005 international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the US dollar has in the United States.” (QOG codebook, 46), while the other is the log scaling of this measure.

Why does all this matter? Compare the following two scatterplots. One shows neonatal mortality by GDP in the original scaling of each variable. Figure 5 displays infant mortality and GDP untransformed. The figure is nearly uninterpretable.

The second shows each variable log transformed. The difference? Without the log transformation the relationship is nearly impossible to detect — because of the presence of skewness and the outliers that affect the resolution of the figure. And inspecting the second figure shows that there is a negative relationship, as would be expected. GDP and neonatal mortality are negatively correlated. Some countries are unusual, however, and these are labelled by their country name.

\(^1\)Within the dataset, I created a new variable called ‘neonate’, q$neonate<-q$ihme_nm, then took the natural log transformation: q$logneonate<-log(q$neonate). The new variable, ‘logneonate’ is the natural log transformation of ‘neonate’.
Figure 2: Natural log of neonatal death rates. Notice how the skewness is reduced. The scaling of the histogram is no longer the original probability of death from birth to one month old, per capita (per 1000 persons), but now the natural log of this rate.

Figure 3: Country level Gross Domestic Product (GDP) in constant 2005 dollars.

The Assignment

Your assignment is to do the following:
1. Identify two variables from the dataset that are numerically scaled – interval or ratio, not a count variable (such as a ranking of countries on a scale of 1, 2 or 3). You may use GDP, infant mortality, or CO2 emissions, but not two or more of these.

2. Inspect the skewness of each variable. As necessary, create a new version of the variables with a log transformation.

3. Produce two scatterplots. One with the untransformed variables. The other with the transformed variables. How do the figures change? Is there evidence to suggest that the two variables are correlated?

4. Post figures for investigating skewness, scatterplots, and a writeup to your lab journal.

I’m working on a screencast with an additional example. I think, though, this assignment sheet should be useful for getting started. If you need help, post questions on the Blackboard Discussion Board. I’ll send you an email when the screencast is available.

**Resources for learning more about logarithms and exponents**

I will not ask you in class or on an exam to calculate a logarithm or exponent, but you should be familiar with the idea of each one. If you need a review, look at the introductory lessons on these subjects at the following two links: For logarithms, https://www.khanacademy.org/math/algebra/logarithms-tutorial and exponents https://www.khanacademy.org/math/algebra/exponent-equations.
Figure 5: Scatterplot of Infant mortality and GDP, untransformed. Notice the vertical line of points on the left hand side of the figure. These points represent all of the very poor countries, with low GDP, and a highly variable infant mortality rate. There is just too little resolution in the graphic, however, to see the relationship, although the green linear trend line suggests there may be a negative relationship between GDP and infant mortality.
Figure 6: Scatterplot of logged infant mortality and GDP. Once both variables are transformed, the relationship between the two is much clearer. The transformations show a much clearer linear, negative, relationship between the two variables. A few countries fall a bit far off the trend line, such as India and China with high GDP and high infant mortality, and Iceland with a modest GDP but very low infant mortality.