Data Transformations

Gener Aviles-R

March 9, 2017

In the process of data mining, after making sure that our data is tidily displayed (Wickham and others 2014), one of the first steps is to see what distribution the values of each variable follow. A normal distribution is prefered becuase it allows us to use a wider variety of tested tools in the analysis. If the values of a variable do not present a normal distribution then **data transformation** techniques can be used to move the behaviour of the values *closer* to normality.

Data Transformation Generalities

Data transformation can be defined as the modification of every value in a data set by a mathematical function of that same value. This is done to modify the shape of a distribution or relationship of a variable (Cox 2007).

Derived from the previous paragraph we can then mention some of the reasons one would want/need to transform data:

1. **Normality assumption of analytics tools.**
    - Most analytic softwares and mathematical tools available and tested work under the assumption that the variables are normally distributed.
2. **Convenience.**
    - This is usually done for understanding purposes. For example, in some cases it is easier to expres percentages rather than the original value.
3. **Reducing skewness.**
    - Dealing with a distribution that is symmetric (the length and slope of it's tales is very similar or equal) is easier to work with and interpret than with a skewed one.
4. **Equal variability.**
    - Data with equal or very similar variability is easier to work with, the results obtained from analysis on it will tend to be more robust. This phenomenon of homogeneity of variance is known as **homoscedasticity**. This is assumed in different tests and procedures throughout the analysis.
5. **Linear relationships.**
    - When working with two or more variables, relationships become very important when running inferential statistics. It is easier to deal with linear relationships than those of other orders.

A violation of these guidelines can produce an increased probability of falling into Type I or Type II errors depending on the analysis and which specific violation was done (Osborne 2010). Despite the clear mathematical arguments on the topic of data transformation, there is evidence that it is still common practice not to put special attention to the normality of data before drawing conclussions and publishing them (Osborne 2008).

Traditional Data Transformations

Square root transformation: x^2

- **What does it do to data?**
    - All values are raised to the $\frac{1}{2}$ power.
- **When is mostly used?**
    - Frecuently used to normalized Poisson distributions.
- **Particulars**
    - It does not work with negative numbers.

  - It behaves different with numbers between 0 and 1, the square root of these numbers is greater than the original value whereas it diminishes with numbers greater than 1.

Logarithms transformations: $log_{10}(100) = 2$

- **What is it?**
  - Logarithm is the exponent a base number must be raised to in order to get the original number.
  - Data transformation with this tool is more a family of transformations than just one approximation (Log, Natural Log, etc).
- **When is mostly used?**
  - Seems to be more common when outcomes are influenced by many independent factors (like biology and the social sciences).
- **Particulars**
  - Does not work with negative numbers.
  - Numbers between 0 and 1 behave different than those ≥ 1.

Inverse transformation: $x = \frac{1}{x}$

- **What is it?**
  - Produces very small numbers.
  - It reverses the order of the scores.
- **When is mostly used?**
  - ...
- **Particulars**
  - Numbers between 0 and 1 are treated differently.

The Cox-Box Transformation

If one looks carefully to the *traditional transformations* previously presented, it becomes clear that they manipulate data through elevating the values to a give power. What the Box-Cox transformation does is precisely this, thourgh brut force it looks for the ideal power to which any given set of values could be elevated producing a more normal distribution of them. Once this ideal power to elevate the values to is found it is then utilized to run the transformation on the whole subset. This *ideal* value is represented by λ. Mathematically the process is expressed as follows:

$$y_t^\lambda = (y_t^\lambda - 1) \text{ where } \lambda \neq 0$$

$$y_t^\lambda = log_e(y_i) \text{ where } \lambda = 0$$

Therefore Box-Cox can be seen more as a family of transformations than a singular process. It wil adjust the value of λ to the specific dynamics of the data in each instance. Some λ values can be identified as the *traditional transformations*:

$\lambda = 1.00$: no transformation needed, produces results identical to original data

$\lambda = 0.50$: square root transformation

$\lambda = 0.33$: cube root transformation

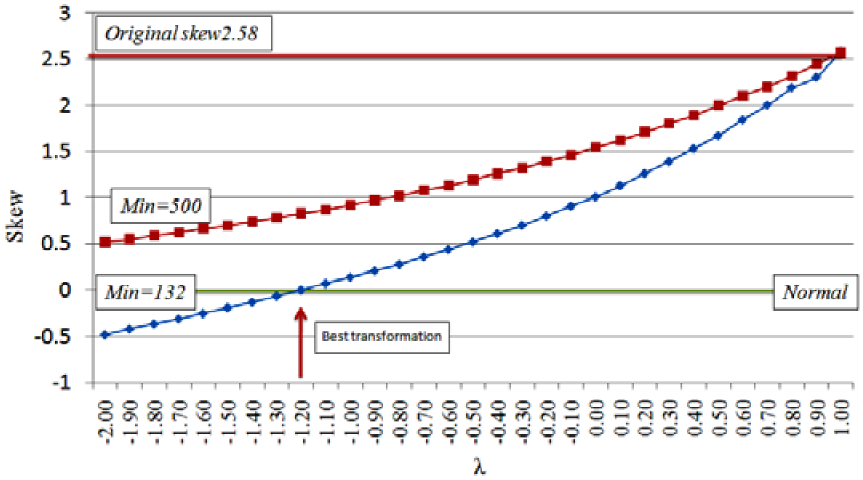
$\lambda = 0.25$: fourth root transformation

$\lambda = 0.00$: natural log transformation

$\lambda = -0.50$: reciprocal square root transformation

$\lambda = -1.50$: reciprocal (inverse) transformation

The following image represents a Box-Cox transformation done by Jason W. Osborne (Osborne 2010) on data from USA Universities size and average faculty salary:



We can observe that the original **skwedness** of the data was 2.58 (very much positively skewed) and when running the transformation it produced some very bad results (far from a normal distribution) but with a λ of -1.20 it produced a distribution with skew values of 0, meaning: **normality**. The red and blue lines represent values to which the whole set of data was *anchored*, meaning that a prvious transformation was performed to have the subset of data starting at a minimum of 500 in their values for the red line and 132 for the green one.

This gives us an exmple of how a Box-Cox transformation can be performed and interpreted. Eventhough it will not produce optimal results in every case, it definately comes in handy during the data transformation phase.

References

Cox, Nicholas J. 2007. "Transformations: An Introduction." July. http://fmwww.bc.edu/repec/bocode/t/transint.html.

Osborne, Jason W. 2008. "Sweating the Small Stuff in Educational Psychology: How Effect Size and Power Reporting Failed to Change from 1969 to 1999, and What That Means for the Future of Changing Practices 1." *Educational Psychology* 28 (2). Taylor & Francis: 151–60.

———. 2010. "Improving Your Data Transformations: Applying the Box-Cox Transformation." *Practical Assessment, Research & Evaluation* 15 (12). Citeseer: 1–9.

Wickham, Hadley, and others. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). Foundation for Open Access Statistics: 1–23.