IDS 702: Module 1.8

Transformations

Dr. Olanrewaju Michael Akande

TRANSFORMATIONS

- As we have already seen, sometimes, we have to deal with data that fail linearity and normality.
- Transforming variables can help with linearity and normality (for the response variable, since we do not need normality of the predictors).
- The most common transformation is the natural logarithm. For the response variable, that is, $\log_e(y)$ or $\ln(y)$.
- This is often because it is the easiest to interpret.
- Suppose

$$\ln(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i.$$

- Then it is easy to see that

$$y_i = e^{(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i)} = e_0^{\beta} \times e^{\beta_1 x_{i1}} \times e^{\beta_2 x_{i2}} \times \ldots \times e^{\beta_p x_{ip}} \times e^{\epsilon_i}.$$

- That is, the predictors actually have a multiplicative effect on y.

Natural log transformation

- The estimated β_j's can be interpreted in terms of approximate proportional differences.
- For example, suppose $\beta_1 = 0.10$, then $e^{\beta_1} = 1.1052$.
- Thus, a difference of 1 unit in x_1 corresponds to an expected positive difference of approximately 11% in y.
- Similarly, $\beta_1 = -0.10$ implies $e^{\beta_1} = 0.9048$, which means a difference of 1 unit in x_1 corresponds to an expected negative difference of approximately 10% in y.
- When making predictions using the regression of the transformed variable, remember to transform back to the original scale to make your predictions more meaningful.

OTHER TRANSFORMATIONS

- While the natural logarithm transformation is the most common, there are several options.
- For example, logarithm transformations with other bases, taking squares, taking square roots, etc.
- Which one should you use?
- Well, it depends on what you are trying to fix.
- For linearity, for example, it is possible to need a logarithm transformation on the response variable but a square root transformation on the one of the predictors, to fix violations of linearity and normality.
- Overall, if you do not know the options to consider, you could try Box-Cox power transformations (to fix non-normality).
- We will not spend time on those in this course but I am more than happy to provide resources to anyone who is interested.
- First, see the boxcox function in R's MASS library.

What's next?

Move on to the readings for the next module!