Jared Yu
Module 8 Discussion

1. Consider a simple linear regression where x is the independent variable and y is the response variable. Suppose that the variance of the random error e is proportional to x. Share your ideas of possible transformations to stabilize the variance of the response variable y.

Ans:
Let the regression model be denoted as follows,

$$y = \beta_0 + \beta_1 x + e.$$

The idea is that the variance of the random error is proportional to x. Let this be denoted as

$$Var(e) \propto x.$$

Then this seems to violate the assumption of constant variance. In this scenario, the variance of the random error is actually quite related to the given x value. Therefore, the notion of constant variance throughout the random errors is being violated. A possible reason then is that the response y follows a distribution where the variance is functionally related to the mean. The textbook gives the example of the Poisson distribution, where the two are identical.

Although it is uncertain, the transformation $y' = \sqrt{y}$ is possibly suitable, which is what's used in the case of Poisson data. The reason is that if $\sigma^2 \propto E(y)$, then this transformation can be used. Ideally, we want $Var(e) = \sigma^2$, but here it is proportional instead to x. If instead the variance σ^2 is proportional to x, then that's similar to being proportional to $E(y)$, since $E(y) = \beta_0 + \beta_1 x$. It could be such that other transformations are more appropriate, where Table 5.1 in the textbook shows a variety of transformations that can be tested.

An alternative is to use a Box-Cox transformation, where we look at y^λ as a suitable transformation. This method would give us a range of λ values that we can test, where we can later plot $SS_{Res}(\lambda)$ versus λ, as seen in Figure 5.9 of the textbook. As the textbook states, we can check whether 1 is within some cutoff to determine whether the transformation is necessary.

It may also first be worthwhile to consider doing an initial lack-of-fit test to see how bad the violation is. There is the chance that doing some transformation or making any changes can do more harm than help in the case of modeling. So, this is just some consideration to see how the model can benefit from looking at the variance correcting techniques.

2. Discuss what pitfalls data transformation could lead to.

Ans:
The initial result of performing transformation means that the results of the model that incorporates transformation is no longer in the same unit of measurement. Therefore, it is possible to do an inverse transformation of the estimates back into their original units of measurement. However, there is the issue that the estimate $E(y)$ is no longer about the mean response, but rather about the median response.

To correct for this issue, instead a confidence interval (CI) instead can be constructed. By creating a CI for the inverse transformation, the results are not impacted by said transformation.

The downfall of this is that the intervals can vary in length so it is not guaranteed that the results will be ideal.

3. In a simple linear regression, if the response variable y is continuous, how likely will we see replicate values on y? If it is unlikely, can lack-of-fit testing be performed for checking the adequacy of the regression model?

Ans:

Given that the response is continuous, it depends on the type of data which allows for the replicate values at some level of x. Another factor is how large the sample size is. If the dataset is relatively small, then there can be a much smaller probability of there being any replicate values. However, if the dataset is large, then there is the chance of not only sampling data from the same x-level, but also having the situation where there are varying corresponding y-values.

Something else to consider is the type of distribution that the x-variable follows. If for example it follows some count-type of pattern, then there is a much larger chance for there to be repeated values of x at the same level, and therefore an increased possibility for there to be varying corresponding y-levels.

If such a situation is unlikely, then a lack-of-fit test can be problematic if not impossible. The reason is that the replicate observations are required in order to obtain a model-independent estimate of σ^2. For example, the test compares the sum of squares for the pure error and sum of squares due to the lack of fit. Then a hypothesis test can be conducted between these two sums of squares. Otherwise, if there are not sufficient replicate values (e.g., one replicate), then the test is much less useful. In such a case, perhaps it'd be helpful to gather a larger sample size so that the test can be done.

4. Discuss what impact "nonconstant variance" have on regression analysis.

Ans:

An interesting note about nonconstant (error) variance is that the least-squares estimators will still have the property of being unbiased. However, it will no longer be considered to have the minimum-variance property from being the best linear unbiased estimator (BLUE). This is important, since minimum variance can be desirable for a model. Therefore, doing some type of transformation can improve the resulting model so that it not only has unbiased coefficients, but that the standard errors of them can be reduced.

By doing some sort of analysis on the resulting model after testing different types of transformations, then it can be such that the researchers will settle on an alternative model where the chosen model is different from ordinary least-squares (OLS) model. However, as noted previously, doing such transformations are not perfect. Furthermore, there are different options and so a choice must be made amongst possible alternative models.

If it can be seen that the transformed model improves the residual analysis plots, then the transformation can be used. There are also important considerations, such as those relevant to the domain knowledge regarding the data itself. It becomes apparent then that there is the possibility that a transformation can improve the diagnostics, but they violate some basic understanding of the data and what is possible. So, it is important that such considerations can be accounted for if possible. There are also many considerations, such as whether the response follows a Poisson or Bernoulli distribution. These can help to direct an appropriate approach for transforming the data.

In the case of Box-Cox, when looking at the resulting response variable that has been transformed to some power, there is also the situation where it becomes apparent that the transformation is quite subtle compared to the original data. In such a case, it’s possible that transformation can be avoided altogether.

Given that transformations are done, it is important to make note of these in the report. This is similar to what happens in outlier removal. In such a case, it’s possible to show the different models and indicate that one is the OLS model and the other is transformed due to the assumption violation. This helps to give a better idea of what the results are and how they can be interpreted and made use of.

A last possibility is to an alternate model such as generalized or weighted least squares. These will give resulting models that have better properties than the OLS model. So, given that the data has some issues with constant variance. Then, it can be that multiple models are presented in the end, and from them each can be interpreted for their own pros and cons. Weighted least squares is used when the random errors have uncorrelated but unequal variance.