Jared Yu
Module 8 Discussion

In the first question, you mention a list of possible transformations to use in the case of the variance being proportional to the x variable. I noticed these also in the table within the textbook. Something interesting that you mention is to do nothing, which I think is also an important point. It is not always the case that the transformation may help the modeling, despite there being some problems with the variance. Something else you mention is the use of various types of distributions. I am not sure what you mean by this, since when I read in the textbook, my understanding was that it was referring to the possibility of the data itself following such a distribution. Therefore, if the data followed such a distribution, then certain types of transformations would be suitable. For example, the square root goes well with the Poisson distribution. Perhaps I misunderstood the textbook though.

I mentioned the same thing in the second part, where if we transform the data, then we have to do an inverse transformation afterwards to get it back into the same unit of measurement. We can also use something like a confidence interval. I also noted that this inverse transformation is not without flaws, and so this is definitely a problem that is factored into the process when we use transformations. Your second point about prior information I think is regarding domain knowledge. I don't think I mentioned that exactly here, but I do agree with this idea. If for example we know something about the data, then we would know what transformation we shouldn't use. However, not having this information means we wouldn't know not to do it.

I think you make a good point about replicate values in the case of continuous variables. Since the variable goes to a fraction, there are many more values and thus less of a chance of replicates for some level of x. We would perhaps need a large sample in order for replicates to start appearing. I think a possibility is to round the data off to maybe one or two decimal places so that replicates appear, however this would cause us to lose information and so I'm not sure it makes sense to do this either. I am also wondering how many samples are required for lack-of-fit to be meaningful. For example, if we have only one or two replicates then it maybe doesn't make so much sense to try and interpret the results.

In the last section, you mention that the regression coefficients will have larger standard errors. I think you are referring to the fact that they no longer have the minimum-variance property and therefore are no longer BLUE like the OLS estimators. In such a case, I think there are possible options such as using generalized least squares or weighted least squares.