Jared Yu
Module 7 Discussion

In the first part, I mentioned the same residuals as you. I like how you compare the standardized residual to a Z-score. With the studentized residuals, I think you make an interesting point that it does better in cases where there is a small sample size. In the case of the PRESS residual, you make an interesting point regarding that it's useful for finding single outliers. I wonder then what would be a good method to identify clusters of outliers. I've seen these outlier plots in regression textbooks, and they seem to show these small textbook datasets. In my experience with today's "big data," the probability is more likely that there will be clusters of outliers in addition to individual outliers. I think identifying these would be quite interesting. You make the interesting point to that the R-student residuals are more sensitive than the PRESS residuals.

I agree with your explanation of the residuals. It makes sense to me that the true values are generally unknown, since they're like population parameters that can't really be measured in any meaningful way. The residuals then I think you are saying is our estimate of the errors by using a fitted model and some sample data. This sort of view of it goes with what I think also.

The last part about our model being unhelpful or misleading I think is important. I think these days there are so many machine learning tools out there so it's easy to run a program and fit a model without thinking too hard about what we're doing. If we make the mistake of incorrectly applying a model when the basic assumptions are not being met, I think that makes room for the same problems you mention. In the worst case we are misled by our models and thus we make faulty decisions that could've easily been avoided in the case that we were more careful.