Jared Yu
Module 14 Discussion

Given a dataset for a response variable and a single regressor, we can fit the data using a regression model on the original values of the response variable or using a regression model on the ranks applied to the response variable. Discuss the advantages and the disadvantages of the two modeling approaches. Discuss the conditions required for either approach, respectively.

Ans:

The question says that there is a response variable which will be denoted y, and a single regressor which will be denoted x. My understanding is that the question is saying that there are potentially two approaches to modeling y. In the first approach, we can treat it as a normal regression type of problem, where we are trying to model some numeric variable (continuous or discrete) like with linear regression. In the second approach, we are applying some sort of ranking to the response. So, for example, we can use some discretization or binning method to fit the response variable into certain categories.

For instance, if the range of a response is between 0-100, they can be ranked 1-10 depending whether they fit into the ranges of 0-10, 11-20, …, 90-100 respectively. This is an interesting possibility, since it could be more practical from an analytical or predictive point of view to estimate this ranking rather than the specific numeric value. So, this will definitely be a choice that is data-dependent and problem-dependent, since it will not always be the best choice for a dataset or a problem statement.

With the first approach, we have the benefits of it being the traditional regression type of problem, where we're generally quite used to understanding how well or not well a dataset or model fit the linear regression paradigm. For instance, analyzing the residuals, we are perhaps more experienced at understanding the deviations from what is expected. Also, using this approach would still force us to check the typical assumptions of linear regression. Therefore, it's not a given that this first approach will always be easier or work better.

In the second approach, we must use some other type of regression other than the normal linear regression. I can think of using logistic regression since the ranking essentially creates categories like a classification problem. On the other hand, if the categories are ranked in some meaningful order, then perhaps some adaptations would be required to account for this nuance. I think the biggest difference for this logistic regression approach is the types of statistical inference that follows. These are of course different from what's done in traditional linear regression, and so there is possibly fewer resources (less research in the area). For example, we will need to try things like a goodness of fit test. In such a case, we would also need to account for the assumptions of logistic regression, like whether the classes are balanced or not.