Linear Models Assignment

Q0:

1. What makes a model "linear"? "Linear" in what?

A model is linear if there is a clear relationship between variables and that there is a constant value relating these variables. The model is linear if the data fits the equation y=mx+b, with m being the slope of the line and b being the y-intercept. There is usually an independent variable (x) and a dependent variable (y). A linear model is technically any model that assumes linearity in the system.

2. How do you interpret the coefficient for a dummy/one-hot-encoded variable? (This is a trick question, and the trick involves how you handle the intercept of the model.)

The intercept in a linear regression model is the expected value of the dependent variable for the category where all of the dummy variables equal 0 (the reference category). The coefficient for the dummy/one-hot-encoded variable is the difference between categories where the dependent variable equals the category coded as 1 compared to the intercept. So, the coefficient is interpreted as the difference (higher, lower, the same, etc.) in the mean value of the dependent variable between the group coded as 1 versus 0. An example of this is the dependent variable is the drug effect, 0 = control group, and 1 = treatment group.

3. Can linear regression be used for classification? Explain why, or why not.

Linear regression is not a good tool for classification. Linear regression takes in continuous values and fits this type of data, whereas classification works with discrete values (like binary outcomes). Additionally, classification often works with probability values which linear regression is not an effective tool for.

4. What are signs that your linear model is over-fitting?

Signs that your linear model is overfitting have low error rates and a high variance. Additionally, if the training data set has a low error rate but the test dataset has a high error rate, this could indicate overfitting. This can happen when the model has too many parameters compared to the number of observations it is taking in which doesn't allow the model to generalize data well and perform well on the test data.

5. Clearly explain multi-colinearity using the two-stage least squares technique.

Multi-colinearity occurs when at least two or more independent variables are highly correlated with each other. This can lead to it being difficult to estimate the relationship/correlation between each individual independent variable and the dependent variables. Two-stage least squares technique is a way to clarify the relationships between each independent variable and the dependent variable to yield more reliable coefficients. This method uses the instrumental variable that effects the problematic independent variable but that isn't related to the outcome you are looking at. The first stage involves running a linear regression on the instrumental variables to predict the independent variable giving you issues. Then in the next stage, you take these values and use them for the main linear regression.

6. What are two ways to incorporate nonlinear relationships between your target/response/dependent/outcome variable y and your features/control/response/independent variables x?

   If the raw data plotted does not yield a linear relationship, you can do a polynomial regression that could better fit this data. A different n value (power of polynomial) can be used depending on the shape of the data (ex. quadratic, cubic, etc.). Additionally, different mathematical functions can be applied to the data. For example, if you want to linearize data that is acting exponentially, you can use the natural logarithmic function (ln). You could also use the square root function for other purposes.

7. What is the interpretation of the intercept? A slope coefficient for a variable? The coefficient for a dummy/one-hot-encoded variable?

The intercept represents the value of the dependent variable (y) when the independent variable (x) equals 0. The slope coefficient for a variable is the rise over run, or the rate of change from one (x,y) pair to a different (x,y) pair in the data (or the average of that across all data points). The coefficient for a dummy/one-hot-encoded variable is the difference in the dependent variable between the dummy and the reference group.