Assignment 5

Regression & Scatterplots

Lacey Hartigan

due 3/9/21

For this assignment, we're going to be using the ELS_training and ELS_testing data introduced in LMS module 5. We will also use code we learned in LMS module 6.

Research topic: We are interested in predicting reading scores as a function of SES (and other covariates) in the ELS dataset.

1. First, identify whether SES and reading scores are correlated (using the *training* data).

2. Plot a scatterplot of SES and reading scores (make sure your IV is on your x-axis, and your DV is on your y-axis). Include a regression line on your plot. Note: it may help to "simplify" your data before plotting (so that we don't have a separate dot for all 8000+ observations).

3. Assuming it makes sense to do so (look at your correlation stats AND your scatterplot), run a simple linear regression predicting reading scores as a function of SES (use the *training* data).

4. Report the RMSE from a validation of your model using the els *testing* data.

5. Revisit your training data. What other covariate (aka, variable) might you add to better predict reading scores? Choose a variable (or variables) and create a correlation matrix including your independent variables and your dependent variable (use the *training* data). Note: don't exceed 3-4 independent variables total. And make sure they all make sense to include!

6. Run an updated regression model using your additional var(s).

7. Report the RMSE from a validation of this updated model using the *testing* data. Did your model improve? If yes, by how much?

EXTRA PRACTICE: Play around with creating some scatterplots of your multiple linear regression model. Choose a display (or displays) that you think would help you explain the results of your model to a non-stats audience.