Goal 6 – Fitting Models

General Instructions

1. Your group will submit a pdf with no more than 3 pages, excel/csv files with ONLY the 4 - 6 variables needed for this assignment., and your RScript (for this make sure that your dataset/s used is the selected dataset you will submit)
    a. Use the select() function to choose these variables, i.e. dataset_new_name=old_dataset_name%>%select(variable1,variable2,…)
2. Your group may submit a knitted document from R but it must be neat and only have necessary information. For example, I do not need to see the packages you loaded and the associated information. If you choose to do this, you will want to knit to a Word document and remove ALL unnecessary lines.
3. The format that is required looks like:

Hypothesis:

Predictor/s: Type of variable:

Response: Type of variable:

Type of model used: (either linear or logistic and simple or multiple)

Answers to questions:

Note: You may have 1 set of these or multiple sets of these based on which route your group decides.

4. If a number is needed to answer any of the questions, e.g. R-squared, slope, P(>|t|), RSE, etc. be sure to include them in your answers. If it requires a plot (for example – is the relationship linear? Or is there a relationship?), please provide a plot.

Specific Instructions

You should have 3 hypotheses. Choose **ONLY ONE** of the following routes:

1. Fit a Simple Linear Regression OR Simple Logistic Regression for each hypothesis
    a. If you have a hypothesis that requires more than one simple linear regressions/ simple logistic regressions, chose ONLY one for each hypothesis
2. If your hypotheses have the same response variable, you are allowed to use a multiple linear regression OR multiple logistic regression. In doing this, you will have one model with 3 predictors and 1 response

3. If 2 of your hypotheses have the same response variable, you can fit one multiple linear regression that combines both hypotheses (i.e. 2 predictors and 1 response). For the last hypothesis, fit one simple linear or simple logistic regression.

The above are the requirements for GOAL 6. You can perform more than what is required for your group's purpose of understanding the data (and being able to talk about it more for your final presentation – Goal 7), **but only submit what is required.**

Route 1

Questions to answer for EACH regression:

1. Is there a relationship between the predictor and the response?
2. Is the relationship linear (if it is a linear regression)?
3. How strong is the relationship between the predictor and the response?
4. Does the predictor contribute to the response?
5. What is the effect of the predictor on the response?
6. How accurately can we predict the response using the predictor with results from lm()?
7. Use one Cross Validation technique (LOOCV or K-fold) for a better method to determine how accurately we can predict.

Questions to answer about all the regressions?

8. Which predictor has a stronger association/relationship?
9. Which predictor has the biggest effect on the response?
10. Do all the predictors contribute to the response?
11. Which predictor best predicts the response using the results from lm() and also from Cross Validation?
12. **Bonus:** Describe the drawbacks of using simple linear regressions instead of multiple linear regression if you are predicting the same response (can be found in the notes)

Total – 25 questions + Bonus

Route 2

Questions to answer for the Multiple Linear/Multiple Logistic Regression

1. Is there a relationship between the variables and the response?
2. How strong is the relationship between the predictors and the response?
3. Do the predictors contribute to the response?
4. What is the effect of each predictor on the response?
5. How accurately can we predict the response using the predictor?
6. Use a model selection procedure to select the best model

    a. forward selection or backward selection using AIC
    b. Cross Validation technique + forward or backward – LOOCV or K-fold to use RSE/RMSE
7. Is the selected model the same as the model with all the variables?
8. **Bonus** – How is the selected model different from the original model?

Total: 7 Questions + Bonus

<u>Route 3</u>

- For the multiple linear/logistic regression, follow **Route 2**
- For the simple linear/logistic regression, follow **Route 1 questions 1-7**

Total: 14 Questions + choose 1 Bonus from the 2 routes