

Assignment 1: Linear Regression

Due date: February 8, 11:59am

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_1.pdf` and `Obama_Barack_1.R`. Your submissions must be based on your own original work. Late submissions will be penalized at 10% per hour.

1. In this problem you will need to analyze the `EggProduction` dataset. The data includes daily information about egg production (number of eggs laid per chicken), feed (amount of feed per chicken) and daily temperature (in degrees Celsius). Several of the questions ask you to interpret the results, you might find it useful to go through the entire analysis before coming back to these questions.
 - (a) Download `EggProudction.csv` to your computer. You can use `setwd()` to set the current working directory. Load the data and print a summary of the variables.
 - (b) Run a regression of eggs on feed and interpret the result. Does the result make sense to you?
 - (c) Now include temperature in the model. Run a regression of eggs on feed and temperature and interpret the result. Does this result make sense to you?
 - (d) Plot feed against temperature to understand how the temperature affects the amount of feed the chickens receive. From the graph you should be able to see that the amount of feed changes when temperature takes certain values. Create a new binary/discrete/categorical variable from the temperature column that captures this phenomenon. Print a summary of the new variable.
 - (e) Regress eggs on feed, temperature, and the new variable you created. Interpret the results.
 - (f) Which of the regression models is best for prediction? Why?
 - (g) For your best model, what is a 95% confidence interval for the regression coefficients. Interpret the results.
 - (h) For your best model, what is a 99% prediction interval if the feed was 26 and the temperature was -2. Interpret the results.
 - (i) Provide an intuitive explanation for what changed between the first two regressions and the last regression.

2. Consider the `NomisB_e-Car_Data.csv` dataset from Session 2. We would like to run an analysis on a different subset of the data. Specifically we want to look at quotes for new cars, 60 month term, FICO between 675 and 725, and amount between \$30,000 and \$40,000.
 - (a) Find a linear regression model to predict the probability of the quote being accepted as a function of the APR offered. What is the model?
 - (b) Is this model a good fit? How does it compare to logistic regression?
 - (c) Try adding the competitor rate to your linear model. Does the new model make sense? Does the fit improve?
 - (d) For the previous model in part (c), compute the RSE of the predictions for customers in Partner Bin 1.