# Homework 1

Due Friday, 2/9 by 5pm

**Turn in to UVA Collab as a PDF**

Instructions: Complete the following homework. Your answers should be turned into collab on a single PDF document. Homework submitted as anything but a PDF will not be accepted. Problem 2 requires you to turn in separate R code, in addition to your writeup. Visualization is highly encouraged wherever it is useful, but don't go overboard. Any visualizations should be neat, legible, and add value to your writeup. If it isn't any of those three, leave it out! You may work together on this assignment, but the writeup should be your own.

1. Breaking down error into its component parts. (For this problem, you may not consult the internet or other textbooks, as the answer to this problem is easy to find online. It is, however, very important to have this under your fingertips as a statistician/data scientist!)
    a. Show that $Var(X) = E(X^2) - E(X)^2$
    b. Let $f^*(x)$ be the bayes predictor for $y$, so we can write $y = f^*(x) + \epsilon$. Let $\bar{f}(x)$ be the expected fit for functions in some restricted class of predictors, and let $\hat{f}(x)$ be the observed best fit for a given dataset in the restricted class. Using a) Show that
    $$E\left[\left(Y - \hat{f}(x)\right)^2\right] = E\left[\left(Y - f^*(x)\right)^2\right] + \left(f^*(x) - \bar{f}(x)\right)^2 + E\left[\left(\hat{f}(x) - \bar{f}(x)\right)^2\right]$$
2. R simulation validation of error decomposition (Please turn in your R code in a separate .R file for this question)
    In this exercise, you will simulate data from the known function (and minimizer of squared error), $y = 0.2\,x + \sin(x) + \epsilon$, with $\epsilon \sim N(0,1)$; simulate 1000 datasets and compute the line that best fits the data in the least squares sense; and record relevant values which will help you calculate the three sources of error for a freshly sampled data point at x=7. Finally, combining your values across the 1000 trials and calculating the relevant error terms, compare your summarization to the theoretical formula in 1b. The steps in your code are:
    a. Write a function gety <- function(x) { # return y value} that computes and returns a y value from a given x value (note you can write the function as taking a single x value, and if you pass that function a vector, R will do the reasonable thing with it. Except you have to be careful with the noise term).
    b. Run the following 1000 times:
        i. Simulate data from x values generated by seq(-10, 10, by=0.1) (never use a function in your homework without knowing exactly what it does and what happens if you change inputs!)

ii. fit the line of best fit, f.hat, using the function lm() (note you first have to create a dataframe using your y and x values)

iii. simulate a new data point $y^{new}$ at x=7. Calculate and record $\left(y^{new} - f.hat(x = 7)\right)^2$, and $f.hat(x = 7)$

c. Using calculate the average prediction at x=7, f.bar, using the 1000 f.hat predictions you obtained.

d. Using the 1000 f.hat predictions and the average prediction, f.bar, calculate the variance of the estimated function at x=7.

e. Calculate the bias

f. Finally, compare to the theoretical quantities.

g. What would happen if we only saw half the x values to find the line of best fit?

3. Data exercise (R code not needed for this problem)

In this problem you will be applying your regression skills to a dataset called lpga, which records the performance of players in the LPGA in the year 2008. The variables are:

i. Golfer: name of the contestant

ii. Driver_Distance: average length of a driver shot

iii. Fairway_Pct: percentage of times the tee-shot lands in the fairway

iv. Greens_in_Reg: percentage of greens hit in the regulation number of shots

v. Avg_putts_per_round: Average putts per round

vi. Sand_per_round: Average number of sand traps hit per round

vii. Sand_saves: percent sand saves

viii. Tot_winnings_per_round: Average amount of winnings per round

ix. Log_tot_win_per_round: The previous number, logged

x. Tot_rounds: total number of rounds played

xi. ID: player specific id number

b. Ignoring the variables “Golfer” and “ID”, use the pairs() function to compare the remaining variables (don’t show this in your writeup). Describe any interesting relationships.

c. In your analyses going forward, you should use either Tot_winnings_per_round or Log_tot_win_per_round

i. Which do you choose, and why?

d. Run a regression of the variable you chose from c. on the rest of the variables (excluding the variable you did not choose!). Look at the summary of the regression and discuss.

e. Compute predictions from your model on the data used to train the model. Compare your predicted values to the true values. Discuss.