# DS 303 Homework 1
# Due: Aug. 28, 2023 on Canvas by 11:59 pm (CT)

**Instructions:** Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

**Code should be submitted with your homework *as a separate file* (for example, a .R file, text file, Word file, or .Rmd are all acceptable).** You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

## Problem 1: Bias-variance decomposition

a. Provide a sketch of typical (squared) bias, variance, expected test MSE, training MSE, and the irreducible error curves on a single plot, as we go from less flexible statistical learning methods towards more flexible methods. The $x$-axis should represent the amount of flexibility in the method, and the $y$-axis should represent the values for each curve. There should be 5 curves. Make sure to label each one.

b. Define in plain language (so that a non-data scientist can understand) what the quantities expected test MSE, training MSE, bias, variance and irreducible error mean.

c. Explain why each of the five curves has the shape displayed in part (a).

d. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for supervised learning? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

e. I collect a data set of ($n = 100$ observations) containing a single predictor and a quantitative response $Y$. I fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 + \beta_2 X^2 + \beta_3 X^3 + \epsilon$. Suppose that the true relationship between $X$ and $Y$ is linear. Consider the training MSE for the linear regression and also the training MSE for the cubic regression. Would we expect one to be lower than other, or is there not enough information to tell? Justify your answer.

f. Answer (e) using test MSE instead of training MSE.

## Problem 2: Interpreting MLR

Suppose we have a data set with five predictors, $X_1$ = GPA, $X_2$ = IQ, and $X_3$ = Level (1 for College and 0 for High School). The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit a multiple linear regression model on our data set and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$ and $\hat{\beta}_3 = 35$.

a. Which answer is correct, and why?

   i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

   ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

   iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

   iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

b. Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

c. True or false: Since the coefficient of IQ is very small, the effect of IQ effect on salary is not very important. Justify your answer.

## Problem 3: Multiple linear regression

For this problem, we will use the `Boston` data set which is part of the `ISLR2` package. To access the data set, install the `ISLR2` package and load it into your `R` session:

```
install.packages("ISLR2") #you only need to do this one time.
library(ISLR2) #you will need to do this every time you open a new R session.
```

To get a snapshot of the data, run `head(Boston)`. To find out more about the data set, we can type `?Boston`.

We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

a. How many rows ($n$) are in the data set? How many variables are in the data set? What does the variable `lstat` represent?

b. Obtain the average per capita crime rate across all suburbs in the data set. Report that here.

c. Obtain the average crime rate only for those suburbs who are not near the Charles river (`chas ==0`) and those suburbs who are near the Charles river (`chas ==1`). Report both values here. Is it safer to be near or away from the Charles river?

d. Do any of the suburbs of Boston appear to have particularly high crime rates? Define what a 'high' crime rate is and provide some summary statistics on the crime rate.

e. Are any of the other predictors in the data set associated with per capita crime rate? Use your exploratory data analysis skills to uncover insights. Describe your findings.

f. Fit a simple linear regression model with `crim` as the response and `lstat` as the predictor. Describe your results. What are the estimated coefficients from this model? Report them here.
Note: a simple linear regression is just a regression model with a single predictor.

g. Explain in words how we fit a model. In other words, what approach did we use to obtain the estimated regression coefficients from this model? Is this approach reasonable?

h. Repeat part (f) for *each predictor in the dataset*. That means for each predictor, fit a simple linear regression model to predict the response. Describe your results and organize them in a table. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

i. Fit a multiple regression model to predict the response using all of the predictors. Summarize your results neatly in a table.

j. How do your results from (h) compare to your results from (i)? Create a plot comparing the simple linear regression coefficients from (h) to the multiple regression coefficients from (i). Describe what you observe. How does this provide evidence that using many simple linear regression models is not sufficient compared to a multiple linear regression model?

k. First `set.seed(1)` to ensure we all get the same values. Then, split the `Boston` data set into a training set and test set. On the training set, fit a multiple linear regression model to predict the response using all of the predictors. Report the training MSE and test MSE you obtain from this model.

l. On the training set you created in part (k), fit a multiple linear regression model to predict the response using only the predictors `zn, indux, nox, dis, rad, ptratio, medv`. Report the training MSE and test MSE you obtain from this model. How do they compare to your results in part (k)?

m. Are these results in part (l) surprising or what you expected? Explain.