# DS 3010 Homework 3

**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 in your solutions.** (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: Statistical Inference

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

```
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(Carseats)`. To find out more about the data set, we can type `?Carseats`.

(a) This question has two parts:

(i) Fit a multiple linear regression model to predict car seat unit sales (in thousands) using all available predictors **except** `ShelveLoc`. Use the entire dataset (do not split the data into training and test sets). Summarize the least squares estimates and their corresponding standard errors in a table.

(ii) Select *one* regression coefficient from the fitted model and test whether it is equal to zero at the significance level $\alpha = 0.05$. Clearly state the null and alternative hypotheses, the test statistic, its null distribution, the $p$-value, and your conclusion.

(b) **(Practice)** What additional assumption did you need to make in part (a) to carry out the hypothesis test?

(c) **(Practice)**Report an estimate for $\sigma^2$. What does this value mean in plain language?

(d) Carefully interpret the estimated regression coefficient associated with `Advertising`. Double check your lecture notes for precise language.

(e) Obtain the RSS for the full model (from part (a)) and the RSS for reduced model (with no predictors). Report them both here. *Hint: The reduced model contains only an intercept. What value should the intercept take ?*

(f) Carry out the F-test at $\alpha = 0.05$. Write out the null/alternative hypothesis, test statistic, null distribution, $p$-value, and conclusion.

(g) Use the fitted model to estimate $f(X)$ under the following predictor values:

- competitor price = average competitor price,
- community income = median income,
- advertising = 15,
- population = 500,
- price = 50,
- average age = 30,
- education level = 10,
- urban location in the U.S.

Report your estimate of $f(X)$ and quantify the uncertainty in this estimate by constructing the appropriate interval.

(h) Same setting as part (g). What is your prediction for $Y$ given these predictors? Quantify the uncertainty surrounding our prediction for $Y$ (given these predictors) by reporting the appropriate interval. *Hint: Think carefully about the difference between estimating $f(X)$ and predicting an individual response $Y$. Which source(s) of uncertainty should be reflected in the interval?*

(i) Prove: in the context of multiple linear regression, $f(X) = E(Y)$ for fixed values of $X$. Therefore, the interval in part (g) can be interpreted as quantifying the uncertainty surrounding our estimate for the expected value of $Y$ ($E(Y)$) for a fixed value of $X$.
Note: proving means to use explicit math notation and logic. It is not sufficient to just write out the idea in words. Reference your lecture notes.

(j) Obtain the prediction for $Y$ using all the same settings as (g), but set the price for car seats at each site to be 450. What is your prediction for $Y$? Does this value make sense? Discuss how this reveals a limitation of our model.

(k) Is the prediction for $Y$ ($\hat{Y}$) an unbiased estimator of $Y$? Justify using statistical concepts.

## Problem 2: The Challenge of Multiple Testing

Think back to our in-class activity related to multiple testing (see `R` script `multiple_testing.R` if you need a refresher). We illustrated in that code that if that set $\alpha = 0.05$, we would expect roughly 10 predictors to be significant just by chance, and we know some of those significant predictors are false positives. This illustrates the multiple testing problem: when testing a large number of null hypothesis, we are bound to get some very small p-values just by chance. If we make a decision about whether to reject each $H_0$, without accounting for the fact that we have performed many tests, we may end up making a large number of type 1 errors (also referred to as false positives or false discoveries).

(a) In general if we wish to test $m$ null hypothesis and we simply reject all null hypothesis for which the corresponding p-value falls below $\alpha$, how many type 1 errors (false positives) should we expect to make?

(b) Repeat the simulation from our in-class activity but now let the number of tests being carried out to vary. This means instead of simulating 150 predictors you can generate data such that $p$ equals 200, 400, 500, 600, and 800. How does the number of false positives change as the number of tests changes? Create a plot where the y-axis is the number of false positives and the x-axis is the number of tests carried out.

## Problem 3: More Simulations

Suppose we know that the true underlying population regression line is as follows :

$$Y_i = 2 + 3 \times X_{1i} + 5 \times X_{2i} + \epsilon_i \quad (i = 1, \ldots, n), \quad \epsilon_i \sim \mathcal{N}(0, 2^2).$$

(a) What are the true values for $\beta_0$, $\beta_1$, and $\beta_2$?

(b) **(Practice)**Generate 100 observations $Y_i$ using the true population regression line. You may use the following code to generate $x_1$ and $x_2$:

```
X1 = seq(0,10,length.out =100) #generates 100 equally spaced values from 0 to 10.
X2 = runif(100) #generates 100 uniform values.
```

(c) **(Practice)**Design a simple simulation to show that $\hat{\sigma}^2$ is an unbiased estimator of $\sigma^2$. Note that $\sigma^2$ **does not** equal 1 in this setup.

(d) **(Practice)**Plot a histogram of the distribution of the $\hat{\sigma}^2$'s you generated. Add a vertical line to the plot showing $\sigma^2$.

(e) **(Practice)**Explain why an estimate of $\sigma^2$ can be so important in the context of multiple linear regression. If you do not have a good estimate of $\hat{\sigma}^2$, what aspect of our pipeline breaks down?

## Problem 4: Consulting

Suppose you are offering data science advice to a team of collaborators. This involves analyzing a dataset and fitting a multiple linear regression model to this dataset.

(a) Your collaborator asks you to carry out hypothesis testing for a regression coefficient $\beta_j$. He sees that you have set the significance level to be $\alpha = 0.05$. He wants to know what this $\alpha = 0.05$ means in the context of hypothesis testing. Explain in plain language.

(b) Suppose that the the $p$-value for the regression coefficient $\beta_j$ is 0.0647. It is not significant at $\alpha = 0.05$ so your collaborator claims that the associated predictor ($X_j$) is not meaningful and suggests fitting a model without this predictor. Do you agree or disagree with their claim? Carefully justify your answer.

(c) **(Practice)**There are future plans to collect additional predictors to better understand factors affecting the response of interest. Suppose there will be a total of 12 predictors collected in the future. The scientist wants to determine whether or not at least one of these predictors is useful in predicting $Y$. He proposes fitting a model with all 12 predictors and then carrying out 12 individual $t$-test for each regression coefficient. If it at least one result is significant, he can conclude at least one of the predictors is useful in predicting $Y$. Explain in plain language why this might be a bad idea. What is the probability of seeing at least one significant result by chance? Use $\alpha = 0.1$.

End of assignment.