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) Fit a multiple linear regression model to predict carseat unit sales (in thousands) using all other variables **except** `ShelveLoc` as your predictors. Use the entire dataset (do not split it into a training and test set). Summarize your least-square estimates and their standard errors in a table. Choose one regression coefficient from the model and test whether it is zero or not at $\alpha = 0.05$. Write out the null/alternative hypothesis, test statistic, null distribution, p-value, and conclusion.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	7.8243876	1.1298792	6.925	1.81×10^{-11}
CompPrice	0.0942545	0.0078637	11.986	$< 2 \times 10^{-16}$
Income	0.0130501	0.0034908	3.738	0.000213
Advertising	0.1369399	0.0210394	6.509	2.33×10^{-10}
Population	-0.0002007	0.0007010	-0.286	0.774780
Price	-0.0924395	0.0050621	-18.261	$< 2 \times 10^{-16}$
Age	-0.0447919	0.0060226	-7.437	6.59×10^{-13}
Education	-0.0423034	0.0373738	-1.132	0.258371
UrbanYes	-0.1559036	0.2133519	-0.731	0.465380
USYes	-0.1062926	0.2832192	-0.375	0.707640

I chose `Income` but students' solutions may vary.

- $H_0 : \beta_{income} = 0$ *versus* $H_1 : \beta_{income} \neq 0$.
- *Test-statistic: 3.738*
- *Null distribution: t distribution with 390 degrees of freedom.*
- *P-value:* 0.000213
- *Conclusion: We have evidence at the* $\alpha = 0.05$ *significance level that the regression coefficient associated with* `Income` *is significantly different from 0.*

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

We had to assume that our errors ($\epsilon_i, i = 1, \ldots, n$) are normally distributed. Alternatively, we could assume our sample size is large enough for Central Limit Theorem to kick in.

(c) Report an estimate for σ^2. What does this value mean in plain language?

Our estimate for σ^2 is $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n - (p + 1)) = 3.73$. This value represents our estimate for the average variability in the response Y (carseat sales). We can think of this as the average amount that the response will deviate from the true population regression line.

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

For a thousand dollar increase in `Advertising`*, average carseat sales will increase by 0.13 holding all other predictors constant.*

(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 does this model predict for every observation?*

RSS for the full model is 1456.031 and RSS for the reduced model is 3182.275.

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

- $H_0 : \beta_1 = \beta_2 = \ldots \beta_9 = 0$ *versus* H_1*: at least one is non-zero*
- *Test-statistic: 51.38*
- *Null distribution: F distribution with df1* $= 9$ *and df2* $= 390$
- *P-value:* $< 2.2 \times 10^{-16}$
- *Conclusion: We have evidence at the* $\alpha = 0.05$ *significance level that at least one of our regression coefficients is different from zero.*

(g) Use the model to estimate $f(X)$ when the price charged by competitor is average (you'll need to find what the average competitor price is), median community income level, advertising is 15, population is 500, price for car seats at each site is 50, average age of local population is 30, education level is 10, and the store is in an urban location within the US. What is your estimate for $f(X)$ given these predictor value? Quantify the uncertainty surrounding our estimate for $f(X)$ by reporting the appropriate interval.

Estimate for $f(X)$ is 15.80707.

$$\begin{aligned}\hat{Y} = 7.82 &+ 0.094 \times (124.975) + 0.013 \times (69) + 0.136 \times (15) \\ &- 0.0002 \times (500) - 0.092 \times (50) - 0.044 \times (30) \\ &- 0.042 \times (10) - 0.155 \times (1) - 0.106 \times (1)\end{aligned}$$

A 95% confidence interval for $f(X)$ is (14.91953, 16.69461).

(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.

Our prediction for Y given these predictor values is 15.80707. A 95% prediction interval (11.90593, 19.70821).

(i) Prove that, 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.

Our model setup is $Y = f(X) + \epsilon$. We assume that $E(\epsilon) = 0$. Therefore it follows immediately that $E(Y) = f(X)$. As such, we can interpret our 'systematic signal' $f(X)$ as the average value of Y for given values of X.

(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.

The prediction for Y is -19.41. This values does not make sense - we cannot have negative carseat sales. This tells us that our model is only good at explaining Y for values of X that has already seen before. In our dataset, the range for price only goes to 191:

```
summary(Carseats$Price)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.0   100.0   117.0   115.8   131.0   191.0
```

If we go outside of the range of X, our model will run into problems. Students may refer to this as extrapolation.

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

No. If $\hat{Y}$ was an unbiased estimator of Y, then we would have $E(\hat{Y}) = Y$. But we can see this is not true. We know that

$$\begin{aligned}E(\hat{Y}) &= E(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) \\ &= \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = f(X) \\ &= E(Y) \\ &\neq Y.\end{aligned}$$

Problem 2: The Challenge of Multiple Testing

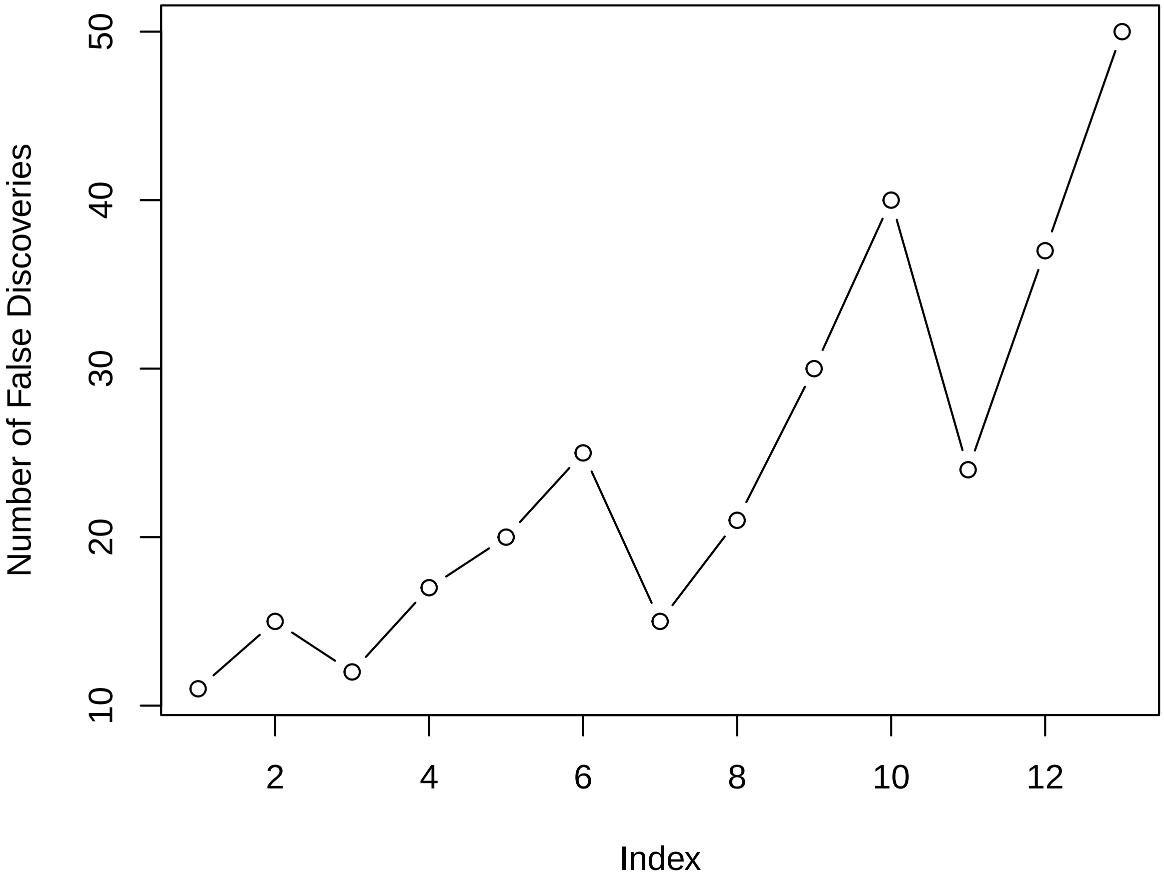
Think back to our in-class activity related to multiple testing (see `R` script `multiple_testing.R` if you need a refresher). We illustrated in that code that if that set $\alpha = 0.05$, we would expect roughly 10 predictors to be significant just by chance, and we know some of those significant predictors 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 α, how many type 1 errors (false positives) should we expect to make?

$m \times \alpha$.

(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.

Students did not necessarily have to use p = 200, 400, 500, 600, and 800. But they should have at least tried different values of p and plotted the number of false positives. Students could have also 5 predictors associated with the response Y or they could have none associated.



```
vec = seq(200,800, by = 50)
result = rep(NA,length(vec))
for(b in 1:length(vec)){
  p = vec[b]
  x = matrix(NA,1000,p)

  for(i in 1:p){
    x[,i] = rnorm(1000)
  }
  y = rnorm(1000)

  data = as.data.frame(cbind(y,x))

  fit = lm(y~.,data=data)
  p_values = summary(fit)$coefficients[,4]
  result[b] = length(which(p_values<0.05))
}

plot(result,type='b',ylab='Number of False Discoveries')
```

Problem 3: More Simulations

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

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

(a) What are the true values for β_0, β_1, and β_2?

$\beta_0 = 2$, $\beta_1 = 3$, and $\beta_2 = 5$.

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

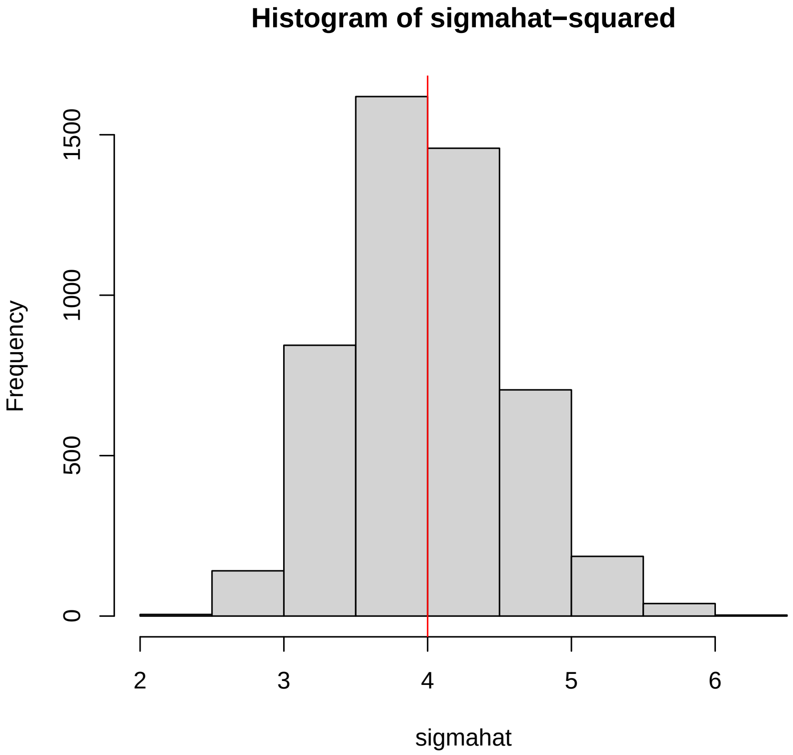
```
 X1 = seq(0,10,length.out =100) #generates 100 equally spaced values from 0 to 10.
 X2 = runif(100) #generates 100 uniform values.
```

```
 beta_0 = 2
beta_1 = 3
beta_2 = 5
X1 = seq(0,10,length.out =100)
X2 = runif(100)
n = 100
error = rnorm(n,0,sd=2) ## students should set  sd = 2
Y = beta_0 + beta_1*X1 + beta_2*X2 + error
```

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

```
 > B = 5000
> sigmahat = rep(NA,B)
>
> for(i in 1:B){
+   error = rnorm(n,0,2)
+   Y = beta_0 + beta_1*X1 + beta_2*X2 + error
+   fit = lm(Y~X1+X2)
+   sigmahat[i] = sum((fit$residuals)^2)/(n-(p+1))
+ }
>
> mean(sigmahat)
[1] 3.996724
```

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



(e) Explain why an estimate of σ^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?

If we don't have a good estimate of σ^2, then we don't have a good sense of what the variability is in our distribution of ϵ, and subsequently Y. This leads to problems. Specifically, all of our inference procedures break down: hypothesis testing (both for t-tests and F-tests), confidence intervals, and prediction intervals. Moreover, we assume constant variance just to fit a least square model. If we don't have a good estimate of σ^2, we cannot verify that this assumption holds.

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 β_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.

A significance level is a threshold that determines whether or not the evidence we see against our null hypothesis is meaningful, while still protecting us from false discoveries. We control our risk of a false discovery (type I error) by setting the significance level to a certain value (for example, 0.05). Formally, a type I error is when we reject the null hypothesis, but we really shouldn't have. If we set $\alpha = 0.05$, *that means we are willing to accept a 5% chance that we incorrectly reject* H_0, *even though the* H_0 *is true.*

(b) Suppose that the the p-value for the regression coefficient β_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.

Students should state they disagree. First, setting $\alpha = 0.05$ *is somewhat arbitrary. What is stopping us from setting* $\alpha = 0.07$*? However, by convention in academia and industry,* α *is usually set to be* 0.05 *(or* 0.01*). Our p-value is* 0.0647*, which is borderline. In terms of practical implications, most likely there is some relationship between anxiety and patient satisfaction, we just may not have enough power to detect it. If we could increase the power of our test (for example, by collecting more data), it's possible we would have obtained significant results here.*

(c) 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 $\alpha = 0.1$.

This essentially boils down to a multiple testing problem: the more tests we carry out, the more likely we are to see a significant result just by chance. If we carry out 12 individual tests, then probability of seeing at least one significant result by chance, for $\alpha = 0.1$*:*

$$\begin{aligned} P(\text{at least one significant result }) &= 1 - P(\text{no significant results}) \\ &= 1 - (1 - 0.1)^{12} \\ &= 0.71 \end{aligned}$$

This shows us that we have a roughly 70% chance of seeing a significant result, even when the H_0 *is true. This is much larger than the type I error we have set at 0.1. This can make our conclusions misleading because we don't know which p-values are false discoveries and which are truly significant.*

End of assignment.