DS 303 Homework 3
Due: Sept. 11, 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: 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.

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

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

(d) Obtain the RSS for the full model and the RSS for reduced model. Report them both here.

(e) Assume that our random errors (ϵ_i) are normally distributed. Carry out the F-test at $\alpha = 0.05$. Write out the null/alternative hypothesis, test statistic, null distribution, p-value, and conclusion.

(f) 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.

(g) Same setting as part (f). 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.

(h) Obtain the prediction for Y using all the same settings as (f), 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.

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 should we expect to make?

(b) Suppose we are carrying out m hypothesis test and we reject H_0 if its corresponding p-value falls below α (i.e. controlling Type 1 error for each null hypothesis at level α). Let V represent the number of type 1 errors. For m hypothesis tests, calculate $P(V \geq 1)$. Your value should be expressed in terms of m and α.
This probability $P(V \geq 1)$ is called the family-wise error rate (FWER). Ideally, we would like the FWER to be controlled at α such that FWER $\leq \alpha$.

(c) Describe a data application/setting where the multiple testing problem could be especially problematic if not addressed properly.

(d) Your colleague suggests a very simple approach to address this multiple testing problem. They suggest that instead of using α as our cutoff for rejecting H_0, we use α/m (where m is the number of hypothesis testings we are carrying out). Does your colleague's approach make statistical sense? Show that your colleague's approach controls the FWER at α (in other words FWER $\leq \alpha$).

(e) Repeat our in-class activity using your colleague's approach in (d). How many predictors are significant using the cutoff of α/m?

(f) List one potential drawback of your colleague's approach in (d).

Problem 3: Diagnostics for MLR

(a) List the four assumptions we make when fitting a multiple linear regression model.

(b) True or False? In order to fit a multiple linear regression model (i.e. obtain least square estimates), we must have distributional assumptions on the random error term ϵ_i. *Briefly justify your answer.*

(c) True or False? The `lm()` output automatically gives you p-value for individual hypothesis tests: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, $j = 1, \ldots, p$. These hypothesis tests assume that the random error term are normally distributed. *Briefly justify your answer.*

(d) Suppose we want to run some visual diagnostics to check whether or not the linearity assumption holds. Besides the fact that it is is time consuming, explain why it is not sufficient to check every pairwise scatterplot between Y and X_j, $j = 1, \ldots, p$. Hint: think back to HW 1, Problem 3(j).

(e) To check whether or not the linearity assumption holds, we can run some diagnostics. Let's use the `Auto` dataset as an example. Run the following code:

```
m1 = lm(mpg~horsepower,data=Auto)
summary(m1)
par(mfrow=c(2,2))
plot(m1)
```

Take a look at the plot in the top left corner (titled Residuals vs Fitted). This is what we call a *residual plot* because it is literally a plot of residuals versus fitted values ($\hat{y}_i$). The red line is a smooth fit to the residuals, which is displayed in order to make it easier to identify any trends. You should observe that the plot exhibits a clear U-shape, which provides a strong indication of non-linearity in the data.

Ideally what we want to see is **no discernible pattern** in the residual plot and a relatively flat red line. A residual plot exhibiting no pattern and a flat red line indicates that the linearity assumption for our model holds. Propose a more flexible model than `m1` that might be able to accommodate some non-linearity. Present your proposed model (call it `m2`) and its corresponding residual plot here. Is it an improvement over `m1`?

Note we'll discuss the remaining plots from `plot(m1)` and their utility in an upcoming lecture.