EXERCISE SET 6

Exercise 1 (Car brands and mpg)**:** In this exercise we will consider the data set containing information about cars and their miles per gallon. This can by loaded by

```
> data = pd.read_csv("https://raw.githubusercontent.com/intro-stat-learning
>                          /ISLP/main/ISLP/data/Auto.csv",encoding = "ISO-8859-1")
> data["name"] = [name.split()[0] for name in data["name"].values]
```

The second line takes the original names (which are the specific models – e.g. Toyota Yaris) and extracts only the brand name (e.g Toyota). We are going to study which brands have the best mpg. Some brands tend to make larger and heavier cars (e.g. pickup tricks) which will have worse mpg, but we want to understand how brands compare within a certain type of car. To determine this we need to control for other factors, such as the the year and weight.

(a) Using all the columns **except** `origin` and `displacement` (since it's not obvious what the units are), write down the regression model which you want to fit to this data to address the question posed in the problem instruction. Assume there are no interactions. Provide an interpretation of each regression coefficient.
(b) Fit the regression model to the data.
(c) What are the 5 best brands for mpg within the same type of car (weight, horsepower etc.).

Exercise 2 (Marginal regression in interactions model)**:** Consider the probability model

$$\begin{aligned} X_1 &\sim \text{Normal}(0, \sigma_1^2) \\ X_2 &\sim \text{Normal}(0, \sigma_2^2) \\ Y|(X_1, X_2) &\sim \text{Normal}(\beta_1 X_1 + \beta_2 X_2 + \beta_{1,2} X_1 X_2, \sigma^2) \end{aligned}$$

(a) Derive the distributions of $Y|X_1$ and $Y|X_2$. Hint: These conditional distributions are both normal, so you only need to determine the mean and variance to find the distributions.
(b) When does the probability model stated in the problem define regression models for Y vs. X_i, $i = 1, 2$? That is, if we ignore one of the predictor variables do obtain a single predictor linear regression model for the other?

Exercise 3 (Predicting the residual plot based on interaction model)**:** Suppose we have 200 data points generated from the following model

$$Y = 4X_1 - 2X_2 + 4X_1X_2 + \epsilon \tag{1}$$

where $\sigma = 0.2$, x_1 is continuous predictor which is uniformly distributed between -1 and 1 and X_2 is a binary predictor (e.g. a Bernoulli random variable). You can assume $X_2 = 0$ for about half the data points. The goal of this problem is to build your intuition about residual plots.

(a) **Without actually fitting a regression**, describe in detail what the residual plot would look like if we fit this data to a linear regression model with NO interaction term. To do so, follow the following procedure
- First, think about what the data looks like when $X_2 = 0$ and $X_2 = 1$ separately. In each case, sketch the regression line and make note of how much variation there is around these lines to get an idea of what the cloud of (X_i, Y_i) points will look like.
- Now consider what the fitted regression line will be based on this picture. What is a very rough estimate of the slopes $\hat{\beta}_1$ and $\hat{\beta}_2$?
- To get a sense for what the residuals look like, take the difference between the true model and this line.

(b) Confirm you answer with simulations.

Exercise 4 (Drug interactions)**:** When treating microbial infections and cancer, combinations of drugs can perform better than individual drugs. However, it can be difficult to identify which combinations are optimal for the reason that identifying very "high order" interactions is difficult. In order to understand the best way to combine M drugs, we construct a regression where Y is the "effect" of the drug and X_i is a Bernoulli random variable representing whether or not the ith drug is present or not. We want to consider the possibility

$$Y = \sum_{i=1}^{M} \beta_i X_i + \sum_{i=1}^{M} \left(\sum_{j>i}^{M} \beta_{i,j} X_i X_j \right) + \sum_{i=1}^{M} \left(\sum_{j>i}^{M} \sum_{k>j}^{M} \beta_{i,j} X_i X_j X_k \right) + X_1 X_2 \cdots X_M + \epsilon$$

For example, with $M = 3$, we would have

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \beta_{2,3} X_2 X_3 + \beta_{1,2,3} X_1 X_2 X_3 + \epsilon$$

(a) Suppose $M = 3$ and

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_{1,2} \\ \beta_{1,3} \\ \beta_{2,3} \\ \beta_{1,2,3} \end{bmatrix} = \begin{bmatrix} 1.2 \\ -0.8 \\ -0.11 \\ 3.48 \\ -2.62 \\ 1.03 \\ 1.66 \end{bmatrix}$$

What is the interpretation of each coefficient?

(b) What is the optimal treatment, meaning which combination of drugs 1, 2 and 3 should we use to maximize Y? There are different ways you can approach this. One way is to make a list of each (X_1, X_2, X_3), compute Y for each one and the find the index of the maximum Y value (using a for loop or `argmax`).

(c) Now additional suppose that $\sigma^2 = 1$. By generating simulated Y values with these parameters for different values of N, determine how many data points are needed to reliably find that all interactions coefficients have p-values below 0.05

(d) Perform the same experiment as in (c) but fit the data to a model with no interactions. What do you find? How does adding the interaction terms influence the p-values.