EXERCISE SET 5

Exercise 1 (Car brands and mpg – **ChatGPT use recommended**)**:** 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 year and weight.

(a) Using all the columns **except** `origin` and `displacement` (since it's not clear 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 in terms of differences in conditional expectation. You don't need to list each brand, but explain in general how the brand predictors are included and give their interpretation.

(b) Fit the regression model to the data. There are slightly different ways to break up the brands and I'll leave it to you to decide exactly how to handle special cases, like brands that are actually subsidiaries of another, or the same brand marketed to different countries.

(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

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

(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, but either one can depend on the predictor we are conditioning on.

(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 you obtain a single predictor linear regression model for the other? Is that fact that the predictors have mean zero import here?

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

$$X_1 \sim \text{Uniform}(-1, 1)$$
$$X_2 \sim \text{Bernoulli}(1/2)$$
$$Y|X \sim \text{Normal}(4X_1 - 2X_2 + 4X_1X_2, 1/4)$$

The goal of this problem is to build your intuition about residual plots.

(a) **Without actually fitting a regression**, describe in detail the residual plot would look like if we fit this data to a linear regression model with NO interaction term. To guide you through this process, here is an outline of the general approach you'll want to take:

- 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 – **ChatGPT use recommended**)**:** Suppose that cancer cells from a live cell biopsy are treated with 3 different drugs, A, B and C and all possible combinations of them. The present of each drug is represented binary variables $X_i \in \{0, 1\}$ with $i \in \{A, B, C\}$. The reduction in cancer cells is measured by Y, the difference in mass before an after treatment. To model the effects of each drug and their interactions, we use the regression model

$$Y = \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_{A,B} X_A X_B + \beta_{B,C} X_B X_C + \beta_{C,A} X_C X_A + \epsilon$$

(a) Suppose it is found that $\beta_A = -1, \beta_B = 2, \beta_C = 1.2, \beta_{A,B} = 3, \beta_{B,C} = -1$ and $\beta_{C,A} = 2$. Which combination of drugs will optimize the reduction in tumor size? Hint: Ask ChatGPT to write python code to generate all possible combinations of 3 binary numbers and put them in a numpy array.

(b) Write a function which generates simulated data from this model using the parameters above and $\sigma_\epsilon^2 = 1/4$.

(c) Using the function you've written, determine how much data would be needed to get a p-value of less than 0.005 on each of the regression parameters. Answer the same question for the model *without* the interactions terms.

Exercise 5 (Bias variance trade-off)**:** As we have seen, if we perform n independent Bernoulli trails $X_1 \ldots, X_N$, our best estimate of q is $\hat{q} = S/N$ where $S = \sum_{i=1}^N X_i$. Laplace's rule of succession is to use the estimate

$$\hat{q}_L = \frac{S+1}{N+2}$$

The idea behind $\hat{q}_L$ is that we assume it is possible for the coin to land on either heads or tails. Hence we assume we have already seen two flips (the $+2$ in the denominator) and one is heads (the $+1$ in the numerator).

(a) Calculate the MSE of $\hat{q}_L$ and decompose it into bias and variance.

(b) For what values of q does $\hat{q}_L$ have a lower MSE? Does this make intuitive sense?

Exercise 6 (Polynomial regression)**:** Consider the model

$$X \sim \text{Uniform}(0, 1)$$
$$Y|X \sim \text{Normal}\left(\sum_{j=1}^K \phi_j(X), \sigma^2\right)$$

where $\phi_j(x) = x^j$.

(a) Calculate the correlation coefficient $\rho_{i,j}$ between $\phi_i(X)$ and $\phi_j(X)$. Hint: You will need to evaluate expectations of the form $\mathbb{E}[X^n]$, which under the assumption that X is uniform on $(0, 1)$ become integrals $\int_0^1 X^n dx = 1/(n+1)$.

(b) For different values if i, plot $\rho_{i,j}$ this as a function of j. What do you notice? Comment on the implications of this for the problem of inferring the true relationship between Y and X.

Exercise 7 (Indicator basis functions)**:** Consider the model

$$Y|X \sim \text{Normal}\left(\sum_{i=1}^K \beta_i \phi_i(X), \sigma^2\right)$$

Suppose that $X \in [0, 1)$ and define the intervals

$$I_i = \left[\frac{i-1}{K}, \frac{i}{K}\right)$$

Notice that

$$[0,1) = I_1 \cup I_2 \cup \cdots \cup I_K.$$

That is, each x in $[0,1)$ is in one of these disjoint intervals. Now introduce the basis functions

$$\phi_i(x) = \begin{cases} 1 & x \in [(i-1)/K, i/K) \\ 0 & x \notin [(i-1)/K, i/K) \end{cases}$$

(a) In general, are ϕ_i orthogonal with respect to a random variable X taking values in $[0,1)$? Does it depend on the distribution of the random variable? Test your conclusion with simulations.

(b) Using `statsmodels`, implement fitting the model with these features. You can make up your simulated data set to fit, or copy the code I used in class to fit the fourier and polynomial models. I recommend writing a function `phi(x,i)` which takes the array of predictors and outputs an array $[\phi_j(X_1), \ldots, \phi_j(X_N)]$. Use $K = 10$ and $N = 100$. **ChatGPT use recommended**

(c) (**Optional**) As usual let $\hat{\beta}_j$ be the fitted value of β_j using least squares, meaning the value that minimizes the squared residuals. Show that in this model $\hat{\beta}_j$ is simply the average value of Y_i among data points where $X_i \in I_j$; that is,

$$\hat{\beta}_j = \frac{1}{N_j} \sum_{i : X_i \in I_j} Y_i$$

where N_j is the number of points in I_j.