229352 Statistical Learning for Data Science 1 **Spring 2020**

Homework 2: due January 22

1. Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 =$ Interaction between GPA and IQ, and $X_5 =$ Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = 10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

2. This question should be answered using the `Carseats` data set.

   (a) Fit a multiple regression model to predict `Sales` using `Price` ,`Urban`, and `US`.

   (b) Provide an interpretation of each coefficient in the model. Be careful–some of the variables in the model are qualitative!

   (c) Write out the model in equation form, being careful to handle the qualitative variables properly.

   (d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

   (e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

   (f) How well do the models in (a) and (e) fit the data?

   (g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

   (h) Is there evidence of outliers or high leverage observations in the model from (e)?

3. This problem focuses on the collinearity problem.

   (a) Perform the following commands in python:

```python
import numpy as np
np.random.seed(1)
x1 = np.random.uniform(size=(100,))
x2 = 0.5*x1 + np.random.normal(size=(100,))/10.
y  = 2 + 2*x1 + 0.3*x2 + np.random.normal(size=(100,))
```

   The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

(b) What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

(c) Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1, and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

(d) Now fit a least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(e) Now fit a least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(f) Do the results obtained in (c)(e) contradict each other? Explain your answer.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1 = np.append(x1, [[0.1]], 0)
x2 = np.append(x2, [[0.8]], 0)
y  = np.append(y, 6.0)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.