COMS 474
Homework 2

Haadi Majeed

Spring 2022

Problem 1

[13 points total (6,3,4)]
Book problem Chapter 3, problem 3 "Suppose we have a data set with five predictors, $X_1 = GPA, X_2 = IQ, X_3 = Level$ 1 for College and 0 for High School, $X_4 =$ Interaction between GPA and IQ, and $X_5 =$ Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$"
Note: for interactions, use products, e.g. $X_4(i) = GPA(i) * IQ(i)$

1. Which answer is correct, and why?

   (i) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

   (ii) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

   (iii) For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

   (iv) For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

   Answer *iii* is correct, by expanding the formula we get

$$\hat{Y} = \beta_0 + \beta_1 * (X_1) + \beta_2 * (X_2) + \beta_3 * (X_3) + \beta_4 * (X_4) + \beta_5 * (X_5)$$

And we know $X_4 = X_1 * X_2$ and that $X_5 = X_1 * X_3$ we can sub these in along with the β values to get

$$\hat{Y} = 50+20*(X_1)+0.07*(X_2)+35*(X_3)+0.01*(X_1*X_2)-10*(X_1*X_3)$$

From there we can plug in values we know, such as X_3 and see how it looks for highschoolers vs college students

Highschool:

$$\hat{Y} = 50+20*(X_1)+0.07*(X_2)+35*(0)+0.01*(X_1*X_2)-10*(X_1*0)$$
$$\hat{Y} = 50 + 20 * (X_1) + 0.07 * (X_2) + 0.01 * (X_1 * X_2)$$

College:

$$\hat{Y} = 50+20*(X_1)+0.07*(X_2)+35*(1)+0.01*(X_1*X_2)-10*(X_1*1)$$
$$\hat{Y} = 85 + 10 * (X_1) + 0.07 * (X_2) + 0.01 * (X_1 * X_2)$$

from this we can get down to if the GPA is less than 3.5, then college students would have more, however with a GPA of 3.5 or greater, a highschool student would have more.

2. Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.
   Using the formula derived from above and assigning the following: $X_2 = 110$ and $X_1 = 4.0$ we get
   $\hat{Y} = 85 + 10 * (4.0) + 0.07 * (110) + 0.01 * (4.0 * 110)$
   $\hat{Y} = 137.1$ or about \$137,100

3. 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.
   False, the coefficient does not provide evidence for or against the interaction. We would need to test for a proper conclusion.

Problem 2

[12 points total (3 points each)]
Book problem Chapter 3, problem 4 "I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$."

1. Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
   For the training data, it is harder to tell since a model with 3 features would be more flexible than a linear one, thus the RSS should be lesser in comparison. However, since it truly is linear, we may be fitting excessively and including noise/error.

2. Answer (a) using test rather than training RSS.
   For the test set of data, it would be expected that the RSS would be smaller for linear regression verses the cubic model and will fit the data better. RSS may have been low, but at the cost of overfitting the training data for better predicting future data.

3. Suppose that the true relationship between X and Y is not linear,but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
   Polynomial models has lower training RSS due to the higher flexibility it has over linear. Because it is more flexible it will fit the training data closer and the RSS will reflect that.

4. Answer (c) using test rather than training RSS.
   Because there is not enough information to determine which RSS would be lesser, we cannot determine how close or far it is to a linear model. If we had that information we could determine if linear's RSS was lower or if the cubic's was lower. Due to this it is not clear which model would fit most ideally.

Problem 3

[5 points]
Suppose we have a data set with one feature X to predict another feature Y. Let n denote the number of samples. Let $\bar{X}$ and $\bar{Y}$ denote the average values of X and Y respectively in the dataset:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X(i) \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y(i)$$

Let β_0^* and β_1^* denote the coefficients for the ordinary least squares (O.L.S) solution,

$$\{\beta_0^*, \beta_1^*\} = \hat{f}(x,y) = \arg\min_{\{\beta_0^*,\beta_1^*\}} \frac{1}{n}\sum_{i=1}^{n}\Big(Y(i) - (\beta_0 + \beta_1 * X(i))\Big)^2$$

Their values are

$$\beta_0^* = \bar{Y} - \beta_1^*\bar{X} \qquad \beta_1^* = \frac{\sum_{i=1}^{n}(X(i)-\bar{X})(Y(i)-\bar{Y})}{\sum_{i=1}^{n}(X(i)-\bar{X})^2}$$

Using those formulas, calculate the O.L.S. model's prediction for $X = \bar{X}$ (i.e., the prediction $\bar{Y}$ for a new sample for whose X feature has the value $\bar{X}$)

$$\beta_0^* = \bar{Y} - \beta_1^* * \bar{X} \rightarrow \frac{1}{n}\sum_{i=1}^{n}(Y(i) - \beta_1^* * X)$$

$$\beta_1^* = \beta_1^* = \frac{\sum_{i=1}^{n}(X(i)-\bar{X})(Y(i)-\bar{Y})}{\sum_{i=1}^{n}(X(i)-\bar{X})^2}$$

$$\rightarrow \beta_1^* = \frac{\sum_{i=1}^{n}(X(i)-X)(Y(i)-\frac{1}{n}\sum_{i=1}^{n}Y(i))}{\sum_{i=1}^{n}(X(i)-X)^2}$$

$$\rightarrow \beta_1^* = \frac{\sum_{i=1}^{n}(X(i))(\frac{1}{n}Y(i)}{\sum_{i=1}^{n}(X(i))^2}$$

Problem 4

[16 points total (3,3,10)]
Book problem Chapter 6, problem 1 "We perform best subset, orward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p + 1 models, containing $0, 1, 2, ...p$ predictors. Explain your answers"
Notes regarding the book's pseudocode for Algorithms 6.1-6.3:

- "RSS" stands for "residual sum of squares" which is the (un-normalized) MSE,
  $RSS = \sum_{i=1}^{n}(Y(i) - \overline{Y})^2$

  In the book's pseudo-codes, "RSS" refers to training set RSS.
- "cross-validated prediction error" - you can read this as "Validation set MSE."
- "R^2" and "adjusted R^2" - you can ignore these for this homework.

1. Which of the three models with k predictors has the smallest *training* RSS?
   A model utilising best subset selection would perform best because it can fit all 2^p possible models with p being a predictor. If the critera is minimizing the RSS, no predictor that forward selection or backward selection could find would not be found by subset selection as well.
2. Which of the three models with k predictors has the smallest *test* RSS?
   Subset selection is most likely to do the best out of the three, as all three will be based on a comboination to minimize the training RSS. As stated before, there is a possiblilty that forward or backward selections can do better than subset, it is unlikely due to how many more models subset selection has to choose from, especially as the number of predictors increases.
3. True or False:

(i) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k + 1)-variable model identified by forward stepwise selection.
True → The k + 1 model will be identical but with one more predictor

(ii) The predictors in the k-variable model identified by back-ward stepwise are a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection.
True → The k + 1 model will be identical but with one less predictor

(iii) The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the (k + 1)-variable model identified by forward stepwise selection.
False → The forward and backward selections will have different start points and paths. Not always true, but has the chance of being true

(iv) The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection.
False → The forward and backward selections will have different start points and paths. Not always true, but has the chance of being true

(v) The predictors in the k-variable model identified by best subset are a subset of the predictors in the (k + 1)-variable model identified by best subset selection.
False → Will not always have the best subset variable of size k+1.