MATH 4322, Intro to Data Science & Machine Learning, Homework # 2.

Instructors: Cathy Poliak, Wendy Wang

DUE: Saturday, September 21st, at 11:59PM.

Instructions: Submit the solutions as a file (type it up and save as a *.pdf* or a *Word*-file, no hand-written solutions) via UH Blackboard. Keep responses brief and to the point. For code & output: include only pieces that are of utmost relevance to the question.

Conceptual.

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond, with response *sales* being regressed on predictors TV, *radio* and *newspaper*. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of *sales*, TV, *radio*, and *newspaper*, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

2. Suppose we have a data set on flights, with three predictors, $X_1 = Distance$ (to destination, in miles), $X_2 = Holiday$ (1 if Yes, 0 for No), $X_3 =$ Interaction between $Distance$ and $Holiday$. The response is the flight ticket price (in dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 100, \hat{\beta}_1 = 0.2, \hat{\beta}_2 = 40, \hat{\beta}_3 = 0.05$.

   (a) Which answer is correct, and **why**?

      i. For a fixed value of $Distance$, on average tickets are more expensive on holidays than on usual days.

      ii. For a fixed value of $Distance$, on average tickets are more expensive on usual days than on holidays.

iii. For a fixed value of *Distance*, on average tickets are more expensive on usual days than on holidays, provided that *Distance* is long enough.

iv. For a fixed value of *Distance*, on average tickets are more expensive on holidays than on usual days, provided that *Distance* is long enough.

(b) Predict the average holiday price of a ticket for a flight that travels 1000 miles to its destination.

(c) True or false: Since the coefficient for the $Distance/Holiday$ interaction term is pretty small, there is no evidence of an interaction effect. Justify your answer.

3. 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. $Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

(a) 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.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we dont 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.

(d) Answer (c) using test rather than training RSS.

Applied.

4. This question involves the use of simple linear regression on the *Auto* data set.

(a) Use the $lm()$ function to perform a simple linear regression with *mpg* as the response and *horsepower* as the predictor. Use the $summary()$ function to print the results. Comment on the output. For example:

i. Is there a relationship between the predictor and the response?

ii. How strong is the relationship between the predictor and the response?

iii. Is the relationship between the predictor and the response positive or negative?

iv. What is the predicted *mpg* associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

(b) Plot the response and the predictor. Use the $abline()$ function to display the least squares regression line.

(c) Use the *plot*() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

5. This question involves the use of multiple linear regression on the *Carseats* data set.

   (a) Produce a scatterplot matrix which includes all of the **numerical** variables in the data set. Exclude all the qualitative variables (Hint: use function *str*(*data*) to determine which variables of data frame *data* are numerical and which are factors).

   (b) Compute the matrix of correlations between the **numerical** variables using the function *cor*().

   (c) Use the *lm*() function to perform a multiple linear regression with *Sales* as the response and all other **numerical** variables as the predictors. Use the *summary*() function to print the results. Comment on the output. For instance:

      i. Is there a relationship between the predictors and the response?

      ii. Which predictors appear to have a statistically significant relationship to the response?

      iii. Interpret the coefficients for *Price* and *CompPrice*, respectively.

   (d) Use the *plot*() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit (see question 4(c)).

6. This question should be answered using the *Carseats* data set.

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

   (b) Provide an interpretation of each coefficient in the model. Be careful - there is a quantitative variable in the model!

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

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

      i. at significance level $\alpha = 0.05$?

      ii. at significance level $\alpha = 0.01$?

   (e) Use " : " symbol to fit an extension of the model from (a) by adding interaction effects $Price \times US$ and $Income \times US$. Do any of the interactions appear to be statistically significant?

   (f) Compare the fits of models (a) and (e). What quantities do you use to compare models? Which model appears to fit better according to those quantities?

7. This problem involves the *Boston* data set. We will try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response?

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H_0: $\beta_j = 0$?

(c) How do your results from (a) compare to your results from (b)? Provide a summary comment on that comparison.

(d) Fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

for each predictor X **individually**. Is there evidence of non-linear association between any of the predictors and the response?