CS 4342 Assignment #2
Total points: 60

Conceptual and Theoretical Questions

1. Describe the null hypotheses to which the p-values given in the below Table correspond. 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.*

Points: 5

2. Carefully explain the differences between the KNN classifier and KNN regression methods.

Points: 5

3. Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = 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.

Points: 5

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. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(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 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.
(d) Answer (c) using test rather than training RSS.
Points: 5

Applied Questions

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

(a) Perform a simple linear regression with mpg as the response and horsepower as the predictor and answer the following questions:

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 95?

(b) Plot the response and the predictor along with the predicted line.

Hints:

- Check https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html
- Check https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Points: 20

2. This question involves the use of multiple linear regression on the Auto data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

(b) Compute the matrix of correlations between the variables. You will need to exclude the name variable which is qualitative.

(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Examine the results, and 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. What does the coefficient for the year variable suggest?

(d) Produce diagnostic plots of the linear 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?

(e) Fit linear regression models with predictors and interaction terms. Do any interactions appear to be statistically significant?

(f) Fit linear regression models with only interaction terms. Do any interactions appear to be statistically significant?

(g) Try a few different transformations of the variables, such as log(X), $\sqrt{X}$, X^2. Comment on your findings.

Hints:

- Check https://pandas.pydata.org/docs/reference/api/pandas.plotting.scatter_matrix.html
- Check NumPy, SciPy, or Pandas for correlation
- Check https://www.statsmodels.org/stable/generated/statsmodels.graphics.regressionplots.influence_plot.html for the leverage plot
- Check http://joelcarlson.github.io/2016/05/10/Exploring-Interactions/ for the interaction term. You can build them manually too.

Points: 20