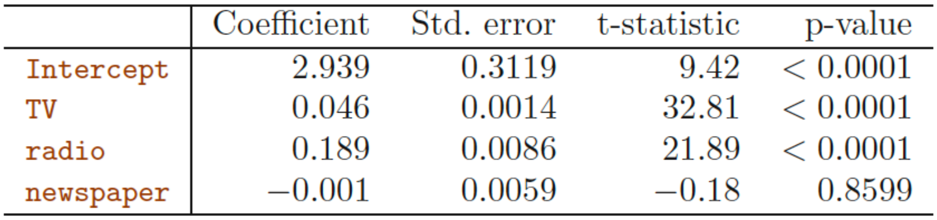
CS 4342 Assignment #2

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 forMale), 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

5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$$\hat{y}_i = x_i \,\hat{\beta},$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right)$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

What is $a_{i'}$?
Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.
Points: 5

6. Using equation (3.4) – shown below, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right) / \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{(3.4.a)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{(3.4.b)}$$

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?

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

(f) 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

3. 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)?

Points: 20

4. This problem involves the Boston data set, which we saw in the previous HW. We will now 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? Create some plots to back up your assertions.

(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)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

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

Points: 10