Data Analysis and Decision Making

(Fall 2018)

Rutgers Business School – Newark and New Brunswick

Instructor: Debopriya Ghosh

Final Exam

Possible Points: 100

Please read all of the following information before starting the examination:

- This is a take home examination. Students are allowed to use their lecture notes and other course resources. However, this is not a group assignment. Please refrain from discussing and collaborating with other colleagues.

  Similar solutions will be penalized.

- This test consists of 5 questions each question is worth 20 points.
- Please turn in you R codes along with the explanations.
- The due date for submission is December 7, 12 pm.

  **No late submission would be accepted.

Good Luck!

Problem 1. (5 + 5 + 10 = 20 points)

(a) Describe the null hypotheses to which the p-values given in the following 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	< 0.001
TV	0.046	0.0014	32.81	< 0.001
radio	0.189	0.0086	21.89	< 0.001
newspaper	-0.001	0.0059	-0.18	< 0.001

(b) It is claimed that in case of simple linear regression of Y onto X, the R^2 statistic is equal to the square of the correlation between X and Y. Prove this is the case. For simplicity assume $\bar{x} = \bar{y} = 0$.

(c) Use the iris data set for predicting iris species based on the other predictor variables. Start by randomly splitting the data into training set (80% for building a predictive model) and test set (20% for evaluating the model). Make sure to set seed for reproducibility.

Problem 2. (20 points)

This question involves the use of simple linear regression on **Auto** dataset (ISLR package)

(a) Build a linear regression model with **mpg** as the response and **horsepower** as the predictor. Comment on the following:

i. Is there a relationship between the predictor and response?
ii. How strong is the relationship between the predictor and the response?
iii. Is the relationship between the predictor and response positive or negative?
iv. What is predicted mpg associated with **horsepower** of 98? What is the associated 95% confidence and prediction intervals?

Hint: Use `predict(model, newdata, interval="predict")`

(b) Plot the response and the predictor. Display the least square regression line.
(c) Use **plot ()** function to produce diagnostic plots of the least square regression fit. Comment on any problems you see with the fit.

Problem 3. (20 points)

This question involves the use of multiple linear regression on the **Auto** dataset.

(a) Produce a scatterplot matrix which includes all of the variables in the dataset.
(b) Compute the matrix of correlations between the variables using the function **cor ()**. You will need to exclude **name** variable which is qualitative.
(c) Perform multiple linear regression with **mpg** as the response and all other variables except **name** as the predictors. Comment on the following:
    i. Is there a relationship between predictors and the response?
    ii. Which predictors appear to have a statistically significant relationship to the response?
    iii. What does the coefficient of the **year** variable suggest?
(d) Produce the diagnostic plots of linear regression fit. Comment on any problems you see with the fit. Does the residual plot suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?
(e) Try different transformations of the variables such as $log\ (X)$, $\sqrt{X}$, X^2. Comment on your findings.

Problem 4. (10 + 10 = 20 points)

(a) Suppose we collect data for a group of students in a statistic class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient $\widehat{\beta_0} = -6, \widehat{\beta_1} = 0.05, \widehat{\beta_2} = 1$.
    i. Estimate the probability that a student who studies 40 hours and has an undergrad GPA of 3.5 gets an A in the class.
    ii. How many hours would the student in part (i) need to study to have a 50% chance of getting an A in the class?

(b) Hawkins hospital software team built a classification model for diagnosing Breast Cancer in women. A sample of 1000 women were studied in a given population and 100 of them with Breast Cancer while the remaining 900 were without it. Hawkins software team trained their model based of this dataset. They split the dataset into 70/30 train/test set. The accuracy was excellent and they deployed the model. A couple of months after deployment, some of the women who were diagnosed by the hospital as having “no breast cancer” started showing symptoms of Breast Cancer. This raised a series of questions and fear amongst the entire population. Hawkins hospital had to do something about this as more and more patient started showing symptoms of Breast Cancer. They decided to hire a Machine Learning Expert, Jane to help them understand what their software team got wrong considering the fact that the model had an accuracy of about 90%. Explain what went wrong.

Problem 5. (20 points)

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the **Auto** dataset.

(a) Create a binary variable, **mpg01**, that contains a 1 if **mpg** contains a value above its median, and a 0 if **mpg** contains a value below its median.
(b) Explore the data graphically in order to investigate the association between **mpg01** and the other features. Which of the other features seem most likely to be useful in predicting **mpg01**? Scatter plots and boxplots may be useful tools to answer this question. Describe your findings.
(c) Split the data into training and test set.
(d) Perform logistic regression on the training data in order to predict **mpg01** using the variables that seemed most associated with **mpg01** in (b). What is the test error of the model obtained?