STAT 847: Assignment 1 - copyright L. Xing, University of Saskatchewan

Date:

(Due date: 11:59pm (Time zone: GMT-6) on Oct 1st, 2023. Late assignments will be accepted only for seven (7) days beyond the due date. The penalty for the delay is 10 percent per day of lateness from the value of the assignment, including weekend days.)

TOTAL MARKS

NSID . . . .

Email . . . .

Question 1: (5 marks) This question involves the use of multiple linear regression on the *Auto* data set.

Auto Data Set Description
-*mpg*: miles per gallon;
-*cylinders*: Number of cylinders between 4 and 8;
-*displacement*: Engine displacement (cu. inches);
-*horsepower*: Engine horsepower;
-*weight*: Vehicle weight (lbs.);
-*acceleration*:Time to accelerate from 0 to 60 mph (sec.);
-*year*: Model year (modulo 100);
-*origin*:Origin of car (1. American, 2. European, 3. Japanese);
-*name*:Vehicle name.

(a) (0.5 mark) Produce a scatterplot matrix which includes all of the variables in the data set.

(b) (0.5 mark) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

(c) (1 mark) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name 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. What does the coefficient for the year variable suggest?

(d) (1 mark) Use the plot() function to 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) (1 mark) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

(f) (1 mark) Try a few different transformations of the variables, such as $log(X), \sqrt{X}, X^2$. Comment on your findings.

Question 2: (10 marks) In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.

```
set.seed (1)
x=rnorm (100)
y=2*x+rnorm (100)
```

(a) (1.5 marks) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command $lm(y \sim x + 0)$.)

(b) (1.5 marks) Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.

(c) (1 mark) What is the relationship between the results obtained in **(a)** and **(b)**?

(d) (4 marks) For the regression of Y onto X without an intercept, the t statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/SE(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38)

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

, and where

$$SE(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2}{(n-1)\sum_{i'=1}^{n} x_{i'}^2)}}$$

(These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as

$$\frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i'=1}^{n} y_{i'}^2) - (\sum_{i'=1}^{n} x_{i'} y_{i'})^2}}$$

(e) (1 mark) Using the results from **(d)**, argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

(f) (1 mark) In R, show that when regression is performed with an intercept, the t-statistic for $H_0 : \beta = 0$ is the same for the regression of y onto x as it is for the regression of x onto y.

Question 3: (10 marks) In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part **(a)** to ensure consistent results.

(a) (1 mark) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a N(0, 1) distribution. This represents a feature, X.

(b) (1 mark) Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.

(c) (1 mark) Using x and ϵ, generate a vector y according to the model

$$Y = -1 + 0.5X + \epsilon$$

What is the length of the vector y? What are the values of β_0 and β_1 in this linear model?

(d) (1 mark) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

(e) (1 mark) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta_0}$ and $\hat{\beta_1}$ compare to β_0 and β_1?

(f) (1 mark) Display the least squares line on the scatterplot obtained in **(d)**. Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.

(g) (1 mark) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

(h) (1 mark) Repeat **(a)-(f)** after modifying the data generation process in such a way that there is less noise in the data. The model in **(c)** should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in **(b)**. Describe your results.

(i) (1 mark) Repeat **(a)-(f)** after modifying the data generation process in such a way that there is more noise in the data. The model in **(c)** should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in **(b)**. Describe your results.

(j) (1 mark) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

Question 4: (8 marks) 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) (2 marks) 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) (2 marks) Answer **(a)** using test rather than training RSS.

(c) (2 marks) 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) (2 marks) Answer **(c)** using test rather than training RSS.

Question 5: (10 marks) 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

$$\hat{y}_i = \left(\sum_{i'=1}^{n} a_{i'} y_{i'} \right).$$

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.