Cpts 575 Data Science

Assignment 5

Jinyang Ruan

011696096

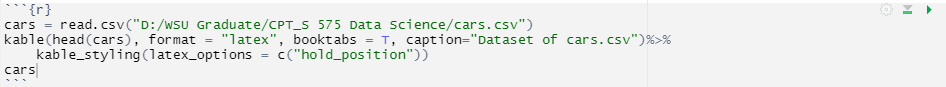
10/26/2021

1. Lode necessary libraries:

Shape, rectangle

Description automatically generated with medium confidence

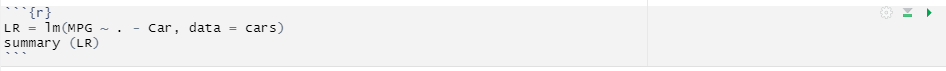
Load data set and present all values in the appropriate types.



Table

Description automatically generated

1. Perform a multiple linear regression with MPG as the response and all other variables except Car as the predictors.



Text

Description automatically generated with low confidence

i) Which predictors appear to have a statistically significant relationship to the response, and how do you determine this?

From the results, we can know that the predictors {Displacement, Weight, Model, Origin} have a statistically significant relationship to the response MGP.

We can determine this though the p-Value which indicate how significant the relationship is (the number of stars).

ii) What does the coefficient for the Displacement variable suggest, in simple terms?

The coefficient of displacement is positive.

It suggests that how much the value of “MPG” will increase when the number of displacements increases by one while keeping all the other predictors constant.

1. 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?

Chart, scatter chart

Description automatically generated with medium confidenceGraphical user interface, chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Comments:

(1) Residuals vs Fitted: the points are almost around a horizontal line, little pattern in residuals.

(2) Normal Q-Q: the points in the Q–Q plot approximately lie on a line, so the distributions are linearly related.

(3) Scale - Location: That the red line is approximately horizontal. Then the average magnitude of the standardized residuals isn’t changing much as a function of the fitted values.

The spread around the red line doesn’t vary with the fitted values. Then the variability of magnitudes doesn’t vary much as a function of the fitted values.

(4) Residuals vs Leverage:

The rightmost point has a high leverage which means that it has a high influence, that is, it determines how much the predicted scores will change if the point is excluded.

1. Fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

Table

Description automatically generated

Interaction such as weight: acceleration appears to be statistically significant.

1. Lode Boston data set:

Table

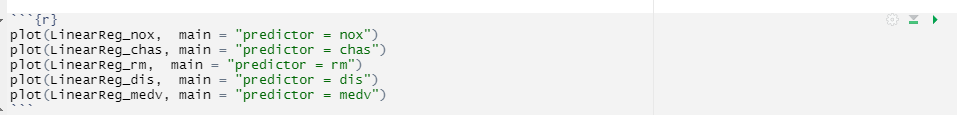
Description automatically generated

1. For each predictor, fit a simple linear regression model to predict the response. Include the code, but not the output for all models in your solution.

Application, table

Description automatically generated

1. In which of the models is there a statistically significant association between the predictor and the response? Considering the meaning of each variable, discuss the relationship between and ,, , and in particular. How do these relationships differ?



1. Predictor = “”

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

1. Predictor = “”

Chart

Description automatically generated with medium confidenceChart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

1. Predictor = “”

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

1. Predictor = “dis”

Chart, line chart

Description automatically generatedChart

Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

1. Predictor = “”

Chart, scatter chart

Description automatically generatedChart, line chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

There is a statistically significant association between the predictor and the response for all variables except .

From the figures above, we can see that there is linear relationship between the and crim. And among all the four predictors, gets the highest R Squared value.

There is no association between and . For the other four predictors , and, it is not a complete straight line in Residuals vs Fitted, so there is little pattern in residuals.

1. Fit a multiple regression model to predict the response using all the predictors. Describe your results. For which predictors can we reject the null hypothesis H0: βj = 0?

Graphical user interface, text, application

Description automatically generated

Comments:

i) There are five predictors including have significant association with crim.

ii) R-squared value is higher for multiple regression when being compared to the simple regressions.

iii) For the predictors , p-values are all less than 0.05, we can reject these predictors.

1. How do your results from (a) compare to your results from (c)?

Text

Description automatically generated

Chart

Description automatically generated

(1) The regression coefficients are different in univariate and multiple regression.

In univariate regression, we only consider the average effect of an increase in the specific predictor, while ignoring other predictors. In multiple regression, we consider the average effect of an increase in the predictor, while holding other predictors fixed.

(2) From the plot we can know the coefficient for most predictors are around 0 in both univariate and multiple regression.

1. Is there evidence of non-linear association between any of the predictors and the response?

Graphical user interface, text, application

Description automatically generated

One output instance:

A screenshot of a computer

Description automatically generated with medium confidence

Looking at the p-value, we can get the following observations:

(1) For predictors , the cubic coefficient is not statistically significant

(2) For predictors, the adequacy of the cubic fit

(3) For predictor , the quadratic and cubic coefficients are not statistically significant, there is no non-linear effect.

1. Suppose we collect data for a group of students in a statistics class with variables:

X1 = hours studied,

X2 = undergrad GPA,

X3 = PSQI score (a sleep quality index), and

Y = receive an A.

We fit a logistic regression and produce estimated coefficient, β0 = −7, β1 = 0.1, β2 = 1, β3 = -.04.

1. Estimate the probability that a student who studies for 30 h, has a PSQI score of 11 and has an undergrad GPA of 3.0 gets an A in the class. Show your work.
2. How many hours would the student in part (a) need to study to have a 60 % chance of getting an A in the class? Show your work.

Assume that the student needs to study h hours.

Then we can get:

1. How many hours would a student with a 3.0 GPA and a PSQI score of 5 need to study to have a 50 % chance of getting an A in the class? Show your work.

Assume that the student needs to study h hours.

Then we can get:

1. For this question, you will use a naïve Bayes model to classify consumer complaints by the category of financial product or service the complaints are related to.

Data set prepare:

Graphical user interface, text, application, email

Description automatically generated

1. Tokenization

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

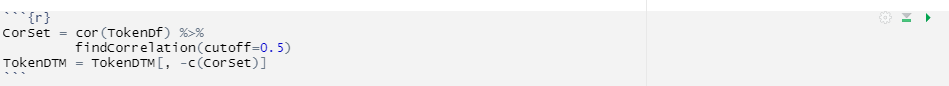
Show a non-zero entries of a random row:

Graphical user interface, text, application, email

Description automatically generated

1. Classification

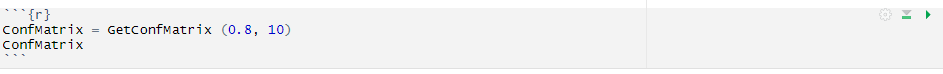
Reduce feature sets and remove correlated features:



Split data into a training set and a test set and Build Naïve Bayes classifier.

Graphical user interface, text, application, email

Description automatically generated



Graphical user interface, text

Description automatically generated

A picture containing table

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Try to get best splitting proportion by accuracy:

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

In this case we get best accuracy when proportion is 0.9.

Then try to get best low frequency by accuracy.

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text, email

Description automatically generated

Finally, we get the best accuracy 36.28 with low frequency = 10 and splitting proportion = 0.90

Show the confusion matrix:

Table

Description automatically generated with medium confidence

Table

Description automatically generated