**Introduction to Statistical Learning**

1) ISLR 2.4 Applied Problem 8.

This exercise relates to the College data set, which can be found in the file College.csv on the book website. It contains a number of variables for 777 different universities and colleges in the US. The variables are • Private : Public/private indicator • Apps : Number of applications received • Accept : Number of applicants accepted • Enroll : Number of new students enrolled • Top10perc : New students from top 10 % of high school class • Top25perc : New students from top 25 % of high school class • F.Undergrad : Number of full-time undergraduates • P.Undergrad : Number of part-time undergraduates 2.4 Exercises 55 • Outstate : Out-of-state tuition • Room.Board : Room and board costs • Books : Estimated book costs • Personal : Estimated personal spending • PhD : Percent of faculty with Ph.D.’s • Terminal : Percent of faculty with terminal degree • S.F.Ratio : Student/faculty ratio • perc.alumni : Percent of alumni who donate • Expend : Instructional expenditure per student • Grad.Rate : Graduation rate Before reading the data into R, it can be viewed in Excel or a text editor.

a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

**Solution:**

In this first we are trying to import our dataset.

Table

Description automatically generated with medium confidence

b) Look at the data using the View() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

> rownames (college) <- college[, 1] > View (college)

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

> college <- college[, -1] > View (college)

Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

**Solution:**

Here we used the head () to know the variables before modification and we removed one variable $X and by using the fix () we fix the college data.

Table

Description automatically generated

Then we finally got the modified dataset after removal of $X as

Table

Description automatically generated

c)

i) Use the summary() function to produce a numerical summary of the variables in the data set.

**Solution:**

summary () it is used to produce a numerical summary of variables in the collage dataset.

A screenshot of a computer

Description automatically generated with medium confidence

ii) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

**Solution:**

pairs () function to produce a scatterplot matrix of the first ten columns or variables of the college dataset.



A picture containing arrow

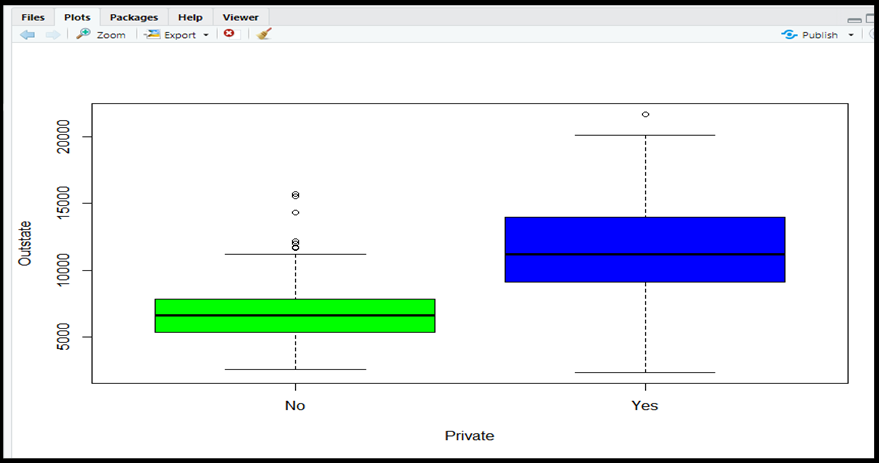
Description automatically generated

iii) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

**Solution:**

plot () function to produce side-by-side boxplots of Outstate versus Private.





iv) Create a new qualitative variable, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

> brilliant <- rep ("No", nrow (college))

> brilliant [college$Top10perc > 50] <- " Yes "

> brilliant <- as. factor (brilliant)

> data <- data.frame (data, brilliant)

Use the summary () function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

**Solution:**

Creating a quantitative variable brilliant, by starting the Top10perc and boxplot of outstate versus brilliant

Text

Description automatically generated

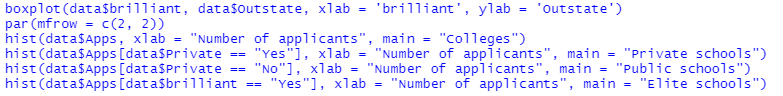
Chart, box and whisker chart

Description automatically generated

v) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow = c(2, 2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

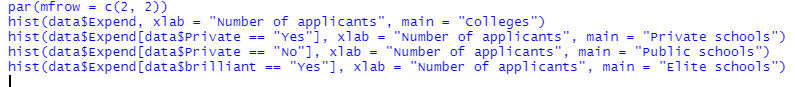
**Solution:**

hist() function to produce histogram with different variables



Chart, histogram

Description automatically generated



Chart, histogram

Description automatically generated

vi) Continue exploring the data and provide a brief summary of what you discover.

**Solution:**

Here I have seen the alumni frequency < 40 is

Text

Description automatically generated

A picture containing logo

Description automatically generated

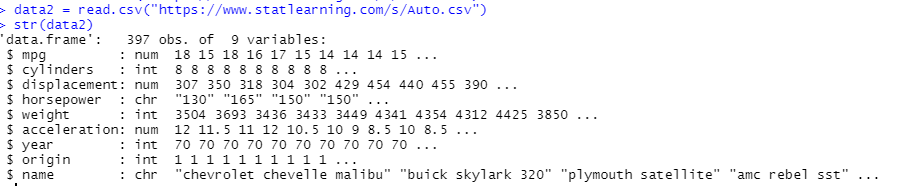
2) ISLR 2.4 Applied Problem 9.

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

a) Which of the predictors are quantitative, and which are qualitative?

**Solution:**

we are loading the Auto dataset and finding which are qualitative and quantitative.



b) What is the range of each quantitative predictor? You can answer this using the range() function.

**Solution:**

We are retrieving the range of each quantitative predictors by using range() function.

Text

Description automatically generated

c) What is the mean and standard deviation of each quantitative predictor?

**Solution:**

Text

Description automatically generated

Text

Description automatically generated

d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

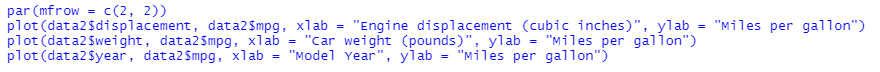
**Solution:**

Text

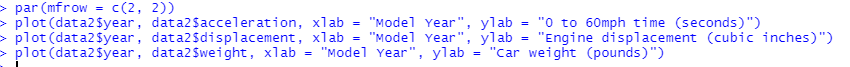
Description automatically generated

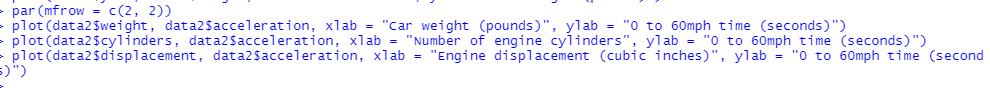
e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

**Solution:**

Chart

Description automatically generated

Diagram

Description automatically generatedGraphical user interface, application

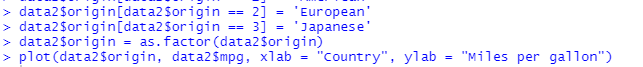
Description automatically generated

We appear to induce more mileage per gallon on a 4 cycle vehicle than the others. Weight, uprooting and drive appear to have a reverse impact with mpg. We see an by and large increment in mpg over a long time. Nearly multiplied in one decade. Japanese cars have higher mpg than US or European cars.

f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

**Solution:**

Based on the scatter plots I made in part 5 which relate miles per gallon to the predictor’s engine displacement, horsepower, car weight, and model year, it seems as if the first three factors would be most helpful in predicting mpg, with model year still potentially being helpful but less so. There are clear relationships that increasing engine displacement/horsepower/car weight results in decreased fuel efficiency. There is also a weak relationship that fuel efficiency generally increased going from 1970 to 1982.

Chart, box and whisker chart

Description automatically generated

Looking at the above box plot, we can also see that there is a relationship between a car's country of origin and fuel efficiency, where on average Japanese cars are the most efficient, followed by European cars and then by American cars.

3) ISLR 2.4 Applied Problem 10

This exercise involves the Boston housing data set.

a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library.

> library (ISLR2)

Now the data set is contained in the object Boston.

> Boston Read about the data set:

> ?Boston

How many rows are in this data set? How many columns? What do the rows and columns represent?

**Solution:**

Loading Dataset

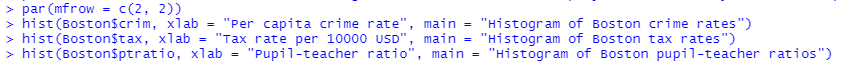
Text

Description automatically generated

b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

**Solution:**

Scatterplots of the predictors

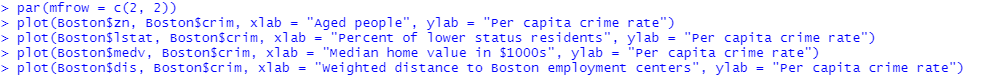


Chart, histogram

Description automatically generated

c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

**Solution:**

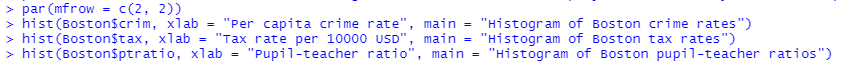


Graphical user interface

Description automatically generated

d) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

**Solution:**

****

Chart, histogram

Description automatically generated

e) How many of the census tracts in this data set bound the Charles River?

**Solution:**



f) What is the median pupil-teacher ratio among the towns in this data set?

**Solution:**

A picture containing text

Description automatically generated

g) Which census tract of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

**Solution:**

****

**A picture containing text

Description automatically generated**

In summary, these two tracts with the lowest median value of owner-occupied homes have predictors generally at the extreme ends of their respective ranges.

h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

**Solution:**

In this data set, there are 64 tracts which average more than seven rooms per dwelling, and 13 of those tracts which average more than 8 rooms per dwelling.

A picture containing logo

Description automatically generated

4) ISLR 3.7 Applied Problem 8.

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

a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary () function to print the results. Comment on the output. For example:

i) Is there a relationship between the predictor and the response?

**Solution:**

Logo

Description automatically generated with medium confidence

Text

Description automatically generated

Simple linear regression gives a model Y^=39.935861−0.157845X1Y^=39.935861−0.157845X1 between the predictor horsepower and the response mpg. A p-value of essentially zero for β^1=−0.157845β^1=−0.157845 gives very strong evidence that there is a relationship between mpg and horsepower Since R2=0.6059*R2=*0.6059, approximately 60.6% of the variability in mpg is explained by a linear regression onto horsepower. This is a modest relationship between the predictor and the response, since as discussed in the chapter we can improve our R2*R2* value to 0.688 by including a quadratic term. The value of β^1*β^1* itself indicates that in the model each increase of 1 horsepower results on average in a decrease of 0.157845 miles per gallon. In other words, in this model there is a negative relationship between the predictor and the response.

ii) How strong is the relationship between the predictor and the response?

**Solution:**

To calculate the leftover blunder relative to the reaction we utilize the cruel of the reaction and the RSE. The cruel of mpg is 23.4459184. The RSE of the lm.fit was 4.9057569 which demonstrates a rate mistake of 20.9237141%. We may moreover note that as the R2 is rise to 0.6059483, nearly 60.5948258% of the inconstancy in “mpg” can be clarified utilizing “horsepower”.

iii) Is the relationship between the predictor and the response positive or negative?

**Solution:**

As the coefficient of “horsepower” is negative, the relationship is additionally negative. The more drive a vehicle has the direct relapse shows the less mpg fuel proficiency the vehicle will have.

iv) What is the predicted mpg associated with a horsepower of 98? What are the associated 95 % confidence and prediction intervals?

**Solution:**

Logo

Description automatically generated with medium confidence

b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

**Solution:**





Chart, scatter chart

Description automatically generated

c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

**Solution:**

Background pattern

Description automatically generated with low confidence

Chart

Description automatically generatedLooking at the Residuals vs. Fitted plot, there is a clear U-shape to the residuals, which is a strong indicator of non-linearity in the data. This, when combined with an inspection of the plot in Part 2, tells us that the simple linear regression model is not a good fit. In addition, when looking at the Residuals vs. Leverage plot, there are some high leverage points (remember that after dropping the rows with null values, there are 392 observations in the data set, giving an average leverage value of 2/392≈0.00512/392≈0.0051) which also have high standardized residual values (greater than 2), which is also of concern for the simple linear regression model. There are also a number of observations with a standardized residual value of 3 or more, which is evidence to suggest that they would be possible outliers if we didn't already have the suspicion that the data is non-linear.

5) ISLR 3.7 Applied Problem 9.

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

Table

Description automatically generated

Note that the origin column actually contains categorical data, even though it is coded using integers. In order to make my life a little easier for performing regression, I'm going replace the values in that column with their meanings and convert it to a factor column. There are also other options for coding categorical variables, such as using the factor() function directly within lm(), or using the C() function to have more control over the contrast coding.

Table

Description automatically generated with medium confidence

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

**Solution:**



Diagram, engineering drawing

Description automatically generated

Since origin and name are categorical columns, I'm excluding them from the scatterplot matrix.

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

**Solution:**

Text, table

Description automatically generated

Since the origin column is also qualitative, I excluded it along with the name column when computing the matrix of correlations.

c) 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?

**Solution:**

Text

Description automatically generated

Text

Description automatically generated with medium confidence

ii) Which predictors appear to have a statistically significant relationship to the response?

**Solution:**

Able to reply this address by checking the p-values related with each predictor’s t-statistic. We may conclude that all indicators are measurably noteworthy but “cylinders”, “horsepower” and “acceleration”.

iii) What does the coefficient for the year variable suggest?

**Solution:**

The coefficient to the “year” variable recommends that the normal impact of an increment of 1 year is an increment of 0.7507727 in “mpg” (all other indicators remaining steady). In other words, cars gotten to be more fuel effective each year by nearly 1 mpg / year.

d) 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?

**Solution:**

A picture containing text

Description automatically generated

Graphical user interface, chart, scatter chart

Description automatically generated

Looking at the Residuals vs. Fitted plot, there appears to be moderate U-shape, which indicates that there might be non-linearity in the data. In addition, when looking at the Residuals vs. Leverage plot we can observe a few things. First, there are a number of observations with standardized residual values with absolute value greater than or equal to 3. Those are likely outliers. This is confirmed by looking at the Scale-Location plot, which has √|Standardized residual| as the *y*-axis. Points with √Standardized residual≥1.732 have |Standardized residual|≥3, which again means that they are likely outliers. Going back the Residuals vs. Leverage plot, we also see that there are a couple points with unusually high leverage. Again, remember that after dropping the rows with null values, there are 392 observations in the data set, giving an average leverage value of 9/392≈0.023*9/392≈0.023*. There is one point with a leverage value of about 0.10, which is almost 5 times greater than the average. There is another point with a leverage of about 0.20, which is almost 10 times greater than the average.

e) Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

**Solution:**

Table

Description automatically generated

Text

Description automatically generatedHere, the interaction term between horsepower and originEuropean has the highest p-value, so we remove the interaction between horsepower and origin from the model to proceed with backward selection. We're going to do this until the adjusted R2 value no longer increases when we remove predictors. The adjusted *R2* value is an adjustment of the regular R2 value to take into account the number of predictors and is a way to compare the fit of models with different numbers of predictors, as the regular *R2* always increases with the inclusion of more predictors.

f) Try a few different transformations of the variables, such as log(X), √ X, X2. Comment on your findings.

**Solution:**

To get a sense of which transformations I want to try out for each quantitative variable, focusing on displacement, horsepower and weight, I'll look at the scatterplots of each one versus mpg.

The book already explored nonlinear transformations of horsepower to predict mpg, so I will first look at transforms of acceleration.

A picture containing text

Description automatically generated

Chart, scatter chart

Description automatically generated

Timeline

Description automatically generated

Company name

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

It appears that there might be heteroscedasticity, or non-constant variances in the error terms, so let's first see how applying a logarithmic transportation affects the model.

A picture containing timeline

Description automatically generated

A picture containing logo

Description automatically generated

Chart, scatter chart

Description automatically generated

While the transformation did bump up the R2 value very slightly, it didn't really do anything to help with the residuals. This is probably due to the fact that two cars with the same 0 to 60 mile per hour time could be quite different in other ways that would affect fuel economy, such has differences in engine efficiency. For the remainder of the problem, let's turn our attention the the relationship between engine displacement and fuel efficiency. From the scatterplot, it is pretty clear that there is a nonlinear relationship between the two quantities. Let's start off by comparing a linear model to one that also includes the quadratic term.

Timeline

Description automatically generated

Timeline

Description automatically generated

Text, letter

Description automatically generated

As we can see, the quadratic term has a p-value of essentially zero, which is quite statistically significant. Moreover, the inclusion of the quadratic term improves the *R2* value from 0.6482 in the linear model to 0.6888. This, along with the above results of using the anova() function to compare the two models, strongly suggests that the model which includes the quadtratic term is a better fit than the model which does not include it. To finish, lets now compare the quadratic model to a quintic one.

Table

Description automatically generated

Text

Description automatically generated

First, we notice that none of the terms above order 2 (i.e. the cubic, quartic, and quintic terms) have statistically significant p-values. In addition, the adjusted R2*R2* value has dropped slightly from 0.6872 in the quadratic model to 0.6861. Lastly, p-value from the anova() function is 0.65, which means that there is not sufficient evidence to reject the null hypothesis that the quintic model is a better fit than the quadratic one. These three pieces of evidence suggest that including terms beyond order 2 does not improve the model.

6) ISLR 3.7 Applied Problem 10.

This question should be answered using the Carseats data set.

Table

Description automatically generated

a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

**Solution:**

Timeline

Description automatically generated

b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

**Solution:**

The coefficient of -0.054459 for Price means that, for a given location (i.e. fixed values of Urban and US), increasing the price of a car seat by $1 results in a decrease of sales by approximately 54.46 units, on average, in the model. The coefficient of -0.021916 for UrbanYes means that, for a given carseat price point and value of US, the model predicts urban areas to have approximately 22 fewer carseat sales on average compared to non-urban areas. The coefficient of 1.200573 for USYes means that, for a given car seat price point and value of Urban, the model predicts that stores in the United States have 1201 more carseat sales on average than stores outside the United States.

c) Write out the model in equation form, being careful to handle the qualitative variables properly.

**Solution:**

The model has the following equation.



Here, *y^* is the estimated carseat sales, in thousands of car seats; x1j*x1j* is the price of the carseat at the jth store, in dollars; and *x2j* and *x3j* are dummy variables to represent whether or not the jth store at is located in an urban area and in the United States, respectively. More concretely, *x2j* and *x3j* use the following coding scheme.

A screenshot of a computer

Description automatically generated with medium confidence

d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?

**Solution:**

The p-values for the intercept, Price, and US Yes are all essentially zero, which provides strong evidence to reject the null hypothesis *H0:βj=0* for those predictors. The p-value for UrbanYes, however, is 0.936, so there is no evidence to reject the null hypothesis that it has a non-zero coefficient in the true relationship between the predictors and Sales.

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.

**Solution:**

Timeline

Description automatically generated

f) How well do the models in (a) and (e) fit the data?

**Solution:**

A picture containing company name

Description automatically generated

Chart, scatter chart

Description automatically generated

A picture containing diagram

Description automatically generated

Chart, scatter chart

Description automatically generated

The models in Part 1 and Part 5 both fit the data about equally well, with identical R2 values of 0.2393. In addition, when comparing the diagnostic plots between the two models, there isn't any discernable visual differences that would strongly indicate that one model is a better fit than the other.

g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).

**Solution:**

Text, letter

Description automatically generated

h) Is there evidence of outliers or high leverage observations in the model from (e)?

**Solution:**

When we look at the residuals vs. leverage plot for the model from Part 5 that I generated in Part 6, we see that there are a number of observations with standardized residuals close to 3 in absolute value. Those observations are possible outliers. We can also see in the same plot that there are number of high leverage points with leverage values greatly exceeding the average leverage of 3/400=0.0075, though those high leverage observations are not likely outliers, as they have studentized residual values with absolute value less than 2.

Group Members:

Kasireddy Sreekar Reddy – [skmvk@umsystem.edu](mailto:skmvk@umsystem.edu)

Mangalaram Dheeraj Reddy - [dm6dx@umsystem.edu](mailto:dm6dx@umsystem.edu)