Homework 2

Claudius Taylor, Junpu Zhao, Thomas Wilson

September 16, 2018

library(MASS)  
library(data.table)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# STAT 5310 – Regression Analysis

# Lab Exercise/Homework 2

## 1.

Develop a regression model to predict the average number of hits (Y) from the average salary for the 10 highest paid hitters from each MLB team (X) based on data from the 2013 season.

The data is in the file Hitters

hitters <- fread('https://raw.githubusercontent.com/wilsonify/AppliedRegression/master/data/Hitters.csv')

names(hitters) <- names(hitters) %>% gsub('[ |(|)|$]','',.)

hitters %>% head()

## Team AverageNumberofHits AverageSalary2013M NewYorkYankees  
## 1: Marlins 41 1.340 0  
## 2: Mariners 51 3.290 0  
## 3: Astros 57 0.759 0  
## 4: Mets 61 2.680 0  
## 5: Pirates 61 3.090 0  
## 6: Padres 59 3.370 0

### a)

Fit a straight line regression model between Y and X.

Decide whether , the coefficient of X, is statistically significantly different from zero or not.

Give a reason to support your answer.

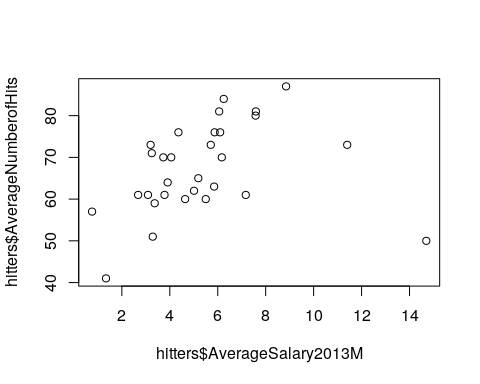
fit <- lm(hitters$AverageNumberofHits ~ hitters$AverageSalary2013M)  
summary\_of\_fit <- summary(fit)  
summary\_of\_fit$coefficients["hitters$AverageSalary2013M","Pr(>|t|)"]

## [1] 0.1328928

is not significatnly different from zero.

Under the null hypothesis that , we would observe a more extreme 13% of the time.

plot( y = hitters$AverageNumberofHits  
 ,x = hitters$AverageSalary2013M  
 )



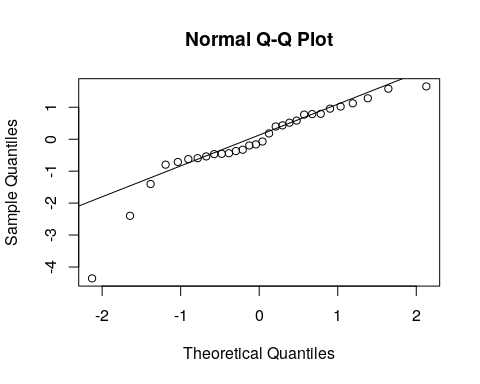
### b)

Examine the studentized residuals for the fitted model in part a.

Identify any points which are outliers (that is, have studentized residuals values more extreme than 3.

studentized\_residuals <- studres(fit)

qqnorm(studentized\_residuals)  
qqline(studentized\_residuals)



There is exactly one outlier with studentized residual of -4.36

## 2.

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands)

The data is in the file Newspaper.

*(Source: Gale Directory of Publications, 1994).*

newspaper <- fread('https://raw.githubusercontent.com/wilsonify/AppliedRegression/master/data/Newspaper.csv')

newspaper %>% head()

## Newspaper Daily Sunday  
## 1: Baltimore Sun 391.952 488.506  
## 2: Boston Globe 516.981 798.298  
## 3: Boston Herald 355.628 235.084  
## 4: Charlotte Observer 238.555 299.451  
## 5: Chicago Sun Times 537.780 559.093  
## 6: Chicago Tribune 733.775 1133.249

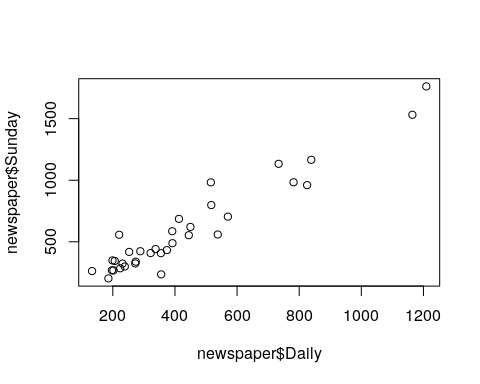
### a)

Construct a scatter plot of Sunday circulation versus daily circulation.

Does the plot suggest a linear relationship between daily and Sunday circulation?

Do you think this is a plausible relationship?

plot(y=newspaper$Sunday,x=newspaper$Daily)



The plot of sunday circulation vs daily circulation does suggest a linear relationship.

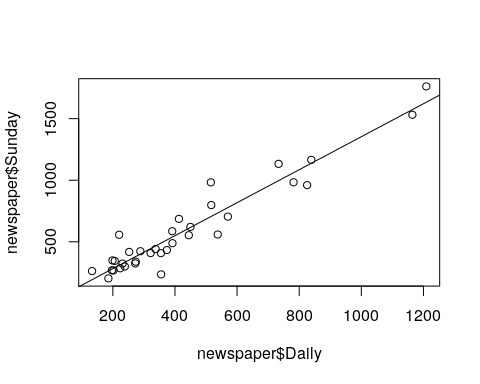
It seems very plausible that, in reality, newspapers with a large circulation on weekdays would also have a large circulation on sunday and vise-versa for newspapers with a small circulation.

### b)

Fit a regression line predicting Sunday circulation from daily circulation.

fit <- lm(data =newspaper, Sunday~Daily)

plot(y=newspaper$Sunday,x=newspaper$Daily)  
abline(fit)



### c)

Obtain the 95% confidence intervals for β0 and β1.

summary\_of\_fit <- summary(fit)  
summary\_of\_fit$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 13.835630 35.80400579 0.3864269 7.017382e-01  
## Daily 1.339715 0.07075395 18.9348403 6.016802e-19

alpha = 1.0 - 0.95  
half\_alpha = alpha / 2.0  
z = qnorm(1-half\_alpha)  
  
  
beta0\_center <- summary\_of\_fit$coefficients["(Intercept)","Estimate"]  
beta0\_SE <- summary\_of\_fit$coefficients["(Intercept)","Std. Error"]  
beta0\_error\_bound <- z \* beta0\_SE  
beta0\_CI <- c(beta0\_center-beta0\_error\_bound,beta0\_center+beta0\_error\_bound) %>% round(2)  
print("95% confidence interval for the intercept, B0 is ")

## [1] "95% confidence interval for the intercept, B0 is "

beta0\_CI

## [1] -56.34 84.01

beta1\_center <- summary\_of\_fit$coefficients["Daily","Estimate"]  
beta1\_SE <- summary\_of\_fit$coefficients["Daily","Std. Error"]  
beta1\_error\_bound <- z \* beta1\_SE  
beta1\_CI <- c(beta1\_center-beta1\_error\_bound,beta1\_center+beta1\_error\_bound) %>% round(2)  
print("95% confidence interval for the slope, B1 is ")

## [1] "95% confidence interval for the slope, B1 is "

beta1\_CI

## [1] 1.20 1.48

### d)

Is there a significant relationship between Sunday circulation and daily circulation?

Justify your answer by a statistical test.

Indicate what hypothesis you are testing and your conclusion.

There a significant relationship between Sunday circulation and daily circulation.

If there were no relationship (H0: ), we would expect a more extreme coefficient only 0.00000000000000006% of the time.

### e)

What proportion of the variability in Sunday circulation is accounted for by daily circulation?

summary\_of\_fit$adj.r.squared

## [1] 0.9154991

91.5% of the variability in sunday ciculation can be explained by a linear relationship between daily circulation and Sunday circulation.

### f)

Provide an estimate (based on a 95% confidence level) for the average Sunday circulation of newspapers with daily circulation of 500,000.

alpha = 1.0 - 0.95  
half\_alpha = alpha / 2.0  
z = qnorm(1-half\_alpha)  
newdata <- data.frame(Daily=c(500000))  
center <- predict(fit, newdata = newdata)  
sigma <- summary\_of\_fit$sigma  
error\_bound <- z \* sigma  
CI <- c(center-error\_bound,center+error\_bound)  
CI

## 1 1   
## 669656.8 670085.7

At a 95% confidence level, the average Sunday circulation of all newspapers with daily circulation of 500,000 will be between 669700 and 670100.

### g)

The particular newspaper that is considering a Sunday edition has a daily circulation of 500,000.

Provide an interval estimate (based on a 95% confidence level) for the predicted Sunday circulation of this paper.

How does this interval differ from that given in part f?

predict(fit, newdata = data.frame(Daily=c(500000)), interval="predict",level = 0.95)

## fit lwr upr  
## 1 669871.2 597872.4 741870

At a 95% confidence level, the Sunday circulation of a particular newspaper with daily circulation of 500,000 will be between 598000 and 742000.

Generally, there is more (tighter) confidence in predicting the average of all newspapers than in predicting the circulation of a particular newspaper.

### h)

Another newspaper considering a Sunday edition has a daily circulation of 2,000,000.

Provide an interval estimate for the predicted Sunday circulation for this paper?

How does this interval compare with the one given in part g?

Do you think it is likely to be accurate?

predict(fit, newdata = data.frame(Daily=c(2000000)), interval="predict",level = 0.95)

## fit lwr upr  
## 1 2679443 2391263 2967623

At a 95% confidence level, the Sunday circulation of a particular newspaper with daily circulation of 200,000,000 will be between 2,400,000 and 2,970,000.

This is a wider interval than predicting at 500,000 in part g.

Generally, the further from the average, less confidence in predicting.

## 3.

Consider the sale price of 1672 residential properties in Ames, Iowa from 2006 to 2010.

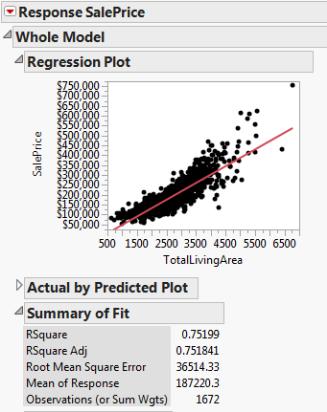
The data included information from the Tax Assessor’s Office along with information on various aspects of each property.

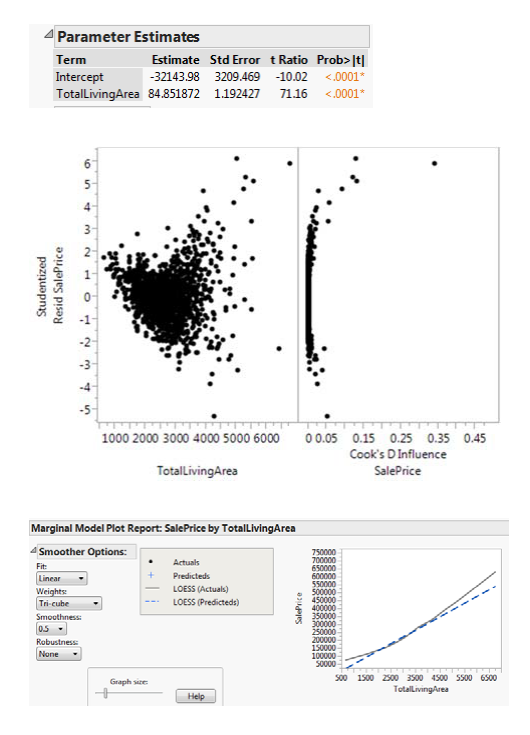
The first model fit to the data was

$ SalePrice = \_0 + \_1 TotalLivingArea + e $

Where TotalLivingArea gives the total area in square feet of the above ground living area and basement for each property.

The following is some output from fitting the first model as well as some plots.





1. Interpret the estimated coefficient of TotalLivingArea in model (1).

The estimated coefficent of is 84.85. Meaning that every unit change in TotalLivingArea is associated with an $84.85 increase in SalePrice.

1. Carefully describe at least one main shortcomings (apart from outliers) evident in model (1).

The residuals are not identically distributed. They tend to get larger as the TotalLivingArea gets larger. It seems that the relationship is more complex than linear.

## 4.

Consider the situation in which 3 regression models of the following form have been fit to a set of data.

Suppose that we are modeling the sale prices of apartments in Austin, TX.

Model (A) is such that y = sale price, x = number of bedrooms

model (B) is such that y = sale price, x = number of bathrooms

model (C) is such that y = log(sale price), x = number of bedrooms

Comparing the models, which one of the following statements is true?

I. It is reasonable to compare the R2 values for models (A), (B) and (C).

1. It is reasonable to compare the R2 values only for models (A) and (B
2. It is reasonable to compare the R2 values only for models (A) and (C).
3. It is not reasonable to compare the R2 values for models (A), (B), and (C).

Give a reason to support your answer.

Because log(sales price) is a one-to-one mapping to sales price itsself and is a proportion relative to the total sum of squares, statement I is true: It is reasonable to compare the values for models (A), (B) and (C).