IST772

Parin Patel
Homework 8
11/25/2019

Attribution Statement:

1. Homework 8 by Parin Patel: I did this homework by myself, with help from the book and the professor. In addition I used the following websites to help with the VIF #7:
    a. http://web.unbc.ca/~michael/courses/stats/lectures/VIF%20articlea.pdf

Exercises:

1. **The data sets package in R contains a small data set called mtcars that contains *n* = 32 observations of the characteristics of different automobiles. Create a new data frame from part of this data set using this command: myCars <- data.frame(mtcars[,1:6]).**

```
> ##Question 1
> myCars<- data.frame(mtcars[,1:6])
```

2. **Create and interpret a bivariate correlation matrix using cor(myCars) keeping in mind the idea that you will be trying to predict the mpg variable. Which other variable might be the single best predictor of mpg?**

The bivariate correlation matrix results show that weight is the variable that might be the best predictor for mpg. This is because it has the strongest inverse correlation (-0.87) to MPG.

IST772

Parin Patel
Homework 8
11/25/2019

```
> ##Question 2
> cor(myCars)
            mpg        cyl       disp         hp
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684
cyl  -0.8521620  1.0000000  0.9020329  0.8324475
disp -0.8475514  0.9020329  1.0000000  0.7909486
hp   -0.7761684  0.8324475  0.7909486  1.0000000
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591
wt   -0.8676594  0.7824958  0.8879799  0.6587479
           drat         wt
mpg   0.6811719 -0.8676594
cyl  -0.6999381  0.7824958
disp -0.7102139  0.8879799
hp   -0.4487591  0.6587479
drat  1.0000000 -0.7124406
wt   -0.7124406  1.0000000
```

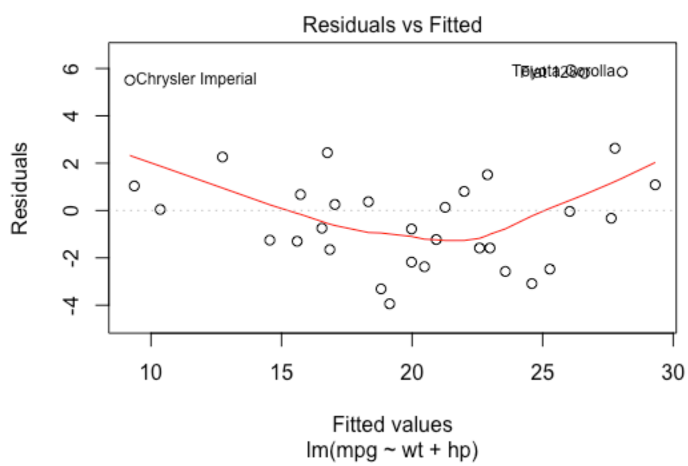
3. **Run a multiple regression analysis on the myCars data with lm(), using mpg as the dependent variable and wt (weight) and hp (horsepower) as the predictors. Make sure to say whether or not the overall *R*-squared was significant. If it was significant, report the value and say in your own words whether it seems like a strong result or not. Review the significance tests on the coefficients (B-weights). For each one that was significant, report its value and say in your own words whether it seems like a strong result or not.**

   The multiple regression analysis on the myCars data with the lm() procedure resulted in a R-squared value of 0.83. Since a R-squared of 1 indicates that the dependent variable is a perfectly predicted by the set of predictors (independent variables). Therefore, we can say that our R-square result of 0.83 is a strong predictor for our data. In addition, the null hypothesis test on the R-squared has F(2,29) = 69.21, p < 0.001 (p = 9.100 e-12), meaning we reject the null hypothesis. Therefore, based on our p-value and R-squared value, we can say we have a strong result with significant variables. A review of the coefficient's significance tests (B-weights) shows that both wt and hp are significant because their v-values are less than the conventional alpha of p <0.05. Therefore, in both these cases we reject the null. The "Estimate" column shows the B-weights for each predictor. Based on this , we can see negative weights for both wt and hp and that sine the weights are not similar, the predictor variables will not have roughly the same contribution to the dependent variable.

IST772

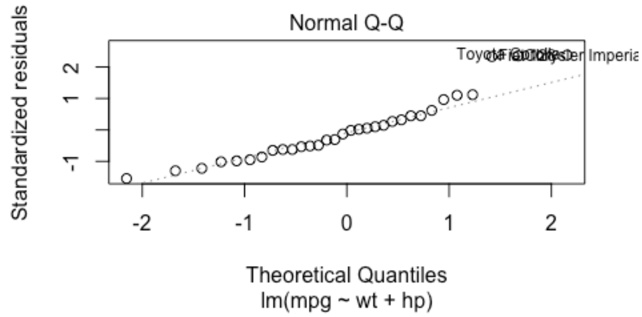
Parin Patel
Homework 8
11/25/2019

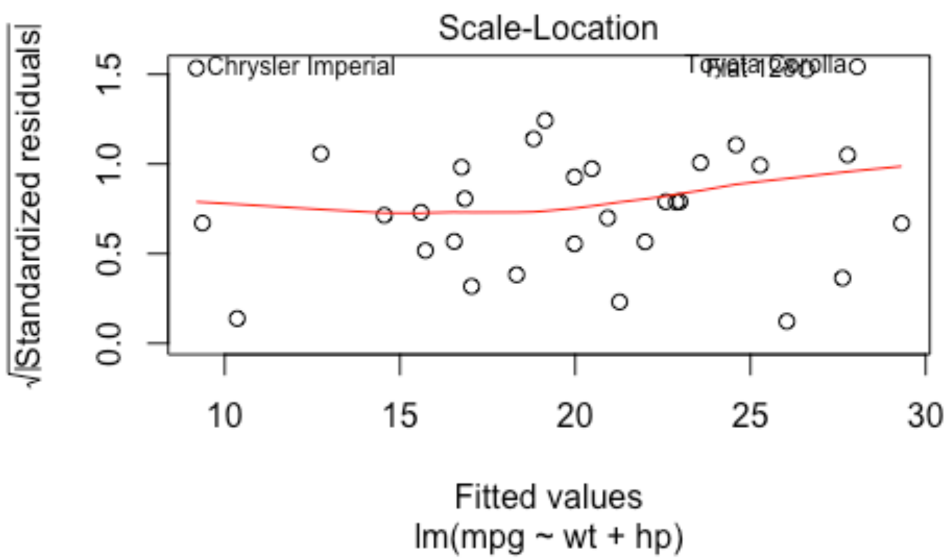
```
> ##Question 3:
> lmResults <- lm(mpg~wt+hp, data = myCars)
> summary(lmResults)

Call:
lm(formula = mpg ~ wt + hp, data = myCars)

Residuals:
   Min     1Q Median     3Q    Max 
-3.941 -1.600 -0.182  1.050  5.854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,	Adjusted R-squared:  0.8148 
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```



IST772

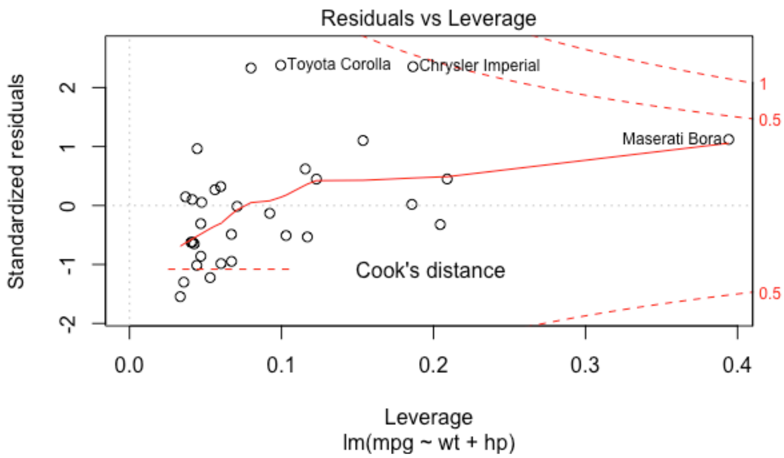
Parin Patel
Homework 8
11/25/2019





IST772

Parin Patel
Homework 8
11/25/2019



4. **Using the results of the analysis from Exercise 2, construct a prediction equation for mpg using all three of the coefficients from the analysis (the intercept along with the two B-weights). Pretend that an automobile designer has asked you to predict the mpg for a car with 110 horsepower and a weight of 3 tons. Show your calculation and the resulting value of mpg.**

   You can see from the results below that the if car wt = 3 tons with a hp = 110, then the project mpg is 22.1.

```
Hit <Return> to see next plot: plot(lmResults)
> #Question 4:
> PredEq1 <- coefficients(lmResults)
> PredEq1
(Intercept)          wt          hp
37.22727012 -3.87783074 -0.03177295
> #Question 4:
> PredMPG <- coefficients(lmResults)
> PredMPG
(Intercept)          wt          hp
37.22727012 -3.87783074 -0.03177295
> weight = 3
> hp = 110
> newMPGpred <- PredMPG[1] + (PredMPG[2] * weight) + (PredMPG[3]  * hp)
> newMPGpred
(Intercept)
   22.09875
```

IST772

Parin Patel
Homework 8
11/25/2019

```
> finalCalc4 <- data.frame(wt = 3, hp = 110)
> predict(lmResults,finalCalc4)
       1
22.09875
>
```

5. **Run a multiple regression analysis on the myCars data with lmBF(), using mpg as the dependent variable and wt (weight) and hp (horsepower) as the predictors. Interpret the resulting Bayes factor in terms of the odds in favor of the alternative hypothesis. If you did Exercise 2, do these results strengthen or weaken your conclusions?**

   The results of the Bayes factor do strengthen the conclusions found in Exercise 2. The results yielded a odds ratio of 788547604: 0. This suggests that the weights and horsepower model would be much stronger at making a prediction of mpg than using an analysis of only the intercept.

```
> ##Question 5:
> library("BayesFactor")
> lmBFout <- lmBF(mpg~wt+hp, data = myCars)
> summary(lmBFout)
Bayes factor analysis
--------------
[1] wt + hp : 788547604 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

6. **Run lmBF() with the same model as for Exercise 4, but with the options posterior=TRUE and iterations=10000. Interpret the resulting information about the coefficients.**

   By looking at the boundaries of our predictors, we are able to evaluate the significance of the values. Therefore, the results of our lmBF() model with the options for posterior = TRUE and 10000 iterations shows first that none of the values overlap with 0. Therefore, we can say that these coefficients are significant predictors for mpg prediction. Additionally, the mean of each variable and the 50% quantile for each variable are about the same as the coefficient value of the frequentist approach.

IST772

Parin Patel
Homework 8
11/25/2019

```
> ##Question 6:
> library("BayesFactor")
> lmBFout2 <- lmBF(mpg~wt+hp, data = myCars, posterior = TRUE, iterations = 10000)
0
      %
|----|----|----|----|----|----|----|----|----|----|
**************************************************|
> summary(lmBFout2)

Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

         Mean        SD  Naive SE Time-series SE
mu   20.09260  0.490045 4.900e-03      4.900e-03
wt   -3.78632  0.668891 6.689e-03      6.886e-03
hp   -0.03091  0.009423 9.423e-05      9.423e-05
sig2  7.48505  2.326105 2.326e-02      2.802e-02
g     3.84716 12.251830 1.225e-01      1.225e-01

2. Quantiles for each variable:

         2.5%      25%      50%      75%   97.5%
mu   19.14192 19.77275 20.09228 20.41045 21.0434
wt   -5.10618 -4.23013 -3.78991 -3.35184 -2.4513
hp   -0.04944 -0.03717 -0.03087 -0.02461 -0.0127
sig2  4.37981  5.97981  7.12569  8.57279 12.5562
g     0.35685  0.94655  1.71100  3.46243 18.4419
```

7. **Run install.packages() and library() for the "car" package. The car package is "companion to applied regression" rather than more data about automobiles. Read the help file for the vif() procedure and then look up more information online about how to interpret the results. Then write down in your own words a "rule of thumb" for interpreting vif.**

```
> ##Question 7:
> library(car)
> help("vif")
>
```

Vif() procedure stands for "Variance Inflation Factors", and detects multicollinearitis between the indepment variables of a model. It is a numerical value of the percentage of variance (standard error squard) for each coefficient. The value will begin at 1, and range upwards. The greater the value from one, the more correlated. The rule of thumb for interpreting Vif is that if the result is 1, then there is no correlation. If the value is between 1 and 5, there is somewhat correlation.. Finally, if it is greater than 5, then there is a high correlation. However, the rule of 10 associated with VIF indicates that the correlation is very high and less reliable. A VIF over 10 is very concerning. In general, a vif between 5-10, indicates a model with a low level of multicollinearity. Note that a VIF<3 would have low multicollinearity.

8. **Run vif() on the results of the model from Exercise 2. Interpret the results. Then run a model that predicts mpg from all five of the predictors in myCars. Run vif() on those results and interpret what you find.**

First, when I run vif() on the results from question 3 (lmResults), we get a vif of 1.77 for wt and hp. Since these vif values are below 5 and above 1, we can say there is a slight correlation, therefore, , the predictors in our model have a low multicollinearity. Secondly, when we run the vif() procedure on our predictors in myCars, we can see that disp and cycl have a high, unreliable correlation (VIC = 7.87 for cyl and 10.46 for disp). Therefore, we would suggest these predictors be removed from the model to increase effectiveness.

```
> ##Question 8:
> vif(lmResults)
      wt       hp
1.766625 1.766625
> vif(lm(mpg~cyl+disp+hp+drat+wt, data = myCars))
      cyl      disp        hp      drat        wt
 7.869010 10.463957  3.990380  2.662298  5.168795
>
```

Appendix A: Final Script:

```
set.seed(321)

##Question 1
myCars<- data.frame(mtcars[,1:6])

##Question 2
cor(myCars)


##Question 3:
lmResults <- lm(mpg~wt+hp, data = myCars)
summary(lmResults)

plot(lmResults)

#Question 4:
```

```
PredMPG <- coefficients(lmResults)
PredMPG

weight = 3
hp = 110

newMPGpred <- PredMPG[1] + (PredMPG[2] * weight) + (PredMPG[3]  * hp)
newMPGpred

finalCalc4 <- data.frame(wt = 3, hp = 110)
predict(lmResults,finalCalc4)


##Question 5:
library("BayesFactor")

lmBFout <- lmBF(mpg~wt+hp, data = myCars)
summary(lmBFout)


##Question 6:
library("BayesFactor")
lmBFout2 <- lmBF(mpg~wt+hp, data = myCars, posterior = TRUE, iterations =
10000)
summary(lmBFout2)

##Question 7:
library(car)
help("vif")

##Question 8:
vif(lmResults)
vif(lm(mpg~cyl+disp+hp+drat+wt, data = myCars))
```

Appendix B: Final Output:

```
> set.seed(321)
>
> ##Question 1
> myCars<- data.frame(mtcars[,1:6])
>
```

IST772

Parin Patel
Homework 8
11/25/2019

```
> ##Question 2
> cor(myCars)
            mpg        cyl       disp         hp       drat
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.7102139
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.0000000
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406
             wt
mpg  -0.8676594
cyl   0.7824958
disp  0.8879799
hp    0.6587479
drat -0.7124406
wt    1.0000000
>
>
> ##Question 3:
> lmResults <- lm(mpg~wt+hp, data = myCars)
> summary(lmResults)

Call:
lm(formula = mpg ~ wt + hp, data = myCars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,  Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12

>
> plot(lmResults)
Hit <Return> to see next plot:
```

IST772

Parin Patel
Homework 8
11/25/2019

```
Hit <Return> to see next plot: #Question 4:
Hit <Return> to see next plot: PredMPG <- coefficients(lmResults)
Hit <Return> to see next plot: PredMPG
>
> weight = 3
> hp = 110
>
> newMPGpred <- PredMPG[1] + (PredMPG[2] * weight) + (PredMPG[3]  * hp)
> newMPGpred
(Intercept)
   22.09875
>
> finalCalc4 <- data.frame(wt = 3, hp = 110)
> predict(lmResults,finalCalc4)
       1
22.09875
>
>
> ##Question 5:
> library("BayesFactor")
>
> lmBFout <- lmBF(mpg~wt+hp, data = myCars)
> summary(lmBFout)
Bayes factor analysis
--------------
[1] wt + hp : 788547604 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS

>
>
> ##Question 6:
> library("BayesFactor")
> lmBFout2 <- lmBF(mpg~wt+hp, data = myCars, posterior = TRUE, iterations = 
10000)
0                                                   %
|----|----|----|----|----|----|----|----|----|----|
**************************************************|
> summary(lmBFout2)

Iterations = 1:10000
```

IST772

Parin Patel
Homework 8
11/25/2019

```
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

         Mean       SD  Naive SE Time-series SE
mu   20.08860  0.48567 0.0048567      4.439e-03
wt   -3.79068  0.67027 0.0067027      6.703e-03
hp   -0.03097  0.00947 0.0000947      9.418e-05
sig2  7.48789  2.17464 0.0217464      2.692e-02
g     3.95125 11.94936 0.1194936      1.195e-01

2. Quantiles for each variable:

         2.5%      25%      50%      75%   97.5%
mu   19.14165 19.77214 20.08624 20.40866 21.0434
wt   -5.11696 -4.22717 -3.80000 -3.35349 -2.4618
hp   -0.04992 -0.03718 -0.03098 -0.02474 -0.0123
sig2  4.35390  5.97118  7.12300  8.62566 12.6598
g     0.36464  0.94529  1.72806  3.49766 20.2555

>
> ##Question 7:
> library(car)
> help("vif")
>
> ##Question 8:
> vif(lmResults)
      wt       hp
1.766625 1.766625
> vif(lm(mpg~cyl+disp+hp+drat+wt, data = myCars))
      cyl      disp        hp      drat        wt
 7.869010 10.463957  3.990380  2.662298  5.168795
```