Kwasi Mensah

2022-11-30

Author: Kwasi Mensah Homwork_Number: 8 Output: pdf_document Attribution statement:

I did the homework by myself, with help from the book and the professor and the following sources: https://www.r-tutor.com/elementary-statistics/multiple-linear-regression/estimated-multiple-regression-equation https://www.displayr.com/variance-inflation-factors-vifs/

#R Markdown #Run these three functions to get a clean test of homework code

```
dev.off() #Clear the graph window

## null device
##           1

cat('\014') #Clear the console

rm(list = ls()) #Clear user objects from the environment
```

#R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: ###

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]).

```
?mtcars

## starting httpd help server ... done

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

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?

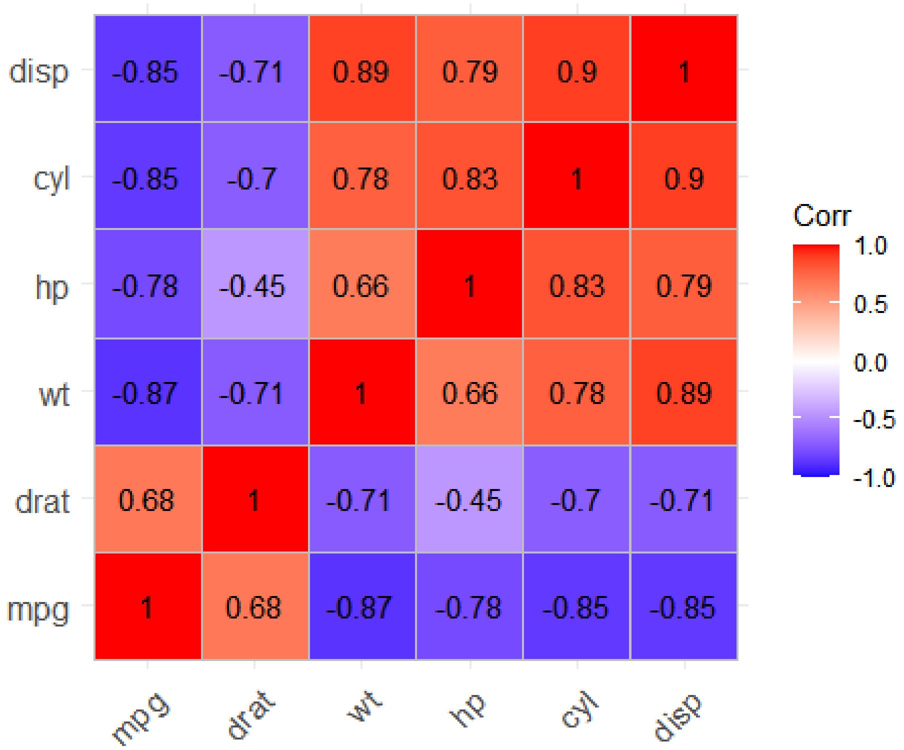
```
cor1 <- cor(myCars)
cor1
```

```
##             mpg        cyl       disp         hp       drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.7102139  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.0000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
```

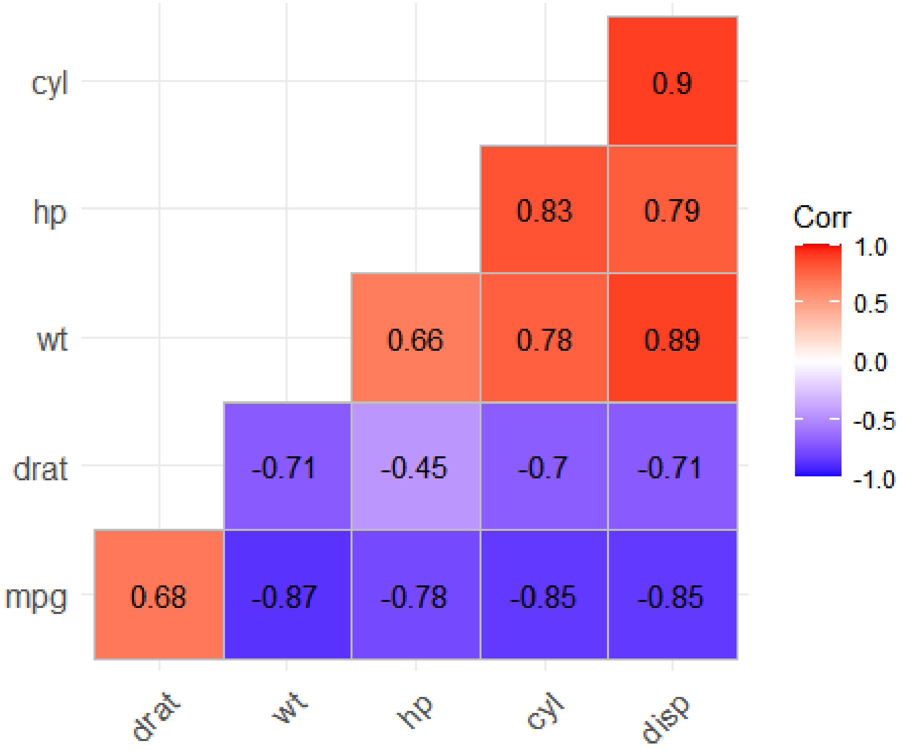
```
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
ggcorrplot(cor1,hc.order = TRUE,
           lab = TRUE)
```



```
ggcorrplot(cor1,hc.order = TRUE,
           type = "lower",
           lab = TRUE)
```



```
# The correlation matrix shows that the wt, cyl, and disp variables have a
high correlation to the mpg variable. The wt (weight) variable has the
highest correlation to mpg being -0.8676594, which is very close to -1.
```

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.

```
reg1 <- lm(mpg ~ wt + hp, myCars)
summary(reg1)

##
## 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

# The F-statistic is 69.21 with a p-value of 9.109e-12. The p-value is below
0.05 so our regression analysis is statistically significant, and this shows
us that there is a relationship between the mpg,weight, and hp variables, so
we will decide to reject the null hypothesis. Thus the results of the
analysis would support an alternative hypothesis. The intercept has a value
of 37.2273, the t-value has a  value of 23.285 and a p-value that is very
close to 0 being < 2e-16. The Multiple R-squared is 0.8268 and the Adjusted
R-squared is 0.8148, which show us the percentage of the weight & hp
variables accounted for the variability in mpg.

#The coefficients are the summary statistics for the performance of the whole
model. The DF is 29,where 2 DF is lost calculating the slope of the
coefficients, and 1 DF is lost calculating the intercept, which is supported
in our equation. The weight coefficient of -3.8778 per increase of 1000 lbs,a
t-value of -6.129 and a p-value of 1.12e-06. The hp coefficient of -.0318 per
incremental increase of Gross horsepower , a t value of -3.519 and a p-value
of 0.0015. Both of the p-values for both variables are below .05, thus
showing that the two variables are directly correlated to the mpg variable.
```

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.

```
w1<-data.frame(wt=6,hp=110)
predict(reg1,w1)

##        1
## 10.46526

mpgPE <- ((37.22727*1) + (-3.87783 * 6) + (-0.03177 * 110))
c(mpgPE)

## [1] 10.46559

# The car would get about 10.465 mpg, which is very bad especially for a new
vehicle.
```

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?

```
library(BayesFactor)

## Warning: package 'BayesFactor' was built under R version 4.2.2

## Loading required package: coda

## Loading required package: Matrix

## ************
## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact
Richard Morey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************

regbf1 <- lmBF(mpg ~ wt + hp, data=myCars)
summary(regbf1)

## Bayes factor analysis
## --------------
## [1] wt + hp : 788547604 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS

# With the Bayes approach to linear regression we get a Bayes Factor of
788547604 which is which is well over the odds cut off of 3:1. This analysis
results support our alternative hypothesis, and we will as a result reject
the null hypothesis. This lines up with our liner regression as our analysis
had a p-value of 9.109e-12, which is lower than .05, thus we will reject the
null hypothesis as well, in support with our alternative hypothesis.
```

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.

```
regbf2 <- lmBF(mpg ~ wt + hp, data=myCars,posterior=TRUE, iterations=10000)
summary(regbf2)

##
## 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   20.09317  0.498584 4.986e-03      4.953e-03
## wt   -3.79012  0.668899 6.689e-03      6.689e-03
## hp   -0.03081  0.009494 9.494e-05      9.494e-05
## sig2  7.52551  2.176727 2.177e-02      2.583e-02
## g     4.07281 24.861724 2.486e-01      2.548e-01
##
## 2. Quantiles for each variable:
##
##          2.5%      25%      50%      75%    97.5%
## mu   19.10610 19.76703 20.09472 20.42304 21.05684
## wt   -5.09581 -4.22937 -3.78208 -3.34911 -2.47196
## hp   -0.04933 -0.03712 -0.03092 -0.02466 -0.01188
## sig2  4.38891  5.98860  7.18404  8.64680 12.67373
## g     0.35825  0.91626  1.67566  3.38431 18.36229

# With the Bayes approach to linear regression the 95% HDI ranges for the
wt(Weight (1000 lbs)) predictor is -5.1071 to -2.47116  with a mean of -
3.78401 which is very close to the coefficient results from the linear
regression of -3.87783. With the Bayes approach to linear regression the 95%
HDI ranges for the hp(Gross horsepower) predictor is -0.0497 to -0.01261,
with a mean of-0.03105 which is very close to the coefficient results from
the linear regression of -0.03177. Since none of the interval pass 0 our
results are statistically significant, and we can reject the null hypothesis,
in support of the alternative hypothesis.
```

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.

```
library(car)

## Loading required package: carData

?vif()
# Variance Inflation Factors calculates variance-inflation and generalized
variance-inflation factors (VIFs and GVIFs) for linear, generalized linear,
and other regression models. A VIF can be computed for each predictor in a
predictive model.


#We could use the vif() function as a diagnostic for
multicollinearity(situations where the independent variable are so highly
correlated with one another that the analysis results are possible
inaccurate.)  a "Rule of Thumb" we could use when interpreting vif() is that
a value of 1 means that the predictor variable is not correlated with other
variables. The higher the value, the greater the correlation of the variable
with other variables.
#Values of more than 4 or 5 are sometimes regarded as being moderate to high,
with values of 10 or more being regarded as very high. If one variable has a
```

```
high VIF it means that other variables will have high VIF values as well.
Thus the variables will be highly correlated, and each will have the same VIF
value.
```

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.

```
reg1 <- lm(mpg ~ wt + hp, myCars)
vif(reg1)

##       wt       hp
## 1.766625 1.766625

reg2 <- lm(mpg ~ ., myCars)
vif(reg2)

##       cyl      disp        hp      drat        wt
##  7.869010 10.463957  3.990380  2.662298  5.168795

 # The VIF value of reg1 predictors both have the same value, and are both
above the value of 1 being 1.766625 meaning that the predictors are
correlated with the dependent variable.
 # The VIF value of reg2 predictors all have different values but are all
above 1. with disp (Displacement (cu.in.)) value being 10.463957. The cyl
(Number of cylinders) and wt variable values were also above 5, cyl was
7.869010 & wt was 5.168795. The hp variable and drat (Rear axle ratio) are
also above 1 with hp VIF value of 3.990380 and drat VIF value of 2.662298,
thus meaning that the predictors are correlated with the dependent variable.
```