Homework 8

Daniel Caley

8/28/2021

Contents

The homework for week 8 is exercises 1-8 on pages 181-182.

Question 1: MyCars DataFrame

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])`.

```
myCars <- data.frame(mtcars[,1:6])
myCars[1:5,1:5]
```

```
##                   mpg cyl disp  hp drat
## Mazda RX4         21.0   6  160 110 3.90
## Mazda RX4 Wag     21.0   6  160 110 3.90
## Datsun 710        22.8   4  108  93 3.85
## Hornet 4 Drive    21.4   6  258 110 3.08
## Hornet Sportabout 18.7   8  360 175 3.15
```

Question 2: MyCars Correlation

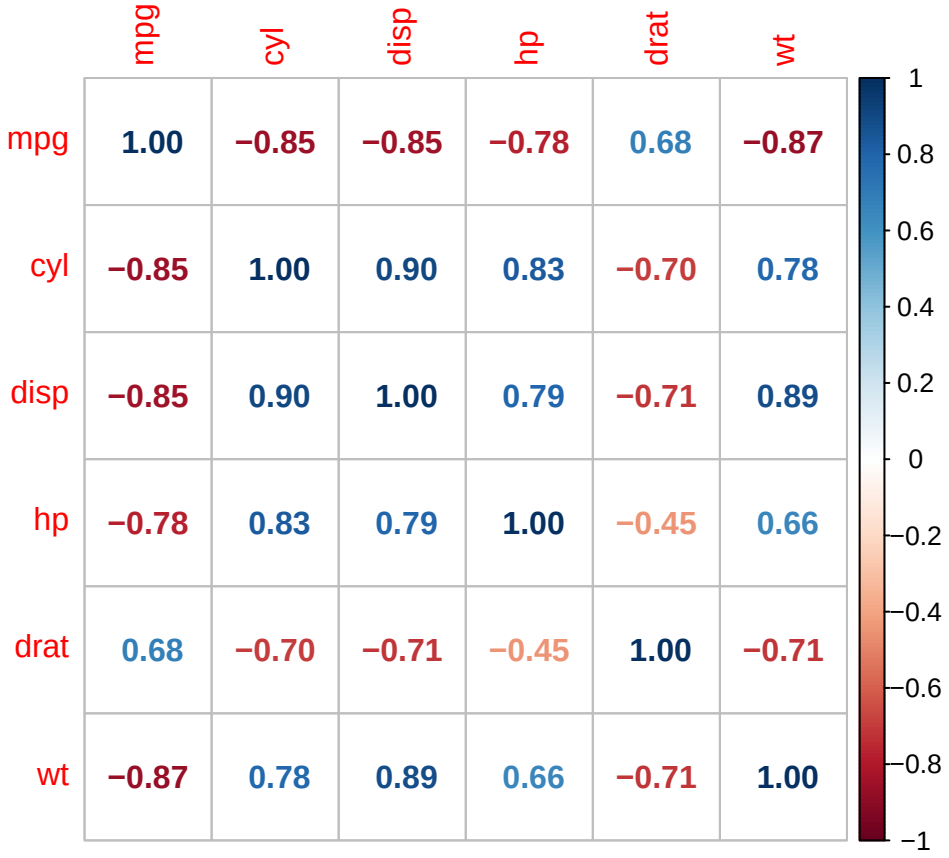
Create and interpret a bivariate correlation matrix using `cor(myCars)` keeping in mind the idea that you will be trying to predict the mpg variable. Which other variable might be the single best predictor of mpg?

Interpreting the Results

Some good predictors might by cyl, disp, and wt. With all of these independent variable below -0.80. We either want a high negative correlation or a high positive correlation. HP might also be a good variable, but does not have the highest correlation. We will run verify when running in the linear regression model.

```
myCars_cor <- cor(myCars)

corrplot(myCars_cor, method = "number")
```



Question 3 MyCar Linear Regression

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.

Interpreting the Results

- The p-value of the F Statistics is well below 0.05 at .0001
- The Adjusted R-squared is at 0.8148 which is very strong.
- The Adjusted R-squared is saying that 81.5% of the dependant variables can be explained by the independent variables.
- The individual p-values are alos very strong with being well below 0.05 at 0.001.

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

```
##
## 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
```

Question 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.

Building an Equation

the equation for the linear regression model would be the following.

mpg = -3.87783(wt) + -0.03177(hp) + 37.22727

22.09908 = -3.87783(3) + -0.03177(110) + 37.22727

MPG for this type of car would be 22.09908.

```
hp <- 110
wt <- 3

-3.87783 * wt + -0.03177* hp + 37.22727
```

```
## [1] 22.09908
```

Question 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?

Interpretting the Results

The odds are extremely high at 788547604:1 in which Bayes Factor overwhelmingly favors a model that includes the three predictors.

```
MyCarsBFLm <- lmBF(mpg ~ wt + hp ,data = myCars, )
MyCarsBFLm
```

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

Question 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.

Interpretting the results

We tested a model of Car specifications that used two variables to predict MPG: Weight and Horsepower, A Bayesian analysis of this model showed a mean posterior estimate for R-squared of 0.794, with the highest density interval ranging from roughly 0.19 to 0.21. The traditional analysis confirmed this result with a slightly more optimistic R-squared of 0.815. The F-test on this value was F(2, 29)=69.21.0, p<.001, so we reject the null hypothesis that R-squared was equal to zero. Both predictors were also significant with B-weights less than 0.001 for weight and horsepower. The Bayes factor of 788547604 was strongly in favor of the two predictor model (in comparison with an intercept-only model).

```
MyCarsBFmcmc <- lmBF(mpg ~ wt + hp ,data = myCars, posterior=TRUE, iterations=100000)
summary(MyCarsBFmcmc)
```

```
##
## Iterations = 1:1e+05
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 1e+05
##
## 1. Empirical mean and standard deviation for each variable,
```

```
##    plus standard error of the mean:
##
##          Mean        SD  Naive SE Time-series SE
## mu   20.09123  0.484270 1.531e-03      1.545e-03
## wt   -3.78590  0.663197 2.097e-03      2.126e-03
## hp   -0.03102  0.009429 2.982e-05      2.982e-05
## sig2  7.48874  2.155430 6.816e-03      8.264e-03
## g     4.19888 41.483466 1.312e-01      1.312e-01
##
## 2. Quantiles for each variable:
##
##          2.5%      25%      50%      75%    97.5%
## mu   19.13147 19.77238 20.09133 20.40986 21.04364
## wt   -5.08805 -4.22187 -3.78938 -3.34854 -2.47901
## hp   -0.04952 -0.03723 -0.03103 -0.02485 -0.01239
## sig2  4.39381  5.97363  7.12101  8.57978 12.72744
## g     0.35796  0.93757  1.70726  3.42322 19.94455
```

```
rsqList <- 1 - (MyCarsBFmcmc[,"sig2"] / var(myCars$mpg))
mean(rsqList)
```

```
## [1] 0.7938356
```

```
quantile(rsqList,c(0.025))
```

```
##      2.5%
## 0.6496145
```

```
quantile(rsqList,c(0.975))
```

```
##     97.5%
## 0.8790388
```

Question 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`.

Interperting vif

The definition from the help function is as followed, calculates variance-inflation and generalized variance-inflation factors (VIFs and GVIFs) for linear, generalized linear, and other regression models.

In simpler terms vif helps in solving for and identify multicollinearity. That is there might be redundancies between predictor variables in which will hurt the model. Vif computes a score called the variance inflation factor in which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

A value that exceed 5 or 10 indicates a problematic amount of collinearity.

```
library(car)
```

```
## Loading required package: carData
```

Question 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.

Interpreting the results.

- The first model from exercise 2 shows that there is no multicollinearity where values are less than 5 and 10.
- The second model from looking at all independent variables show that cycl, wt, disp all have multicollinarity with being above 5 and 10.
- We either need to drop variables or due some sort of mutation in order to reduce multicolinarity.

```
vif(MyCarsLm)
```

```
##       wt       hp
## 1.766625 1.766625
```

```
FivePredictor_Model <- lm(mpg ~., myCars)

summary(FivePredictor_Model)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = myCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7014 -1.6850 -0.4226  1.1681  5.7263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.00836    7.57144   4.756  6.4e-05 ***
## cyl         -1.10749    0.71588  -1.547  0.13394
## disp         0.01236    0.01190   1.039  0.30845
## hp          -0.02402    0.01328  -1.809  0.08208 .
## drat         0.95221    1.39085   0.685  0.49964
## wt          -3.67329    1.05900  -3.469  0.00184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 26 degrees of freedom
## Multiple R-squared:  0.8513, Adjusted R-squared:  0.8227
## F-statistic: 29.77 on 5 and 26 DF,  p-value: 5.618e-10
```

```
vif(FivePredictor_Model)
```

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