IST 777

Joe Hernandez

8/29/2019

Homework #7

1. Create new data frame from part of dataset command

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

2. Create and interpret bivariate correlation matrix trying to predict mpg variable

```
> cor(myCars)
            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
```

According to the bivariate correlation matrix, the only positive predictor in predicting the mpg variable was the rear axle ratio. The number of cylinders, displacement, and weight had nearly similar inverse correlation at predicting the mpg with horsepower trailing behind; however, the single best predictor would be the weight at ~ -0.8677.

3. Run multiple regression analysis on myCars

```
> carregOut1 <- lm(formula = mpg ~ wt + hp, data = myCars)
> summary(carregOut1)

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
```

Median value of -0.182 in residuals suggest a slight symmetrically left-skewed distribution of the data. The coefficient estimates provide us the mpg offset of 37.22727 with the B-weight for wt at -3.87783 and B-weight for hp at -0.03177. The standard error spreadthe estimates for the mpg intercept by 1.59879, wt by 0.63273, and hp by 0.00903. The t-value shows each estimated coefficient is not equal to zero with an associated probability of $p < 0.01$ we strongly reject the null hypothesis for each coefficient.
The effect size for this regression shows a multiple R-squared value of 0.8268 (accounting for 82% of variability in mpg). This is very strong for predictive modeling.
Finally the summary statistics shows the residual standard error as 2.593 on 118 degrees of freedom and an F-statistic the overall p-value is close to zero confirming the rejection of the null hypothesis for this predictive model.

4. Construct prediction equation for mpg using all three coefficients

```
mpg = 37.22727 - 3.87783*(wt) - 0.03177*(hp)
Predict mpg for car with 110 hp and weight 3 tons
> 37.22727 - 3.87783*(3) - 0.03177*(110)
[1] 22.09908
```

With the aforementioned specifications of the car the predicted mpg is 22.09908.

5. Run multiple regression analysis on myCars

```
> library(BayesFactor)
> carmultregOut1 <- lmBF(formula = mpg ~ wt + hp, data = myCars)
> summary(carmultregOut1)
Bayes factor analysis
--------------
[1] wt + hp : 788547604 ±0%

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

The odds for the alternative hypothesis is very strong toward the predictor model over the intercept only. The "frequentist" NHST we reject the null hypothesis for the two predictors as well as the overall R-squared. The HDI from the MCMC output shows estimates for the coefficients and R-squared that concur with the frequentist model as well. By integrating these three types of evidence we can determine the variables are significant in predicting mpg.

6. Run multiple regeression analysis again using posterior = TRUE and iterations = 10000

```
> carmultregOut2 <- lmBF(formula = mpg ~ wt + hp, data = myCars,
posterior = TRUE, iterations = 10000)
|----|----|----|----|----|----|----|----|----|----|
**************************************************|
> summary(carmultregOut2)

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   20.099   0.495923 4.959e-03      4.959e-03
wt   -3.796   0.663621 6.636e-03      6.902e-03
hp   -0.031   0.009464 9.464e-05      9.664e-05
sig2  7.517   2.495672 2.496e-02      2.917e-02
g     4.138 16.157408 1.616e-01      1.616e-01

2. Quantiles for each variable:

         2.5%      25%      50%      75%    97.5%
mu   19.14941 19.77413 20.10112 20.41785 21.05197
wt   -5.09669 -4.24000 -3.80160 -3.36975 -2.47572
hp   -0.04935 -0.03722 -0.03103 -0.02479 -0.01236
sig2  4.36033  5.97722  7.13340  8.60261 12.73988
g     0.36497  0.94603  1.70688  3.41430 21.11898
```

We tested a model of vehicle range 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 0.81, with the highest density interval for each B-weight ranging from roughly -5.10344 to -2.47169 for weight and -0.0497 to -0.01209 for horsepower. The traditional analysis confirmed this result with a slightly more optimistic R-squared of 0.83. The F-test on this value was F(2, 29)=69, p<.001, so we reject the null hypothesis that R-squared was equal to zero. All three predictors were also significant with B-weights of -3.88 (wt) and -0.032 (hp). The Bayes factor of 788547604 was strongly in favor of the two predictor model (in comparison with an intercept-

only model).

7. Install and load car package

```
> library("car")
> # Read help file for vif()
> ?vif()
```

Variable Inflation Factor is a way to determine if a variable should be included in a regression model. Doing so would significantly affect the prediction's stability. A rule of thumb for interpreting vif is if it has a factor greater than or equal to 5 (or 10 based on threshold) it would be better to leave out the variable from the regression.
https://www.youtube.com/watch?v=0SBIXgPVex8
https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/

8. Run vif() on model

```
> vif(carregOut1)
      wt       hp
1.766625 1.766625
```

Based on the Variable Inflation Factor criteria the current B-weights for predicting mpg using only weight and horsepower have stable factors less than 5 and would be included.

Run model that predicts from all 5 predictors in my cars

```
> carregOut2 <- lm(formula = mpg ~ wt + hp + cyl + disp + drat,
data = myCars)
> summary(carregOut2)

Call:
lm(formula = mpg ~ wt + hp + cyl + disp + drat, 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 ***
wt          -3.67329    1.05900  -3.469  0.00184 **
hp          -0.02402    0.01328  -1.809  0.08208 .
cyl         -1.10749    0.71588  -1.547  0.13394
disp         0.01236    0.01190   1.039  0.30845
drat         0.95221    1.39085   0.685  0.49964
---
```

```
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(carregOut2)
       wt        hp       cyl      disp      drat
 5.168795  3.990380  7.869010 10.463957  2.662298
```

Based on the Variable Inflation Factor criteria using all the B-weights for predicting mpg using all variables in the myCars data set, we find only horsepower and rear axle ratio as the most stable variables to include in predicting mpg. The weight of the vehicle is slightly over the vif greater than or equal to five threshold and can be speculated that by removing the two additional variables well above the threshold would lower the vif such that weight would become a stable B-weight.

```
> carregOut3 <- lm(formula = mpg ~ wt + hp + drat, data = myCars)
> summary(carregOut3)

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.3598 -1.8374 -0.5099  0.9681  5.7078

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.394934   6.156303   4.775 5.13e-05 ***
wt          -3.227954   0.796398  -4.053 0.000364 ***
hp          -0.032230   0.008925  -3.611 0.001178 **
drat         1.615049   1.226983   1.316 0.198755
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.561 on 28 degrees of freedom
Multiple R-squared:  0.8369,     Adjusted R-squared:  0.8194
F-statistic: 47.88 on 3 and 28 DF,  p-value: 3.768e-11
> vif(carregOut3)
      wt       hp     drat
2.869445 1.769308 2.033837
```

After removing number of cylinders and displacement the Variable Inflation Factor for weight, horsepower, and rear axle ratio is stable below the threshold.