Quinn Knudsen

Homework 8

Topic: **Multiple Regression/Linear Prediction**

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

2)

> 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

I believe the best predictor of MPG from the correlation matrix will be weight (wt) as it has the highest absolute value in terms of it’s correlational strength.

3)

> lm.1 <- lm(mpg~wt+hp,data=myCars)

> summary(lm.1)

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 results of this model indicate a that 81% of the variance in mpg can be predicted by weight and horsepower p<.001. This is a strong result because it leaves only 19% of the variance unexplained by unmeasured variables. Each of the independent variables, weight (-3.88, p<.001) and horsepower (-0.031, p<.01) were significant negative predicters.

> QuantPsyc::lm.beta(lm.1)

wt hp

-0.6295545 -0.3614507

I would consider these predictors to be strong due to the level of significance and standardized Beta weight which indicates fairly robust predictive power. I used the Beta weight to standardized these predictors because the B-weight estimate looked to be small for horsepower, but artificially so due to the variability on the scale horsepower and weight reside on. It might be good practice to center the predictors before running the model in the future.

4)

> (-3.87783\*(3))+(-0.03177\*(110))+37.22727

[1] 22.09908

I would predict that this vehicle should have a mpg of ~22 based upon the model we developed.

5)

> stateOutBF <- lmBF(mpg~wt+hp,data=myCars, posterior=FALSE)

> stateOutBF

Bayes factor analysis

--------------

[1] wt + hp : 788547604 ±0%

Against denominator:

Intercept only

---

Bayes factor type: BFlinearModel, JZS

This strengthens my interpretation of a very strong effect from the frequentist approach run earlier as this odds ratio falls well beyond the very strong parameter (788547604 ±0%).

6)

> stateOutBF <- lmBF(mpg~wt+hp,data=myCars, posterior=TRUE, iterations=100000)

|----|----|----|----|----|----|----|----|----|----|

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*|

> summary(stateOutBF)

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 20.08962 0.482576 1.526e-03 1.483e-03

wt -3.78568 0.662771 2.096e-03 2.140e-03

hp -0.03104 0.009416 2.978e-05 2.993e-05

sig2 7.49451 2.164023 6.843e-03 8.246e-03

g 4.05626 30.964955 9.792e-02 9.792e-02

2. Quantiles for each variable:

2.5% 25% 50% 75% 97.5%

mu 19.13670 19.77112 20.09060 20.40765 21.04319

wt -5.08932 -4.22373 -3.78803 -3.34947 -2.46924

hp -0.04942 -0.03727 -0.03103 -0.02489 -0.01239

sig2 4.38299 5.97391 7.12742 8.59337 12.75126

g 0.35889 0.94312 1.70718 3.42954 19.49346

The results here demonstrate a population mean estimate of 20 which is slightly more conservative than the linear model run previously. It also demonstrates through the 95% HDI that both independent variables are negative and do not cross the threshold of zero.

7)

> vif(lm.1)

wt hp

1.766625 1.766625

* 1 = not correlated.
* Between 1 and 5 = moderately correlated.
* Greater than 5 = highly correlated.

These variables, although somewhat correlated, fall within an acceptable VIF range (<2). Some research I have seen on the topic is a generous as to allow a VIF of up to 10, but after 5 there is grounds for concern for the distinctness of the independent variables. Dimension reduction (PCA, additive linear combination, removing one variable) might be worthwhile exploring at that point.

8)

> lm.vif <- lm(mpg~.,data=myCars)

> summary(lm.vif)

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(lm.vif)

cyl disp hp drat wt

7.869010 10.463957 3.990380 2.662298 5.168795

Here we see the independent variables are too highly correlated and VIF values > than 5 for weight, cylinder, and displacement.

> 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

As seen previously, there are high correlations between weight and displacement and cylinder and they overlap in the variance explained. This creates instability in the prediction and a less favorable model for use. Also, it is important to see that the adjusted r squared for this model using all of the variables had little to no incremental variance gain over the weight and horsepower model (which had acceptable VIF).