Data-Driven Model Selection by Cross Validation

Yogindra Raghav

December 5, 2018

```
library(mdsr)
library(broom)
```

1. Sum of Squared Errors:

```
mod = lm(mpg~. , data = mtcars)
residuals = augment(mod, mtcars) %>% select(.resid)
residuals^2 %>% sum()
```

```
## [1] 147.4944
```

2. Coefficients that are significantly different from zero:

```
summary(mod)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

According to the p-values, none of the coefficients are significantly differ from zero.

3. Row numbers for the first partition for the training data are rows 1-4,6-8, 10-16, 18-19, 21-26, 30-32.

4. Based on the SSE value from cross-validation, we get 593. However, in excercise 1 we got an SSE of 147.4944.