Sums of Squares and Multicollinearity

For this tutorial, we will use the `mileage.txt` data set. The data come from 32 classic automobiles. The variables are:

- y: gas mileage (miles/gallon)
- x_1: Displacement (cubic in.)
- x_2: Horsepower (ft-lb)
- x_3: Torque (ft-lb)
- x_4: Compression ratio
- x_5: Rear axle ratio
- x_6: Carburetor (barrels)
- x_7: No. of transmission speeds
- x_8: Overall length (in.)
- x_9: Width (in.)
- x_{10}: Weight (lb)
- x_{11}: Type of transmission (automatic/manual)

Read the data in and fit MLR model regressing the response variable against four predictors: x_1, x_2, x_6, x_{10}.

```
Data <- read.table("mileage.txt", header=TRUE)

##regress gas mileage against 4 predictors
result<-lm(y~x1+x2+x6+x10, data=Data)
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x6 + x10, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7167 -1.7716 -0.4098  1.7031  6.2422
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.628823   3.496822   9.903 1.75e-10 ***
```

```
## x1          -0.044284   0.021167  -2.092    0.046 *
## x2           0.005186   0.048889   0.106    0.916
## x6           0.708581   0.974944   0.727    0.474
## x10         -0.001207   0.002040  -0.592    0.559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.102 on 27 degrees of freedom
## Multiple R-squared:   0.79,  Adjusted R-squared:  0.7589
## F-statistic:  25.4 on 4 and 27 DF,  p-value: 8.246e-09
```

Looking at the t tests for the coefficients, the coefficients for x_2, x_6, x_{10} are insignificant. We know that the t tests do not inform us if we can drop all of these predictors simultaneously from the model, so we need to conduct a partial F test.

The null and alternative hypotheses are:

$H_0 : \beta_2 = \beta_6 = \beta_{10} = 0$,

H_a : at least one of the coefficients in H_0 is not 0.

In words, the null hypothesis supports going with the reduced model by dropping x_2, x_6, x_{10}, whereas the alternative hypothesis supports the full model by not dropping x_2, x_6, x_{10}.

We will explore two different approaches to conduct this partial F test using R.

1. Partial F test: Approach 1

In this approach, we fit the reduced model, and then use the `anova()` function to compare the reduced model with the full model

```
##fit the reduced model with just displacement
reduced<-lm(y~x1, data=Data)

##perform the partial F test to see if we can
##drop the last 3 predictors
anova(reduced,result)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2 + x6 + x10
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     30 281.82
## 2     27 259.86  3    21.966 0.7608 0.5259
```

The F statistic from this test is 0.7608, with a p-value of 0.5259. So we fail to reject the null hypothesis, so there is little evidence of supporting the full model. We go with the reduced model over the full model.

2. Partial F test: Approach 2

In this other approach, we use the `anova()` function on the full model to obtain the sequential sums of squares associated with the full model

```
anova(result)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 955.72  955.72 99.3021 1.531e-10 ***
## x2         1   6.55    6.55  0.6810    0.4165
## x6         1  12.04   12.04  1.2510    0.2732
## x10        1   3.37    3.37  0.3503    0.5588
## Residuals 27 259.86    9.62
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The values under the column "Sum Sq" give the sequential SS_Rs. So,

- for the first line, we have $SS_R(\beta_1) = 955.72$,
- the second line, we have $SS_R(\beta_2|\beta_1) = 6.55$,
- then $SS_R(\beta_6|\beta_1, \beta_2) = 12.04$,
- and then $SS_R(\beta_{10}|\beta_1, \beta_2, \beta_6) = 3.37$.
- The very last line refers to $SS_{Res}(\beta_1, \beta_2, \beta_6, \beta_{10})$.

The test statistic for the partial F test is

$$F_0 = \frac{(6.55 + 12.04 + 3.37)/3}{259.86/27}$$
$$= 0.7605634$$

```
F0<-((6.55+12.04+3.37)/3)/(259.86/27)
F0
```

```
## [1] 0.7605634
```

```
##find the p-value for this F statistic
1-pf(F0,3,27)
```

```
## [1] 0.526051
```

```
##find the critical value for the F(3,27) distribution at 0.05 sig level
qf(0.95,3,27)
```

```
## [1] 2.960351
```

So we fail to reject the null hypothesis, and go with the reduced model over the full model. Notice that both approaches result in the same F statistic (other than rounding off).

3. Multicollinearity

With MLR, we need to consider the presence of multicollinearity in our model. From the `summary()` output of the full model, notice that:

1. the t tests for the coefficients were almost all insignificant (and the one significant test was barely significant), yet the ANOVA F test was highly significant.
2. predictors that should relate to gas mileage such as horsepower (x_2) and weight (x_{10}) had insignificant t tests.

These observations support the presence of multicollinearity in our model. There are other ways to check if multicollinearity is present.

We can produce a matrix of correlations involving all quantitative predictors, using the `cor()` function

```
cor(Data[,c(2,3,7,11)])
```

```
##            x1        x2        x6       x10
## x1  1.0000000 0.9452080 0.6590601 0.9456621
## x2  0.9452080 1.0000000 0.7719099 0.8834004
## x6  0.6590601 0.7719099 1.0000000 0.5206424
## x10 0.9456621 0.8834004 0.5206424 1.0000000
```

Notice the high correlations between some pairs of predictors.

We can also find the variance inflation factors (VIFs) for the model. We use the `vif()` function from the `faraway` package

```
library(faraway)
vif(result)
```

```
##        x1        x2        x6       x10
## 19.836177 15.576450  3.527631 12.043521
```

Another way to calculate the VIF, for example, for x_1

```
result2<-lm(x1~x2+x6+x10, data=Data)
summary(result2)
```

```
##
## Call:
## lm(formula = x1 ~ x2 + x6 + x10, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.256 -20.202   6.908  17.320  41.318
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -120.72107   21.31247  -5.664 4.53e-06 ***
## x2             1.16453    0.37695   3.089   0.0045 **
## x6             4.09390    8.67003   0.472   0.6405
## x10            0.06573    0.01332   4.935 3.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.7 on 28 degrees of freedom
## Multiple R-squared:  0.9496, Adjusted R-squared:  0.9442
## F-statistic: 175.8 on 3 and 28 DF,  p-value: < 2.2e-16
```

```
1/(1-0.9496)
```

```
## [1] 19.84127
```