Stat 6021 R Tutorial: Extra Sums of Squares and Variance Inflation Factors

For this tutorial, we will use the "mileage.txt" data set. The data come from 32 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)

1. Fit the multiple regression model using the predictors x_1, x_2, x_6, and x_{10}. Use the `summary()` function and make the relevant conclusions from the t tests and the ANOVA F test.

2. We want to investigate if we can drop the insignificant predictors from the model. There are two approaches to carry out the partial F test in R. Our null hypothesis is $H_0 : \beta_2 = \beta_6 = \beta_{10} = 0$.

    - Approach 1: Fit the reduced model, and compare the SS_R with the full model that contains all the predictors. E.g.:

```
result<-lm(y~x1+x2+x6+x10)
reduced<-lm(y~x1)
anova(reduced,result)
```

From this output, what is the F statistic for testing $\beta_2 = \beta_6 = \beta_{10} = 0$? What is the relevant conclusion?

- Approach 2: Use the `anova()` function on the full model to produce all the sequential sum of squares, i.e. `anova(result)`. It is crucial that the predictors you are testing are listed last in `lm()`. Your output should look like this:

```
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
```

The values under the column "Sum Sq" give the sequential SSRs. 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 partial F statistic for this test is

$$\begin{aligned} F_0 &= \frac{SS_R(\beta_2, \beta_6, \beta_{10}|\beta_1)/r}{MS_{Res}(\beta_1, \beta_2, \beta_6, \beta_{10})} \\ &= \frac{\{SS_R(\beta_2|\beta_1) + SS_R(\beta_6|\beta_1, \beta_2) + SS_R(\beta_{10}|\beta_1, \beta_2, \beta_6)\}/3}{SS_{Res}(\beta_1, \beta_2, \beta_6, \beta_{10})/27} \end{aligned}$$

3. Compare your F statistic from both approaches. Do you recall how to use R to obtain the p-value? How about the critical value?

4. Checking all pairwise correlations. You might suspect multicollinearity. One informal way to check is to look at all pairwise correlations. Type

```
preds<-cbind(x1,x2,x6,x10)
cor(preds)
```

The function `cbind()` combines the vectors for each predictor by columns. The function `cor()` provides a matrix of correlations between all possible pairs of variables stored in the columns. If you find there are too many decimal points, you can try `round(cor(preds),3)`, which will round off all values to three decimal places.

5. VIFs. The function `vif()` does not belong to the base R package. A search on the World Wide Web will show that a function to calculate VIFs belong to various packages. The one we will use belongs to the faraway package. You need to install this package before proceeding. Follow these steps to install a package:

- Choose "Install package(s)" from the "Packages" menu.
- Select a CRAN Mirror. By convention, choose the location closest to you.
- Select a package, e.g., faraway, and click OK.

Once a package is installed, you can load it anytime by typing `library(faraway)`.

To obtain the VIFs for our regression model, type `vif(result)`. What is the interpretation of these values?

6. VIF calculation. Consider the VIF for the predictor *Displacement*. Regress *Displacement* against the other predictors, and find the R^2 for this model.

$$\text{VIF}_{\text{Displacement}} = \frac{1}{1 - 0.9496} = 19.84,$$

which is the VIF for the predictor *Displacement*.

7. One way to deal with multicollinearity is to remove some of the highly correlated predictors from the model. What will be your choice here? This decision is sometimes guided by scientific reasons, and sometimes by practical reasons.