Homework 6 Solutions:

Mon, Nov 4: Chapter 7: #1, 2, 5, 6
Wed, Nov 6: Assignment on handout

Assigned Mon, Nov 4:

7.1 In each case, df = the number of terms being added to the model. So, they are:
(1) 1 **(2)** 1 **(3)** 2 **(4)** 3

7.2 In general, the term "extra sums of squares" is the amount of SSTotal that gets moved from SSE to SSR when an explanatory variable is added to the model. *SSR(X_1)* measures that reduction when X_1 is added as the only explanatory variable. In other words, it measures the amount that gets moved from SSTotal to SSR when comparing the model with the intercept only to the model with the intercept and X_1.

7.5 **a.** To decompose the sum of squares in the requested order in R, you need to enter the variables into the linear model function in the order X_2, X_1, X_3. Here are the R code and results:

```
Problem75 <- lm(Y~ X2 + X1 + X3, data = Ch7Pr6)
summary(Problem75)
anova(Problem75)
Response: Y
               Df Sum Sq Mean Sq F value    Pr(>F)
     X2         1 4860.3  4860.3 48.0439 1.822e-08 ***
     X1         1 3896.0  3896.0 38.5126 2.008e-07 ***
     X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
```

Therefore, $SSR(X_2) = 4860.3$, $SSR(X_1|X_2) = 3896.0$ and $SSR(X_3|X_2, X_1) = 364.2$.

b. To test H_0: $\beta_3 = 0$ versus Ha: $\beta_3 \neq 0$, reject H_0 if p-value < $\alpha = 0.025$. From the output in (a), F* = 3.5997, p-value = 0.06468. Thus, we cannot reject Ho, and conclude that there is insufficient evidence to indicate that $\beta_3 \neq 0$. So the model without X_3, anxiety level, is not significantly better than the model that already contains X_1, patient's age, and X_2, severity of illness, in predicting patient's satisfaction. (As a side note, the p-value is relatively small, so it would still make sense to include this variable in the model, assuming it is easy to collect!)

7.6 The hypotheses are H_0: $\beta_2 = \beta_3 = 0$; H_a: At least one of β_2 and β_3 is not 0.
Fit the Full model using all 3 variables, and the reduced model using X_1 only, then compare.

```
Full <- lm(Y~X1+X2+X3, data=Ch7Pr6)
Reduced <- lm(Y~X1, data=Ch7Pr6)
anova(Reduced,Full)
Analysis of Variance Table

Model 1: Y ~ X1
Model 2: Y ~ X1 + X2 + X3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     44 5093.9
2     42 4248.8  2    845.07 4.1768 0.02216 *
```

The output shows that F* = 4.1768 and the p-value = 0.02216. Therefore, using $\alpha = 0.025$, we can reject the null hypothesis (just barely!) and conclude that at least one of the variables X_2 and X_3 is needed, even with X_1 already in the model.

Assigned Wed, Nov 6, on handout:

1. Find the correlation between these sets of variables:
   Solution:
   The entire correlation matrix is shown below. You only needed to find the ones requested:
   a. Male and Female: −1.000
   b. RtFoot and LeftFoot: 0.944
   c. HeadCirc and RtFoot: 0.475
   d. HeadCirc and LeftFoot: 0.467

```
         Height LeftArm  RtArm LeftFoot RtFoot LeftHand  RtHand HeadCirc   nose Female   Male
Height    1.000   0.748  0.695    0.819  0.809   0.3690  0.4150   0.4236  0.284 -0.711  0.711
LeftArm   0.748   1.000  0.885    0.603  0.593   0.3345  0.3558   0.3044  0.320 -0.569  0.569
RtArm     0.695   0.885  1.000    0.561  0.620   0.2877  0.3203   0.3496  0.297 -0.511  0.511
LeftFoot  0.819   0.603  0.561    1.000  0.944   0.3105  0.3741   0.4666  0.304 -0.773  0.773
RtFoot    0.809   0.593  0.620    0.944  1.000   0.2845  0.3760   0.4754  0.287 -0.717  0.717
LeftHand  0.369   0.334  0.288    0.310  0.285   1.0000  0.9353   0.0413  0.174 -0.495  0.495
RtHand    0.415   0.356  0.320    0.374  0.376   0.9353  1.0000   0.0927  0.217 -0.531  0.531
HeadCirc  0.424   0.304  0.350    0.467  0.475   0.0413  0.0927   1.0000  0.137 -0.489  0.489
nose      0.284   0.320  0.297    0.304  0.287   0.1744  0.2170   0.1367  1.000 -0.384  0.384
Female   -0.711  -0.569 -0.511   -0.773 -0.717  -0.4950 -0.5309  -0.4894 -0.384  1.000 -1.000
Male      0.711   0.569  0.511    0.773  0.717   0.4950  0.5309   0.4894  0.384 -1.000  1.000
```

2. Try fitting the model with *only* the two predictors Male and Female. Look at the `summary` of the model in R. What kind of message does R give? Explain why.
   Solution*: The output is shown below, with the message highlighted. The problem is that Male and Female have a perfect negative correlation, so it is redundant to have them both in the model. (The X matrix is singular.)*

```
lm(formula = HeadCirc ~ Male + Female, data = PhysicalData)
Residuals:
   Min     1Q Median     3Q    Max
-4.076 -1.231  0.113  1.019  5.113
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   55.887      0.361  154.68  < 2e-16 ***
Male           2.189      0.536    4.09  0.00015 ***
Female            NA         NA      NA       NA
Residual standard error: 1.98 on 53 degrees of freedom
Multiple R-squared: 0.239,    Adjusted R-squared: 0.225
F-statistic: 16.7 on 1 and 53 DF,  p-value: 0.00015
```

The relevant output for Questions 3 to 8 is shown below. Also, Adjusted R-squared is needed for Question 10.

```
lm(formula = HeadCirc ~ LeftFoot + RtFoot + Male, data = PhysicalData)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.247      4.056   12.64   <2e-16 ***
LeftFoot      -0.113      0.358   -0.32    0.753
RtFoot         0.306      0.310    0.99    0.328
Male           1.478      0.843    1.75    0.086 .

Residual standard error: 1.97 on 51 degrees of freedom
Multiple R-squared: 0.273, Adjusted R-squared: 0.23
F-statistic: 6.38 on 3 and 51 DF,  p-value: 0.000939

> anova(Full)
Analysis of Variance Table
Response: HeadCirc
          Df Sum Sq Mean Sq F value  Pr(>F)
LeftFoot   1   59.4    59.4   15.27 0.00028 ***
RtFoot     1    3.1     3.1    0.79 0.37901
Male       1   12.0    12.0    3.07 0.08565 .
Residuals 51  198.5     3.9
```

3. Test the coefficient corresponding to each of the three X variables. I.e., test the hypotheses H_0: $\beta_j = 0$ versus H_a: $\beta_j \neq 0$ for $j = 1, 2, 3$. Give a test statistic and *p*-value and make a conclusion for each test. Use $\alpha = 0.05$. Are any of them statistically significantly different from 0?
   ***Solution**: These are given in the first part of the output.*
   H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, $t = -0.32$, $p = 0.753$, do not reject H_0.
   H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$, $t = 0.99$, $p = 0.328$, do not reject H_0.
   H_0: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$, $t = 1.75$, $p = 0.086$, do not reject H_0.
   None of them are statistically different from 0.

4. Get the Anova table for the full model using `>anova(Full)`. Use the Anova table to test the null hypothesis H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, i.e., to test whether LeftFoot should be in the model. Use $\alpha = 0.05$.
   ***Solution**:* From the output above, F-value = 15.27, $p = 0.00028$. Clearly reject H_0.

5. Compare the results of your test of H_0: $\beta_1 = 0$ in Questions 3 and 4. Did you reach the same conclusion? If so, what was it? If not, explain why not.
   ***Solution**:* The conclusion was different. In Question 3 the null hypothesis H_0: $\beta_1 = 0$ was not rejected, and in Question 4 it was. That's because the hypothesis in Question 3 is that LeftFoot is not needed in the model *given* that RightFt and Male are already in the model. In Question 4 the hypothesis is that LeftFoot is not needed when there are no other explanatory variables already in the model.

6. Test the overall hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. Give a test statistic and *p*-value. What is your conclusion? Use $\alpha = 0.05$.
   ***Solution**:* Use this part of the output: `F-statistic: 6.38 on 3 and 51 DF,  p-value: 0.000939`
   Clearly reject H_0 and conclude that at least one of the variables should be in the model.

7. Compare the results in parts Questions 3 and 6. In both cases, you tested all 3 of the coefficients. Did you reach the same conclusion in both questions? If so, what is the conclusion? If not, explain why not.
   ***Solution**:* No, the conclusion in Question 3 was that none of the β_j were significantly different from 0, while in Question 3 the conclusion was that at least one of them was. That's because in Question 3, the hypothesis in each case was that the variable does not significantly add to the model *given* that the other two variables are already in the model.

8. Find the Variance Inflation Factor for each of the 3 variables. Use the command `>vif(Full)` where "Full" is what you called the full model when you fit it. Based on these VIF values, what do you suggest about what variable(s) to include in your model?
   ***Solution**:* The results are shown below. Both LeftFoot and RtFoot have vif values greater than 5, which says that they are highly correlated with the combination of the other two variables, so they should not both be in the model. (Remember that if a predictor variable X has a vif value greater than 5, it indicates that R^2 would be at least 80% if you tried to predict X as the response, using the others as explanatory variables.)

```
> vif(Full)
LeftFoot   RtFoot     Male
   11.07     9.19     2.49
```

9. Fit two more models, call the first one Left and use the variables LeftFoot and Male. Call the other one Right and use the variables RtFoot and Male. Are any of the variables statistically significant predictors based on the tests of the individual coefficients? Use $\alpha = 0.05$.
   ***Solution**:* Relevant output is shown below. None of the variables is statistically significant based on the *t*-tests. Again, it's because each one is tested *given* that the other variable is in the model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.174      4.054   12.62   <2e-16 ***
LeftFoot       0.198      0.169    1.17    0.248
Male           1.430      0.842    1.70    0.095 .
```

```
Residual standard error: 1.97 on 52 degrees of freedom
Multiple R-squared: 0.259,    Adjusted R-squared: 0.23
F-statistic: 9.08 on 2 and 52 DF,  p-value: 0.000414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.621      3.510   14.42   <2e-16 ***
RtFoot         0.219      0.145    1.51    0.138
Male           1.367      0.760    1.80    0.078 .
Residual standard error: 1.96 on 52 degrees of freedom
Multiple R-squared: 0.271,    Adjusted R-squared: 0.243
F-statistic: 9.68 on 2 and 52 DF,  p-value: 0.000266
```

10. Which of the 3 models (Full, Left, Right) would you recommend be used to predict Head circumference? Give a statistical reason for choosing that model.
    Solution: To compare models, use the one with highest Adjusted R-squared or lowest Residual standard error. In either case the best model is the one with RtFoot and Male. It has Adjusted R-squared of 0.243 (24.3%) and Residual standard error of 1.96.