Extra Sums of Squares in R

We will work again with the data from Problem 6.9, "Grocery Retailer." Recall that after you fit a linear regression model, say `Retailer`, and obtain the ANOVA table using the function

```
> anova(Retailer)
```

you get a row of sum of squares for each predictor variable in the model:

```
Analysis of Variance Table

Response: Hours
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Cases      1  136366  136366  6.6417   0.01309 *  
Costs      1    5726    5726  0.2789   0.59987    
Holiday    1 2034514 2034514 99.0905 2.941e-13 ***
Residuals 48  985530   20532                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For our model, which I named "Retailer," we had X_1 = Cases, X_2 = Costs, and X_3 = Holiday. The ANOVA table given by R provides the extra sum of squares for each predictor variable, *given that* the previous predictors are already in the model. Thus the Sum of Squares given for "Cases" is $SSR(X_1) = 136366$, while the Sum of Squares given for "Costs" is $SSR(X_2 \mid X_1) = 5726$, and the Sum of Squares given for "Holiday" is $SSR(X_3 \mid X_1, X_2) = 2034514$. This corresponds to Table 7.3 on p.261 of the text.

Now you have $SSR(X_1)$, $SSR(X_2 \mid X_1)$, and $SSR(X_3 \mid X_1, X_2)$ and their corresponding degress of freedom and mean squares. If you sum them together you get $SSR(X_1, X_2, X_3)$, which has 3 degrees of freedom. Divide $SSR(X_1, X_2, X_3)$ by 3 to get $MSR(X_1, X_2, X_3)$. To get $SSE(X_1, X_2, X_3)$, its degrees of freedom, and $MSE(X_1, X_2, X_3)$, use the line beginning with "Residuals." To calculate and store these in R, use the commands

```
> SSR <- sum( anova(Retailer)[1:3,2] )
> MSR <- SSR / 3
> SSE <- anova(Retailer)[4,2]
> MSE <- anova(Retailer)[4,3]
```

You can obtain alternate decompositions of the regression sum of squares into extra sum of squares by running new linear models with the predictors entered in a different order. For example, if we want $SSR(X_3)$, $SSR(X_1|X_3)$ and $SSR(X_2|X_1,X_3)$, we could try:

```
> Model2 <- lm( Hours ~ Holiday+Cases+Costs, data=Grocery)
> anova(Model2)
```

which gives us the ANOVA table we need:

```
> Model2 <- lm( Hours ~ Holiday+Cases+Costs, data=Grocery)
> anova(Model2)
Analysis of Variance Table

Response: Hours
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
Holiday    1 2077646 2077646 101.1913 2.086e-13 ***
Cases      1   92285   92285   4.4947   0.03920 *  
Costs      1    6675    6675   0.3251   0.57123    
Residuals 48  985530   20532                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If you want $SSR(X_2, X_3 \mid X_1)$, for example, use equation **(7.4b)** on p.260, which gives you $SSR(X_2, X_3 \mid X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$. This means you will have to run a linear model involving only X_1 to obtain $SSE(X_1)$. Other combinations of error sums of squares can be obtained by fitting reduced models (where you leave out one or two variables), obtaining the values for *SSE* or *SSR*, and applying the various formulae given in section 7.1 of the text.

Suppose we want to test whether one or more of the variables can be dropped from the original linear model (henceforth called the *full model*). The easiest way to accomplish this in R is to just run a new model that excludes the variables we are considering dropping (henceforth called the *reduced model*), then perform a **general linear test**. For instance, if we are considering dropping `Costs` (X_2) from the `Retailer` model, we run a reduced model which uses only the other two predictors `Cases` and `Holiday`:

```
> Reduced <- lm( Hours ~ Cases + Holiday, data=Grocery)
```

Then to perform the *F* test, just type

```
> anova(Reduced, Retailer)
```

to get the ANOVA comparison:

```
> Reduced <- lm( Hours ~ Cases+Holiday, data=Grocery)
> anova(Reduced, Retailer)
Analysis of Variance Table

Model 1: Hours ~ Cases + Holiday
Model 2: Hours ~ Cases + Costs + Holiday
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 992204                           
2     48 985530  1      6675 0.3251 0.5712
>
```

Note that the first argument to the `anova()` function must be the reduced model, and the second argument must be the full model (the one with all the original predictors). In this example, we are testing H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$, and, since $F^* = 0.3251$ is small, we obtain a very large *P*-value of 0.5712. So we would not reject H_0 at any reasonable level of significance, and thus conclude that the variable `Costs` can be dropped from the linear model. This agrees with the result of the *t*-test from the summary table, which had the same *P*-value, and in fact this value of F^* is the square of the value of t^* in the summary table. Likewise, this result agrees exactly with the second ANOVA table above. This test statistic may also be obtained using the all-purpose formula given by **(2.9)** on p.73 of the text.

Now suppose we want to test H_0: $\beta_2 = 0$, $\beta_3 = 600$ against its alternative. In this case the reduced model, corresponding to H_0, is $Y_i = \beta_0 + \beta_1 X_{i1} + 600X_{i3} + \varepsilon_i$, which may be rewritten as $Y_i - 600X_{i3} = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$. To obtain the reduced model in R, use the formulation:

```
> Reduced <- lm(Hours – 600*Holiday ~ Cases, data=Grocery)
```

However, you will get an error message if you attempt to use the `anova()` function to compare this model with the full model, because the two models do not have the same response variable. Instead, you will need to obtain the *SSE* for this reduced model, along with its degrees of freedom, from its ANOVA table, and the *SSE* from the full model, along with its degrees of freedom, from the ANOVA table for the full model, then calculate *F** using equation **(2.9)** in the textbook.

To obtain the coefficients of partial determination, you will need to use formulae like those in section 7.4. You may also need to run several different models, with the predictors in various different orders, in order to obtain values for the needed forms of *SSE* and the extra sums of squares.