How to Compute Typical F-tests from Splus Output

1 The ANOVA table

For the multiple regression model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \;\; \forall\, i = 1, \ldots, n$$

a standard "ANOVA" table shows the sums of squares and mean-squares necessary to compute an omnibus F-test of fit of the regression model.

The table looks like this

```
Source        df        SS         MS                          F
-----------------------------------------------------------------
Regression    k         RegSS      RegMS = RegSS/k             RegMS/RMS
Residuals     n-(k+1)   RSS        RMS   = RSS/(n-(k+1))
----------------------------
Total         n-1       TSS
```

and to test the hypotheses

$$\begin{aligned} H_1: &\quad Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon \\ H_0: &\quad Y = \alpha + \varepsilon \end{aligned}$$

we construct the F-statistic

$$F = \frac{RegSS/k}{RSS/(n-(k+1))}$$

which under H_0 will be distributed as $F(k, n-(k+1))$. In general we expect F to be near 1, so values of F much larger than one are bad news for H_0.

In Splus a typical ANOVA table is constructed using the `anova()` command, and it looks like this:

```
> fuel _ read.table("fuel.dat")
> fuel[1:4,]
    Pop Tax Nlic   Inc  Road Fuelc Dlic Fuel
ME 1029 9.0  540 3.571 1.976   557 52.5  541
NH  771 9.0  441 4.092 1.250   404 57.2  524
VT  462 9.0  268 3.865 1.586   259 58.0  561
MA 5787 7.5 3060 4.870 2.351  2396 52.9  414
> myfit _ lm(Fuel ~ Tax + Inc + Dlic,data=fuel)
> anova(myfit)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
Tax        1   55443.5  55443.5  7.51758 0.008674482
Inc        1   56250.5  56250.5  7.62700 0.008235139
Dlic       1  190469.6 190469.6 25.82576 0.000006674
Residuals 46  339258.2   7375.2
```


The RegSS term has been split into a separate sum of squares for each identified regressor in the problem. To recreate the "standard" ANOVA table, we sum these individual sums of squares back up. For reasons that will become clear later in the course, the order in which regressors are entered into the model affects the particular values of these individual sums of squares, but their total is always the same, RegSS. From the table above we get

```
Source        df        SS         MS                          F
-----------------------------------------------------------------
Regression    3         302163.6   100721.2                    13.65678
Residuals     46        339258.2   7375.178
----------------------------
Total         49        641421.8
```

and you can verify that the F statistic here agrees with the output from `summary(myfit)` (and $SE_{resid} = \sqrt{7375.2} = 85.88$ agrees also).

2 The Extra-Sum-of-Squares Principle (Incremental Sum of Squares)

Sometimes it is desirable to test to see if just some of the regression coefficients in a model are different from zero, instead of all of them. By suitably renumbering the X's we can arrange that the hypotheses to be compared are

$$\begin{aligned} H_1: &\quad Y = \alpha + \beta_1 X_1 + \cdots + \beta_{k-q} X_{k-q} + \cdots + \beta_k X_k + \varepsilon \\ H_0: &\quad Y = \alpha + \beta_1 X_1 + \cdots + \beta_{k-q} X_{k-q} + \varepsilon \end{aligned}$$

Note that the difference in the number of parameters in this model is q, and another way to state the null hypothesis is to say that the last q β's are equal to zero:

$$H_0: \;\; \beta_{k-q+1} = \cdots = \beta_k = 0$$

It turns out that there is an F test for this comparison also. The "extra sum of squares" principle (or "incremental sum of squares" principle) says that the relevant F test is

$$F = \frac{(RegSS_{big} - RegSS_{small})/q}{RSS_{big}/(n-k-1)} = \frac{(RSS_{small} - RSS_{big})/q}{RSS_{big}/(n-k-1)}$$

where the "big" model is the H_1 model and the "small" model is the H_0 model, and the switch from RegSS to RSS follows from the equation

$$TSS = RegSS + RSS$$

applied to both the "big" and the "small" model.

There are three ways to compute this F statistic using the `anova()` command in Splus.

- By looking at RSS and RMS values in `anova()` applied to each model.

```
> big _ lm(Fuel ~ Tax + Inc + Dlic,data=fuel)
> small _ lm(Fuel ~ Tax,data=fuel)
> #
> #  Test to see if the coefficients of Inc and Dlic can be set to zero
> #
> anova(big)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
Tax        1   55443.5  55443.5  7.51758 0.008674482
Inc        1   56250.5  56250.5  7.62700 0.008235139
Dlic       1  190469.6 190469.6 25.82576 0.000006674
Residuals 46  339258.2   7375.2
> anova(small)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
Tax        1   55443.5 55443.48 4.541614 0.03822929
Residuals 48  585978.3 12207.88

> (585978.3 - 339258.2)/2 ) / (339258.2 / 46 ) # F = ...
[1] 16.72638
```

- By computing RegSS for the smaller model by arranging to have the terms by which the models differ appear last in the `anova()` output.

```
> anova(big)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
Tax        1   55443.5  55443.5  7.51758 0.008674482
Inc        1   56250.5  56250.5  7.62700 0.008235139
Dlic       1  190469.6 190469.6 25.82576 0.000006674
Residuals 46  339258.2   7375.2

> ( (56250.5 + 190469.6)/2 ) /  (339258.2 / 46 ) # F again...
[1] 16.72638
```


- By using `anova(model1,model2)` to compare the models directly.

```
> anova(big,small)
Analysis of Variance Table

Response: Fuel

             Terms Resid. Df      RSS      Test Df Sum of Sq  F Value
1 Tax + Inc + Dlic        46 339258.2
2              Tax        48 585978.3 -Inc-Dlic -2 -246720.1 16.72638
         Pr(F)
1
2 3.474748e-06
```