Regression analysis Module 15: Advanced interactions

(The attached PDF file has better formatting.)

Selecting the optimal model using sums of squares and degrees of freedom (*F* test)

- Tables 7.1 and 7.2 on page 139 are tested on the final exam.
- This posting explains the computations for the *F* test in these tables.

The variables are: I = income, E = education, and T = type

The regression sums of squares are

Model	*Terms*	*Sum of Squares*	*df*
1	I, E, T, I × T, E × T	24,794	8
2	I, E, T, I × T	24,556	6
3	I, E, T, E × T	23,842	6
4	I, E, T	23,666	4
5	I, E	23,074	2
6	I, T, I × T	23,488	5
7	E, T, E × T	22,710	5

For each model,

- The residual sum of squares is $\sum\left(Y-\hat{Y}\right)^2$.
- The regression sum of squares is $\sum\left(\bar{Y}-\hat{Y}\right)^2$.
- The total sum of squares is $\sum\left(\bar{Y}-Y\right)^2$.

The total sum of squares does not depend on the model; it is 28,347 in this illustration.

Jacob: All three formulas for the sums of squares use only Y values, not X value or ß's.

Rachel: The regression sum of squares and the residual sum of squares use the fitted Y values, which depend on the X values. They vary by model.

The degrees of freedom in Table 7.1 on page 139 are the number of explanatory variables in the model (k). The degrees of freedom are actually N-k-1. This illustration shows the degrees of freedom for the numerator of the F test, which is the difference in the number of variables in the full vs reduced models. N-1 is the same for all models, so it drops out of the difference.

For the number of explanatory variables:

- I and E are one explanatory variable each.
- T, I × T, and E × T are two explanatory variables each.

Table 7.2 shows the degrees of freedom and sum of squares in the numerator of the F test.

Source	*Models*	*Sum of Squares*	*df*	*F*
Income	3 – 7	1,132	1	28.35
Education	2 – 6	1,068	1	26.75
Type	4 – 5	592	2	7.41
Income × Type	1 – 3	952	2	11.92
Education × Type	1 – 2	238	2	2.98
Residuals		3,553	89	
Total		28,347	97	

The total sum of squares is 28,347. The sample has 98 data points, so the total sum of squares has 98 – 1 = 97 degrees of freedom.

The full model (Model 1) has a regression sum of squares of 24,794, so it has a residual sum of squares of 28,347 – 24,794 = 3,553. This residual sum of squares has 98 – 8 – 1 = 89 degrees of freedom.

The denominator of the F ratio (for all tests) is 3,553 / 89 = 39.921.

Illustration: To test the significance of income, we contrast models 3 and 7.

The sum of squares is 23,842 for Model 3 and 22,710 for Model 7. The difference in the sum of squares is 23,842 – 22,710 = 1,132.

Model 3 has 6 explanatory variables and Model 7 has 5 explanatory variables. The degrees of freedom in the numerator of the F test is 6 – 5 = 1.

- The numerator of the F ratio is 1,132 / 1 = 1,132.
- The F ratio is 1,132 / 39.921 = 28.356.

Illustration: To test the significance of education × type, we contrast models 1 and 2.

The sum of squares is 24,794 for Model 1 and 24,556 for Model 2. The difference in the sum of squares is 24,794 – 24,556 = 238.

Model 1 has 8 explanatory variables and Model 2 has 6 explanatory variables. The degrees of freedom in the numerator of the F test is 8 – 6 = 2.

- The numerator of the F ratio is 238 / 2 = 119.
- The F ratio is 119 / 39.921 = 2.981.

To find the *p*-values in Table 7.2, use a table of the F-distributions or statistical software, such as Excel. If an exam problem asks for a *p*-value, it will give a table.