Module 15: Advanced interactions

Fox Module 15 Practice problems on F test

(The attached PDF file has better formatting.)

Fox forms regression models and infers significance for both continuous and discrete explanatory variables. He uses the Canadian prestige data: prestige is regressed on two quantitative explanatory variables (income and education) and one discrete explanatory variables (type of occupation).

Fox chooses this illustration because the explanatory variables are correlated, and the partial effects derived from the multiple regression differ from the marginal effects.

- Income is correlated with education: high vs low income track high vs low education.
- Type of occupation is correlated with both income and education: Professional workers have (on average) higher income and education than white collar or blue collar workers.

These correlations cause the partial effects to differ from the marginal effects.

Illustration: Prestige is higher for professional workers than for blue collar workers. This is a marginal effect; we don't know if the type of occupation is the causative factor. Perhaps professional workers have higher prestige than blue collar workers because they have higher incomes and more education.

Illustration: Prestige is positively correlated with both income and education. These are marginal effects; we must examine whether both explanatory variables affect prestige or only one of them does.

The partial effects hold all other variables constant. They ask: *If we take professional, white collar, and blue collar occupations with the same income and education, which has the highest or lowest prestige?* Since the correlations of occupation type with income and prestige are positive and the correlations of prestige with income and education are positive, the partial effects of occupation type are less than the marginal effects.

Statistical inference tests if the partial effects are significant. Fox shows how to use *F*-tests to evaluate these partial effects.

Illustration of interaction effects: An actuary examines the effects of sex of the driver and mileage driven on personal auto claim frequency. Men have higher average claim frequencies than women, and more miles driven leads to higher claim frequency. These are marginal effects. If men and women drive different average miles a week, the partial effects differ from the marginal effects. Partial effects ask: *For men and women who drive the same number of miles, what are the relative claim frequencies?*

Jacob: Simple linear regression shows the marginal effects; multiple regression shows the partial effects. If the ordinary least squares estimate of the coefficient is positive, then the explanatory variable has a positive effect on the response variable (and vice versa if it is negative).

Rachel: Random fluctuations cause any estimate to differ from zero. We must test if the difference stems from random fluctuations or reflects the influence of the explanatory variable: that is, we test the significance of the effect.

Jacob: To test significance, we use *t* values. Earlier modules show how to form *t* values for both simple and multiple regression.

Rachel: If the explanatory variables are all continuous and they have no interaction effects, we use *t* values. For factors what take discrete values and for explanatory variables that interactions, we use incremental *F* tests. This module deals with testing for significance with factors and interaction effects.

ANOVA (analysis of variance) for the Canadian prestige data set.

This posting explains Fox's ANOVA (analysis of variance) for the Canadian prestige data set and then shows the format of the final exam problems for this module.

Fox's explanation is difficult for some candidates. The explanations and dialogues in this posting clarify the analysis of variance.

The data are 98 occupations classified along three dimensions:

- The type of occupation (T): professional / managerial; white collar; and blue collar.
- The average education (E) of workers in the occupation.
- The average income (I) of workers in the occupation.

Fox examines the 7 regression models listed in the table below. He computes the regression sum of squares for each model (not shown explicitly in the chapter) and uses the RegSS to infer the significance of the main effects and the interaction effects.

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

A. What are the major effects in these models? What are the linear regression equations using only one explanatory variable at a time (to give marginal effects)?
B. What are the interaction effects?
C. How many degrees of freedom does the total sum of squares (TSS) have?
D. How many degrees of freedom does model #5 (*I, E*) have?
E. How many degrees of freedom does model #4 (*I, E, T*) have?
F. How many degrees of freedom does model #6 (*I, T, I × T*) have?
G. How many degrees of freedom does model #2 (*I, E, T, I × T*) have?
H. How many degrees of freedom does model #1 (*I, E, T, I × T, E × T*) have?

Part A: The major effects are income (I), education (E), and type of occupation (T).

- Income has one parameter, β_1. A linear regression of prestige on income is $Y = \alpha + \beta_1 \times X_{inc} + \epsilon$.
- Education has one parameter, β_2. A linear regression of prestige on education is $Y = \alpha + \beta_2 \times X_{edu} + \epsilon$.

(The explanatory variables differ in the two equations above: one is income and one is education.)

Type of occupation has two parameters: γ_1 and γ_2. A linear regression of prestige on type of occupation is

$$Y = \alpha + \gamma_1 \times D_1 + \gamma_2 \times D_2 + \epsilon.$$

Jacob: What are D_1 and D_2?

Rachel: Type of occupation is a factor: it take one of three discrete values. For a factor with N levels, we need N-1 regressors, called D_1,D_2, … .

- The explanatory variable T has three levels. Every occupation is in one of the three levels (professional, white collar, or blue collar).
- We need only two regressors, such as one for white collar and one for blue collar. If an occupation is not white collar or blue collar, it is professional.

Jacob: What the values of these regressors?

Rachel: The values depend on the type of regressor. Fox uses two types of regressors. Equation 7.2 on page 125 shows the most common method:

Category	D_1	D_2
Professional	1	0
White Collar	0	1
Blue Collar	0	0

- D_1 = 1 for professional occupations and 0 for white collar and blue collar occupations.
- D_2 = 0 for professional occupations,1 for white collar occupations, and 0 for blue collar occupations.

A second type of regressor (called “sigma” or “zero-sum” regressors) is discussed in later modules. Both types of regressors are tested on the final exam.

Jacob: Does it matter which types of occupation get the one’s and which types get zeros?

Rachel: If the hypotheses are properly formulated, the regression results will be the same, though the meaning of the coefficients and their values will differ.

For quantitative explanatory variables, we test whether the β coefficient differs from zero (or from the value in the null hypothesis). If we regress motor insurance claim frequency on distance driven, the null hypothesis (that distance driven does not affect claim frequency) implies that β = zero.

For factors (qualitative explanatory variables), we test whether differences between levels are significant. If we regress motor insurance claim frequency on sex (male/female), we test whether men and women differ. The absolute value of the difference does not depend on which sex is the base class.

The meaning of the coefficient is the difference from the base class, so it depends on which is the base class.

Illustration: Suppose territory has three classes: urban vs suburban vs rural.

- If urban is the base class, the β for suburban is the suburban claim frequency minus the urban frequency.
- If rural is the base class, the β for suburban is the suburban claim frequency minus the rural frequency.

If sub-urban = rural and both differ from urban, the β for sub-urban alone depends on the base class. But the proper test is whether the three levels differ, not whether sub-urban is the same as rural.

Degrees of freedom

- The degrees of freedom for quantitative explanatory variables is one.
- The degrees of freedom for factors (qualitative explanatory variables) is the number of level minus one.

Jacob: This seems counter-intuitive. Income and education are powerful indicators of prestige. They divide the observations finely: actual income can be any positive number. Type of occupation has only three values. Yet income and education have one degree of freedom, and type of occupation has two degrees of freedom.

Rachel: The degrees of freedom says how much random fluctuations affect the results. If income does not affect prestige, the one β causes a small apparent affect. If type of occupation does not affect prestige, the to dummy variables have a greater effect on prestige.

Intuition: For degrees of freedom, do not think of the observed relation of the explanatory variable with the response variable. Assume the explanatory variable is independent of the response variable, and consider how the number of regressors may show an apparent effect from random fluctuations.

Degrees of freedom for factors is the number of levels minus one.

- If we regress personal auto claim frequency on age of the driver, we have one parameter.
- If we group the ages into youthful, adult, and retired, we have two parameters.

Jacob: Why not three parameters for the grouping into youthful, adult, and retired?

Rachel: If the drive is not youthful or adult, the driver is retired. We need only two parameters: one for youthful and one for adult.

Part B: The interaction effects are income by type ($I \times T$) and education by type ($E \times T$).

Jacob: If we have education and type, what does education by type add?

Rachel: The education explanatory variable may impliy that each additional unit of education adds 4 units of prestige. But this relation may differ by type of occupation. Each additional unit of education may add 3.1 units of prestige for professional workers, 6.0 units for white collar workers, and 1.7 units for blue collar workers.

Jacob: How many parameters does the interaction of education by prestige add?

Rachel: The interaction of education by prestige adds two parameters. Let the two parameters be white collar by education and blue collar by education.

- Education alone adds 3.1 units of prestige for each unit of education.
- White collar by education adds 6.0 – 3.1 = 2.9 units of prestige for each unit of education.
- Blue collar by education adds 1.7 – 3.1 = –1.4 units of prestige for each unit of education.

Intuition: The interaction of a factor with a quantitative explanatory variable adds as many degrees of freedom as the factor alone adds. Fox calls these parameters δ_{11} and δ_{12} for the interaction of income and type and δ_{21} and δ_{22} for the interaction of education and type.

Jacob: The way you set this up works only if we also have education as a separate explanatory variable.

Rachel: That is Fox's *principle of marginality*. If we use education by type, we should also use education alone and type alone.

Jacob: You explained why we must include education alone. Why do we include type of occupation alone?

Rachel: The intercept differs by type of occupation. The intercept may be 17.63 for professional workers, –31.26 for white collar workers, and 2.276 for blue collar workers.

Jacob: Do we have three parameters for type of occupation?

Rachel: We also have an intercept parameter, so two parameters for type of occupation. Let the intercept parameter α be 17.63, and the two parameters for type of occupation are white collar and blue collar. For white collar workers, the parameter is –31.26 – 17.63 = -48.89 and for blue collar workers, the parameter is 2.276 – 17.63 = -15.35.

Part C: The data have 98 occupations, so 98 data points. The total sum of squares has one parameter (the overall mean), so it has 97 degrees of freedom.

Part D: Model #5 has two explanatory variables, income and education. Each explanatory variable adds one parameter, so the regression model has two explanatory variables and one intercept. 98 data points minus three parameters gives 95 degrees of freedom. The three parameters are α, β_1, β_2.

Fox presents this in slightly different words. The total sum of squares has 97 degrees of freedom. Income and education each add one parameter, so Model #5 has 97 – 2 = 95 degrees of freedom.

In Table 7.1, Fox shows the degrees of freedom for Model #5 as 2. Model #5 adds two parameters, so it *reduces* the degrees of freedom by 2.

Jacob: You said Model #5 has three parameters (if we include the intercept).

Rachel: The total sum of squares also has an intercept. Model #5 *adds* two more parameters.

Part E: Model #4 (*I, E, T*) has 97 – 1 – 1 – 2 = 93 degrees of freedom. In Fox's terms, Model #4 adds 4 parameters, so it has four fewer degrees of freedom. The 4 + 1 = 5 parameters are α, β_1, β_2, γ_1, γ_2.

Part F: Model #6 (*I, T, I* × *T*) has 97 – 1 – 2 – 2 = 92 degrees of freedom. In Fox's terms, Model #6 adds 5 parameters, so it has five fewer degrees of freedom. The 5 + 1 = 6 parameters are α, β_1, γ_1, γ_2, δ_{11}, δ_{12}.

Part G: Model #2 (*I, E, T, I* × *T*) has 97 – 1 – 1 – 2 – 2 = 91 degrees of freedom. Model #2 adds 6 parameters, so it has 6 fewer degrees of freedom. The 6 + 1 = 7 parameters are α, β_1, β_2, γ_1, γ_2, δ_{11}, δ_{12}.

Part H: Model #1 (*I, E, T, I* × *T, E* × *T*) has 97 – 1 – 1 – 2 – 2 – 2 = 89 degrees of freedom. Model #1 adds 8 parameters, so it has 8 fewer degrees of freedom. The 8+1=9 parameters are α, β_1, β_2, γ_1, γ_2, δ_{11}, δ_{12}, δ_{21}, δ_{22}.

Jacob: Can we compare Model #6 with Model #7 to see whether income or education is a better predictor?

Rachel: The *F*-test is used for *nested* models: one model has all the terms of the other model plus some additional terms. We can not compare models with different terms. One model may have a higher *t*-value but a weaker effect of its explanatory variable.

Table 7.1 from page 139 is reproduced below. The *df* column is the number of additional parameters (that is, not including the intercept α) in the model. We use this table for the next exercise (the analysis of variance).

Model	*Terms*	*Parameters*	*Regression Sum of Squares*	*df*
1	*I, E, T, I × T, E × T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	24,794	8
2	*I, E, T, I × T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	24,556	6
3	*I, E, T, E × T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	23,842	6
4	*I, E, T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$	23,666	4
5	*I, E*	α, β_1, β_2	23,074	2
6	*I, T, I × T*	$\alpha, \beta_1, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	23,488	5
7	*E, T, E × T*	$\alpha, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	22,710	5

Jacob: What is the regression sum of squares?

Rachel: The models give fitted prestige values for each occupation. The overall average prestige value for all occupations combined does not differ by model. For each model, the regression sum of squares is the sum of the squares of (the fitted value by occupation minus the overall average).

Take heed: Final exam problems are modeled on the exercise below. If you understand the solution to this exercise, you can solve the final exam problems.

Marginal effects vs partial effects and main effects vs interactions

Jacob: Fox distinguishes marginal effects from partial effects and main effects from interactions. How do the two pairs differ? Is a main effect like a marginal effect and an interaction like a partial effect?

Rachel: The two dichotomies are not related to each other.

- Marginal effects vs partial effects deals with correlations among explanatory variables.
- Main effects vs interactions deals with interactions of two explanatory variables on the response variable.

A motor insurance example shows the difference. Motor insurance claim frequency depends on the distance driven (a quantitative explanatory variable) and the driver's attributes (qualitative factors such as sex, age, residence, and credit score). We consider distance and gender here, though many other explanatory variables affect claim frequency.

- More distance driven increases the expected number of claims. We assume the relation is additive: claim frequency = $\alpha + \beta \times$ distance.
- Male drivers have more accidents than female drivers.

Suppose all men drive more than 10,000 kilometers a year, and all women drive less than 10,000 kilometers a year. Male vs female is correlated with distance driven, so the partial effects (of sex or distance driven) do not equal the marginal effects (of sex or distance driven). In fact, some people believe that the differences in claim frequency by sex of the driver reflect primarily the average distance driven by men vs women.

Interactions refer to the effects on the response variable, not the correlation of the explanatory variables. For the average driver, each 1,000 kilometers of annual driving may increase the annual claim frequency by one percentage point: claim frequency = $\alpha + 0.02 \times$ distance (in thousands of kilometers). But the β parameter (of 0.02 here) may differ for men vs women: it might be 0.01 for women and 0.03 for men. This is an interaction: whether or not sex and distance driven are correlated, the β parameter for distance driven differs for men vs women.

Later modules show how generalized linear models with log-link functions deal with these interactions. This module uses classical regression analysis with additional parameters for the interaction effects.

** Exercise 15.1: Analysis of variance

We have computed Table 7.1 (the table above) for the Canadian prestige data. We want to know if income by occupation (that is, the interaction of income and type of occupation) affects prestige.

Jacob: What do you mean by the interaction of income and type of occupation?

Rachel: If prestige increases by one point for each 10,000 of annual income *for all types of occupation*, the regression has no interaction. If prestige increases by one point for each 10,000 of annual income for blue-collar workers, by two points for each 10,000 of annual income for white-collar workers, and by three points for each 10,000 of annual income for professional workers, then income and type of occupation interact.

Similarly, we may want to know if income (a main effect) affects prestige. That is, does income affect prestige for all types of occupation combined, even if there are no differences by type of occupation.

A. What two models do we compare for each test: the main effect of income and the interaction of income by type of occupation?
B. What is the difference in the regression sum of squares?
C. What are the degrees of freedom in the numerator of the *F*-ratio?
D. What is the residual sum of squares?
E. What are the degrees of freedom in the denominator of the *F*-ratio?
F. What is the *F*-ratio?

Jacob: For the main effect of income, why not compare a model with just an intercept against a model using just income?

Rachel: We want the *partial* effect of income and the partial effect of income by type of occupation, *not* the *marginal* effect of income or the marginal effect of income by type of occupation.

- We compare the full model *including all parameters* vs a model *excluding the effect of income* by type of occupation to test the significance of the interaction of income by type of occupation.
- We compare the full model *including all parameters* excluding the interaction of income by type of occupation vs a model *excluding the man effect of income* and the *interaction effect of incomeby type of occupation* to test the significance of the main effect of income.

Jacob: The full model is Model #1, with terms *I, E, T, I* × *T, E* × *T*. The table is missing some models, such as *E, T, I* × *T, E* × *T*. To test the main effect of income, why not test the full Model #1 against the model using the terms *E, T, I* × *T, E* × *T*?

Rachel: Fox's *principle of marginality* says we don't include the interaction of income and type (*I* × *T*) unless we include the major effect of income and type. The model *E, T, I* × *T, E* × *T* contradicts the principle of marginality. The model without income is *E, T, E* × *T*, so the model with income is *I, E, T, E* × *T*. We compare model #3 (with income) to model #7 (without income) to test the main effect of income.

Jacob: Why not compare model #1 (*I, E, T, I* × *T, E* × *T*) with model #7? This compares a model with all the effects of income against a model with no effects of income?

Rachel: Fox does this in two parts. Model #3 vs Model #7 tests the major effect of income. Model #1 vs Model #3 tests the interaction of income with type. Model #1 vs Model #7 tests the combined major effect of income plus the interaction of income and type. When Fox says *test the effect of income*, he means the major effect of income, not the interaction effect.

Take heed: The final exam problems follow Fox's procedure. The exam problem may say: "What is the *F*-ratio for the main effect of income? What is the *F*-ratio for the interaction effect of income with type of occupation?" Know which models Fox compares.

Part B: The difference in the regression sum of squares is 23,842 – 22,710 = 1,132. The difference in the regression sum of squares appears in the *numerator* of the *F*-ratio.

Jacob: What if the exam problem gives the residual sum of squares for each model?

Rachel: All models have the same total sum of squares, since all models use the same set of y-values (the response variable). The difference in the regression sum of squares is the negative of the difference in the residual sum of square.

Take heed: In this chapter, Fox uses the RegSS to mean the regression sum of squares for that model, and the residual sum of square means the residual sum of square for the full model.

Part C: The degrees of freedom in the numerator of the *F*-ratio is the difference in the number of parameters in the two models. Model #3 has 6 parameters (besides the intercept); Model #7 has 5 parameters. The degrees of freedom in the *numerator* of the *F*-ratio is 6 – 5 = 1.

Jacob: Can we use an *F* test to compare Model #5 (I, E = 2 degrees of freedom) with Model #6 (I, T, I × T = 5 degrees of freedom)? Can we use an *F* test to compare Model #2 (I, E, T, I × T = 6 degrees of freedom) with Model #7 (E, T, E × T = 5 degrees of freedom)?

Rachel: No: the *F* test compares nested models: one model must be nested in the other. The nested model excludes one or more explanatory variables used in the larger model, but it does not include any explanatory variables not used in the larger model.

Model	*Terms*	*Parameters*	*Regression Sum of Squares*	*df*
1	*I, E, T, I × T, E × T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$	24,794	8
2	*I, E, T, I × T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	24,556	6
3	*I, E, T, E × T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	23,842	6
4	*I, E, T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2$	23,666	4
5	*I, E*	α, β_1, β_2	23,074	2
6	*I, T, I × T*	$\alpha, \beta_1, \gamma_1, \gamma_2, \delta_{11}, \delta_{12}$	23,488	5
7	*E, T, E × T*	$\alpha, \beta_2, \gamma_1, \gamma_2, \delta_{21}, \delta_{22}$	22,710	5

Jacob: Why can't we test two models that are not nested to see which explains the observations better?

Rachel: Statistics focuses on the significance of observed results. If the two models are not nested and the F value is not not significant, we can't say whether the parameters of both models are not significant or the parameters of both models are significant and are explaining different aspects of the response variable.

Illustration: The residence of the driver (urban, rural, or suburban) and the age of the driver (youthful, adult, or retired) both influence motor insurance loss costs. Comparing these two models (residence vs age) does not say anything about the significance of either explanatory variable.

Residual sum of squares

Part D: The residual sum of squares is the sum of squares not explained by the regression equations. The total sum of squares is 28,347.

Jacob: For the residual sum of squares in the denominator of the *F*-ratio, do we use 28,347 – 23,842 = 4,505 or 28,347 – 22,710 = 5,637? That is, do we use Model #3 or Model #7?

Rachel: We use Model #1, not Model #3 or Model #7. Fox emphasizes this point; use his method for the final exam problems. The residual sum of squares is 28,347 – 24,794 = 3,553. (We explain Fox's rationale below.)

Part E: The degrees of freedom in the denominator of the *F*-ratio is the number of data points (98) minus the number of parameters in Model #1 (9 including the intercept): 98 – 9 = 89. Alternatively, it is the degrees of freedom in the total sum of squares (97) minus the *additional* parameters in Model #1: 97 – 8 = 89.

Jacob: You say *Fox's* principle of marginality and you stress *Fox's* computation of the residual sum of squares.

Rachel: Not all statisticians agree with Fox. Fox uses the same residual sum of squares for all the *F* tests. Some other statistics texts use the residual sum of squares for the larger model in this *F*-test.

- Fox uses the residual sum of squares for Model #1 in all the *F* tests.
- Some other statisticians use the residual sum of squares for Model #3 to test the main effect of income, the residual sum of squares for Model #2 to test the main effect of education, and the residual sum of squares for Model #4 to test the main effect of type of occupation.

Jacob: Fox's chapter on the incremental *F* test uses the sums of square for the two models, not the residual sum of square for a model including other explanatory variables. What is the rationale for Fox's treatment in this chapter?

Rachel: Fox examines the effects of each explanatory variable given that all other explanatory variables are used, but he excludes interaction effects if the main effect is being tested.

The final exam problems follow Fox's textbook. The F test *for the analysis of variance uses the residual sum of square for the full model, even when a main effect is being tested.*

The F test

Part G: The *F*-ratio is (1,132 / 1) / (3,553 / 89) = 28.35576.

Jacob: What does the *F*-ratio mean?

Rachel: The *F*-ratio gives a *p*-value, which is the likelihood that one might get an *F*-ratio this large or larger by random fluctuations even if the explanatory variable being tested has no effect on the response variable.

The *p*-value depends on the degrees of freedom in the numerator and denominator of the *F*-ratio. Excel has a built-in function to give the *p*-value. Older statistics books often showed tables of *F* ratios by degrees of freedom with their *p* values.

<<** for intuition: with one explanatory variable (&slregr), *F* test = square of t value.

For large data sets, the t distribution is approximated by the normal distribution. A Z value of 2 is significant for a 95% two-sided confidence interval, so an F value of 4 is a lower bound for significance at a 95% confidence interval with one degree of freedom in the numerator.

Fox's Table 7.2 from page 139 is reproduced below. To prepare for the final exam problems, be sure you understand each column.

Source	*Models*	*Sum of Squares*	*df*	*F*
Income	3 – 7	1,132	1	28.356
Education	2 – 6	1,068	1	26.753
Type	4 – 5	592	2	7.415
Income × *Type*	1 – 3	952	2	11.923
Education × *Type*	1 – 2	238	2	2.981
Residuals		3,553	89	
Total		28,347	97	

Source is the explanatory variable that is being tested. For example,

- Source = Education means we test the *main effect* of education: the full model without education vs the full model with education. By the principle of marginality, the full model does not include the interaction of education with type of occupation, since we cannot include the interaction unless we include the main effect. Models 2 and 6 both exclude the interaction of E × T.

- Source = Education × Type means we test the interaction effect of education by type of occupation. Both the full model and the nested model include the main effects of education and of type of occupation. Models 1 and 2 both include the main effects E and T, but only model 1 includes the interaction E × T.

Models is the models being compared. if we don't include the major effect of education, we don't include the interaction effect of education with type, so both models exclude *E* × *T*. We compare Model #2 (with education) vs Model #6 (without education).

Sum of squares is the difference in the regression sum of squares of the models being compared.

The column *df* is the *difference* in degrees of freedom of the models being compared. The degrees of freedom for each model depends on its parameters. The final exam problems give the explanatory variables and the number of levels for factors (such as 2 for gender and 3 for urban vs rural vs suburban). You must derive the degrees of freedom in the denominator of the *F* test.

F is the *F*-ratio. The denominator of *every F*-ratio is 3,553 / 89. The numerator of the *F*-ratio is the difference in the regression sum of squares divided by the difference in the degrees of freedom.

The regression sum of squares for the full model is 24,794, and the total sum of squares is 28,347. Note that

$$24{,}794 + 3{,}553 = 28{,}347, \text{ or } 28{,}347 - 24{,}794 = 3{,}553$$

The final exam problem may give the total sum of squares and test a main effect (income, education, or type of occupation). Using the notation in the table above, the regression sums of squares come from Models 2,3, 4, 5, 6, and 7, but the residual sum of square comes from Model #1.

Jacob: What does it mean to test *Income* × *Type*?

Rachel: We compare the full Model #1 with Model #3, which excludes only *Income* × *Type*. Both of these models include *income*.

Jacob: Can we test the combined effect of *Income* and *Income* × *Type*?

Rachel: Yes; we compare Model #1 with Model #7. Fox tests main effects separately from interaction effects. He believes it is worthwhile to test the significance of the two pieces separately, but the principle of marginality does not prohibit testing the two parts together.

Jacob: Do the final exam problems ask if the *F*-ratio is significant?

Rachel: It is worth knowing the approximate *p*-value for any given *F*-ratio. The exact *p*-value for a given *F*-ratio depends on the degrees of freedom in the numerator and denominator. Some statistics textbooks give tables of the *F*-distribution. Now you will use Excel (or some other software) for practical work.

Understand the implications of each piece of the *F* test

The numerator uses the difference in the regression sum of squares of the two models. If this difference is large, one model explains more of the variation in the response variable than the other model does, so the *F* test is significant. If this difference is small, the extra parameters in the larger model are not contributing much explanatory power, so the *F* test indicates that they are not significant.

The regression sum of squares depends on the number of observations and the units of measurement. A meter is 100 times a centimeter, so a meter squared is 10,000 times a centimeter squared. If a statistician switches from centimeters to meters, each regression sum of squares is multiplied by 10,000. The same is true for the total sum of squares and the residual sum of square. Dividing by the residual sum of square in the denominator of the *F* test eliminates the distortions from units of measurement.

The RSS in the denominator shows the value of the difference of the regression sum of squares.

Illustration: Suppose one model is the full model and the other model does not use one explanatory variable.

Jacob: What should we expect for the final exam problems?

Rachel: The final exam problems will be like the exercise here. The problem will give one or more continuous variables and one discrete variable with two or more levels. For example, the response variable might be personal auto claim frequency, and the explanatory variables might be territory with three levels (urban, rural, and suburban) and distance driven (a continuous variable).

The final exam problem will give the regression sum of squares for each model, the total sum of squares, and the number of data points. It will ask: what is the *F*-ratio for the effect of distance driven or of territory or the interaction of distance driven with territory?

The final exam problems may use other scenarios, such as life insurance or property insurance, with discrete factors and quantitative variables for each. If you can reproduce Fox's analysis of variance for the Canadian prestige data, you can solve the final exam problems.

Jacob: Why are we concerned with the degrees of freedom? We want to know which model has the greatest regression sum of squares. Why not just pick the model with the greatest regression sum of squares?

Rachel: Suppose we an explanatory variable called occupation with 97 levels for the first 97 occupations. If the occupation is not one of the first 97, it is the 98^{th}. This regression model explains all the variance in the data. Its regression sum of squares = the total sum of squares and its $R^2 = 1$. But this model is useless. It has 97 occupation parameters and one intercept, so its degrees of freedom = 98 – 97 – 1 = 0.

Jacob: Who would use such a model?

Rachel: In past generations, actuaries often used such model. For example, workers' compensation pricing used 650+ classes with separate rates for each. The experience data were the loss costs by class, and the model generated pure premiums by class. The model generated results, but since the number of parameters equaled the number of observations, one could not test the significance of the pricing results.