Chapter 7: Dummy Variable Regression

7.1 A Dichotomous Factor

Common slope model with a binary factor

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i$$

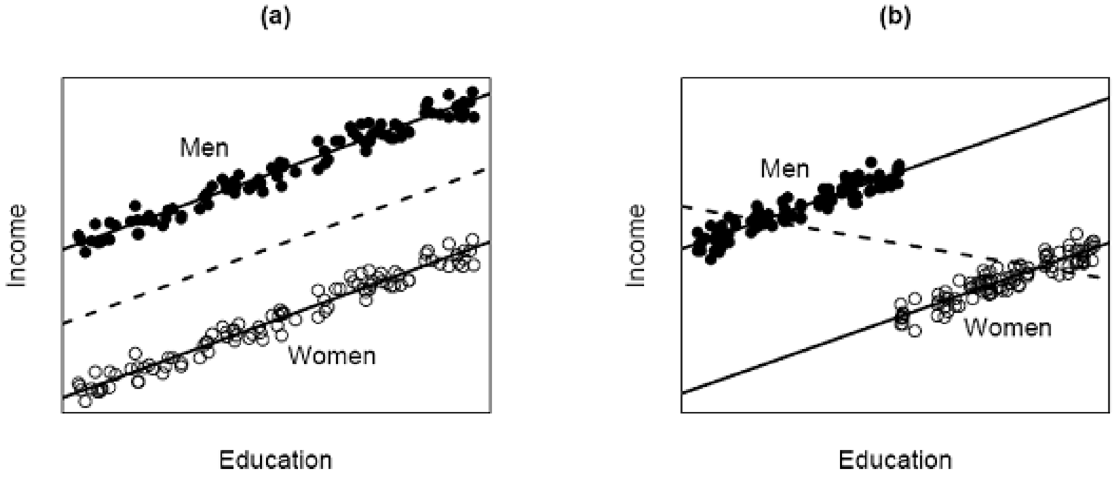
where

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

So, for men: $Y_i = \alpha + \beta X_i + \gamma \cdot 1 + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$

for women: $Y_i = \alpha + \beta X_i + \gamma \cdot 0 + \varepsilon_i = (\alpha) + \beta X_i + \varepsilon_i$

Figure 7.1



The additive dummy variable regression model.

The geometric view of the multiple regression on one quantitative and one binary regressor.
(Fox: “the geometric ‘trick’, as the linear regression plane is defined only at D=0 and D=1)

Figure 7.2

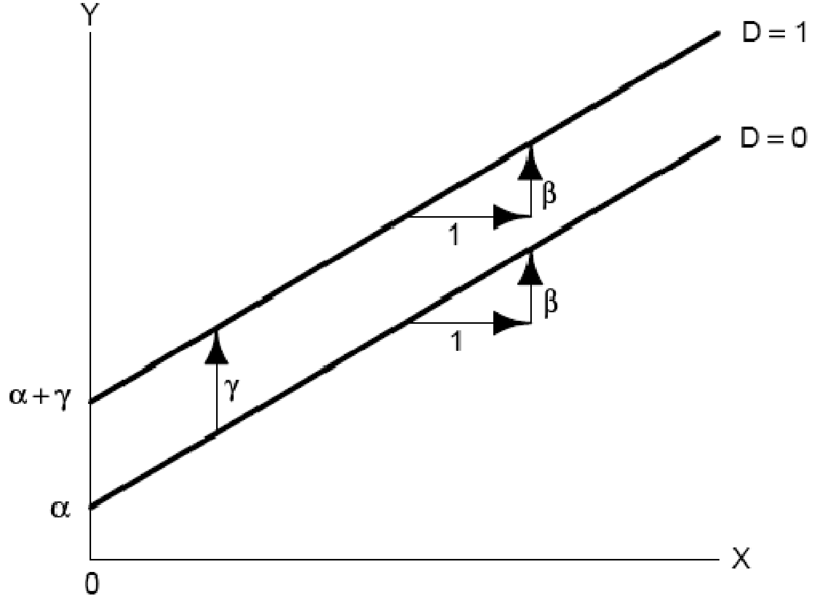
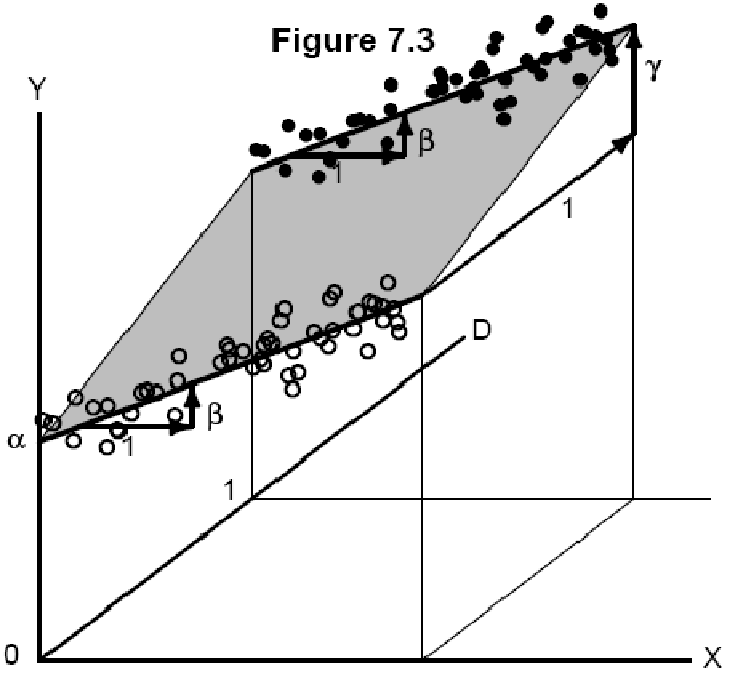


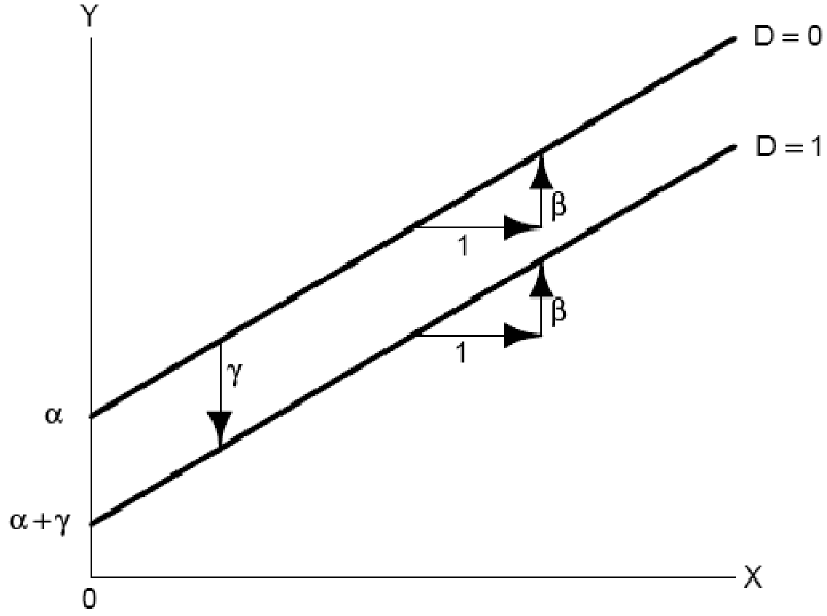
Figure 7.3



Changing the "reference category" so that

$$D_i = \begin{cases} 0 & \text{for men} \\ 1 & \text{for women} \end{cases}$$

Figure 7.4



We'll see additional approaches to coding factors in R.

7.2 Polytomous Factors

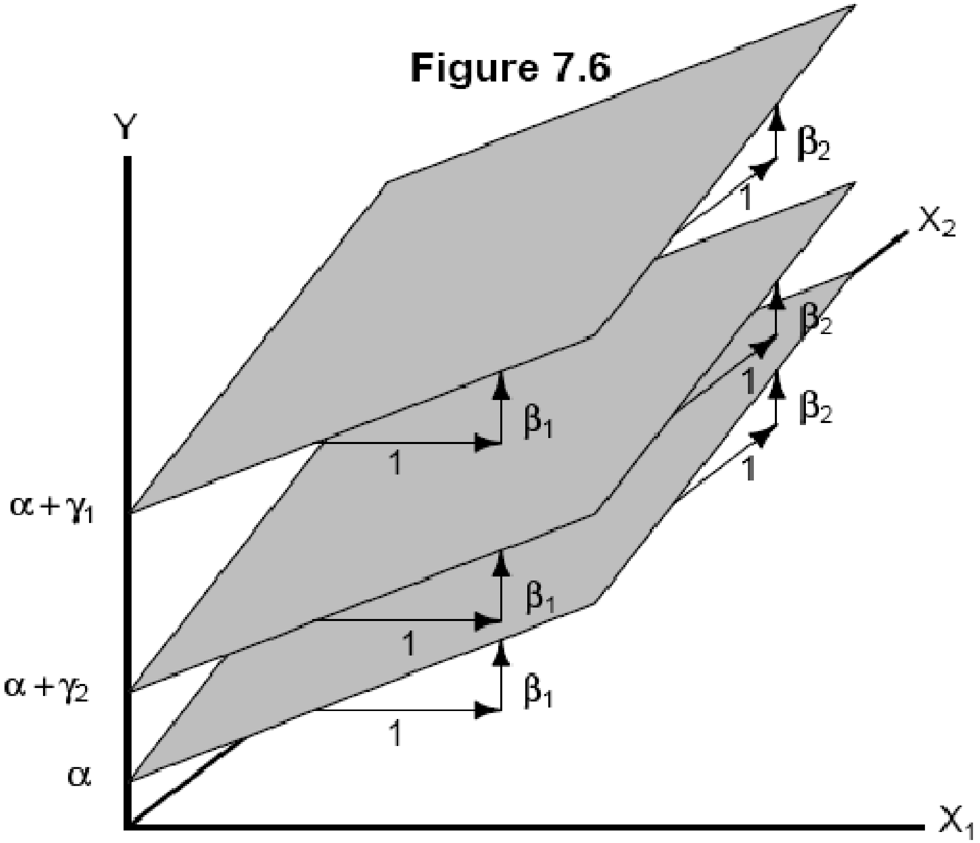
Category	D_1	D_2	D_3
Professional	1	0	0
White Collar	0	1	0
Blue Collar	0	0	1

Why won’t this work?

Conventional approach: pick one category as the “reference” category and use only (m-1) dummy variables for a factor with m levels.

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i$$

Figure 7.6



Dummy variable coding in R and

Testing contrasts (<u>not</u> using "quasi-variances" described in Fox's text)

Remember that this is just a multiple regression that can be written $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{X} = \begin{bmatrix} \underset{\sim}{1} & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{d}_1 & \mathbf{d}_2 \end{bmatrix}$ and $\boldsymbol{\beta}^T = \begin{pmatrix} \alpha & \beta_1 & \beta_2 & \gamma_1 & \gamma_2 \end{pmatrix}$.

To test $H_0 : \gamma_1 = \gamma_2$, i.e., the Professional and White-Collar jobs have the same prestige "adjusting for education level and income", we want to compute

$$t = \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{se(\hat{\gamma}_1 - \hat{\gamma}_2)} \text{ where } se(\hat{\gamma}_1 - \hat{\gamma}_2) = \sqrt{\hat{V}ar(\hat{\gamma}_1 - \hat{\gamma}_2)} \text{ and}$$

$$\hat{V}ar(\hat{\gamma}_1 - \hat{\gamma}_2) = \hat{V}ar(\hat{\gamma}_1) + \hat{V}ar(\hat{\gamma}_2) - 2 \cdot \hat{C}ov(\hat{\gamma}_1, \hat{\gamma}_2)$$

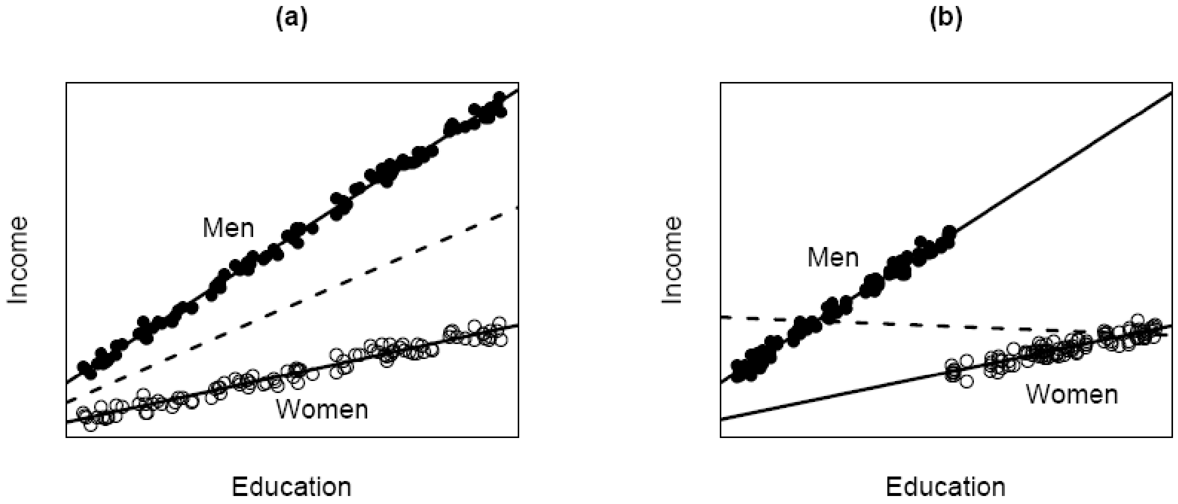
We extract these variances and diagonal elements from the off-diagonal elements of the variance-covariance matrix

$$\hat{V}ar(\hat{\boldsymbol{\beta}}) = s_E^2 \left(\mathbf{X}^T \mathbf{X} \right)^{-1}$$

In R . . .

7.3 Modeling Interactions

Figure 7.7



Interaction: The (partial) effect of one variable (or factor) depends on the value or level of another variable (factor). In this (contrived) example,

(a) the effect of education depends on whether we are considering men or women, and

(b) the difference between men and women ("effect of sex") depends on the level of education.

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

So for women, with $D_i = 0,$ $\quad Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i$

and for men, with $D_i = 1,$ $\quad Y_i = (\alpha + \gamma) + (\beta + \delta) X_i + \varepsilon_i$

I.e., we have fit two separate regression lines in one model.

The Principle of Marginality:

- In general, we do not test or interpret main effects of explanatory variables that interact, and
- We do not fit models with interaction terms without the main effects

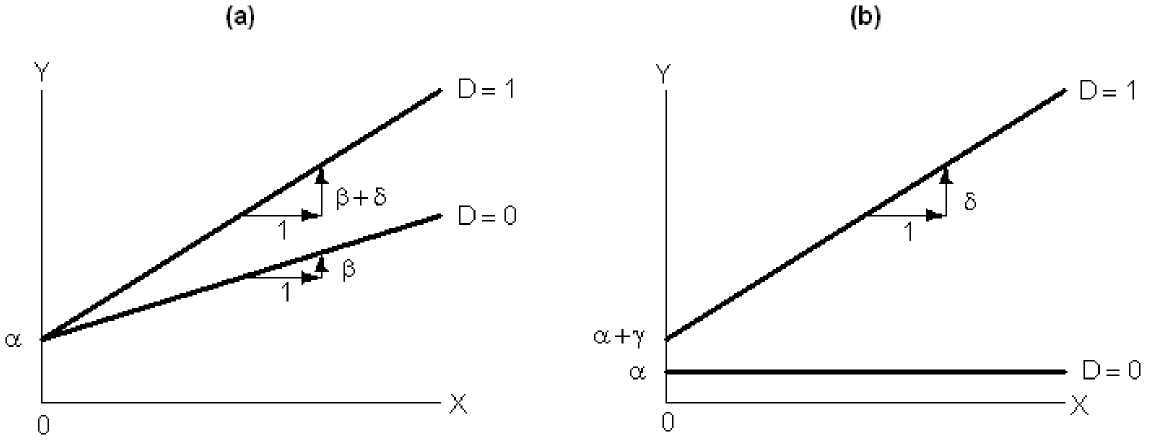
For example, we do not usually fit

$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$ -- two lines with common intercept

This can make sense in some applications, but less likely is

$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$ --- one line has zero slope

Figure 7.10



With multiple quantitative explanatory variables and polytomous factors, consider products of explanatory factors with dummy variables, with R and all other statistical analysis programs do automatically.

Back to R . . .

Tests of hypotheses

The default, "Type II" tests computed by the car "Anova" function honor the "Principle of Marginality":

- a <u>main effect</u> is tested by the incremental F-test procedure---comparing models with and without the effect of interest---only considering models <u>not</u> including interactions with the main effect

  --- *but you wouldn't usually test a <u>main effect</u> if you believed an interaction was appropriate as it would refer to an <u>average effect</u> over levels of the other factor.*
- an <u>interaction effect</u> is always tested by the incremental F-test procedure comparing models including the main effects
- <u>Note</u>: the denominator mean square (Table 7.2) is the “biggest” model with all main effects and interactions. This estimate is always unbiased for σ_{ε}^{2} and it is what you want to do unless the sample size is so small that there are very few degrees of freedom.

"Type III" tests, computed by lots of programs (and by an option in the “Anova” function) will test a given term in a model, main effect or interaction, against the model including all other terms. I.e., main effects are tested against models without the main effect but with interactions, which we don't usually want to do that.

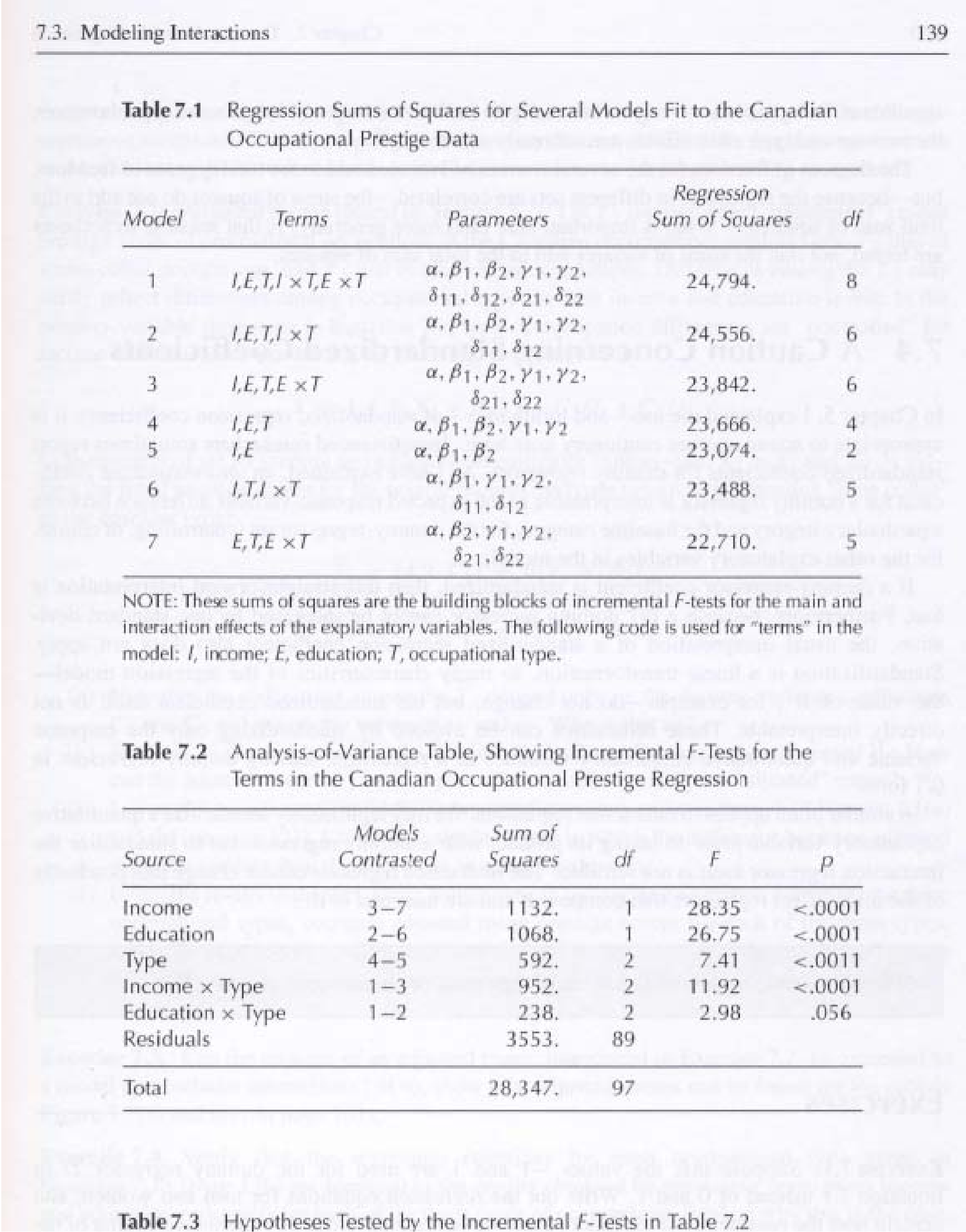


Table 7.1 Regression Sums of Squares for Several Models Fit to the Canadian Occupational Prestige Data

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

NOTE: These sums of squares are the building blocks of incremental *F*-tests for the main and interaction effects of the explanatory variables. The following code is used for "terms" in the model: *I*, income; *E*, education; *T*, occupational type.

Table 7.2 Analysis-of-Variance Table, Showing Incremental *F*-Tests for the Terms in the Canadian Occupational Prestige Regression

Source	*Models Contrasted*	*Sum of Squares*	*df*	*F*	*p*
Income	3−7	1132.	1	28.35	<.0001
Education	2−6	1068.	1	26.75	<.0001
Type	4−5	592.	2	7.41	<.0011
Income × Type	1−3	952.	2	11.92	<.0001
Education × Type	1−2	238.	2	2.98	.056
Residuals		3553.	89		
Total		28,347.	97		

Table 7.3 Hypotheses Tested by the Incremental *F*-Tests in Table 7.2