7 Dummy-Variable Regression

One of the serious limitations of multiple-regression analysis, as presented in Chapters 5 and 6, is that it accommodates only quantitative response and explanatory variables. In this chapter and the next, I will explain how qualitative explanatory variables, called *factors*, can be incorporated into a linear model.[1]

The current chapter begins with an explanation of how a *dummy-variable regressor* can be coded to represent a *dichotomous* (i.e., two-category) factor. I proceed to show how a set of dummy regressors can be employed to represent a *polytomous* (many-category) factor. I next describe how interactions between quantitative and qualitative explanatory variables can be represented in dummy-regression models and how to summarize models that incorporate interactions. Finally, I explain why it does not make sense to standardize dummy-variable and interaction regressors.

7.1 A Dichotomous Factor

Let us consider the simplest case: one dichotomous factor and one quantitative explanatory variable. As in the two previous chapters, assume that relationships are *additive*—that is, that the partial effect of each explanatory variable is the same regardless of the specific value at which the other explanatory variable is held constant. As well, suppose that the other assumptions of the regression model hold: The errors are independent and normally distributed, with zero means and constant variance.

The general motivation for including a factor in a regression is essentially the same as for including an additional quantitative explanatory variable: (1) to account more fully for the response variable, by making the errors smaller, and (2) even more important, to avoid a biased assessment of the impact of an explanatory variable, as a consequence of omitting another explanatory variable that is related to it.

For concreteness, suppose that we are interested in investigating the relationship between education and income among women and men. Figure 7.1(a) and (b) represents two small (idealized) populations. In both cases, the within-gender regressions of income on education are parallel. Parallel regressions imply additive effects of education and gender on income: Holding education constant, the "effect" of gender is the vertical distance between the two regression lines, which—for parallel lines—is everywhere the same. Likewise, holding gender constant, the "effect" of education is captured by the within-gender education slope, which—for parallel lines—is the same for men and women.[2]

In Figure 7.1(a), the explanatory variables gender and education are unrelated to each other: Women and men have identical distributions of education scores (as can been seen by projecting the points onto the horizontal axis). In this circumstance, if we ignore gender and regress income on education alone, we obtain the same slope as is produced by the separate within-gender

[1] Chapter 14 deals with qualitative *response* variables.

[2] I will consider nonparallel within-group regressions in Section 7.3.

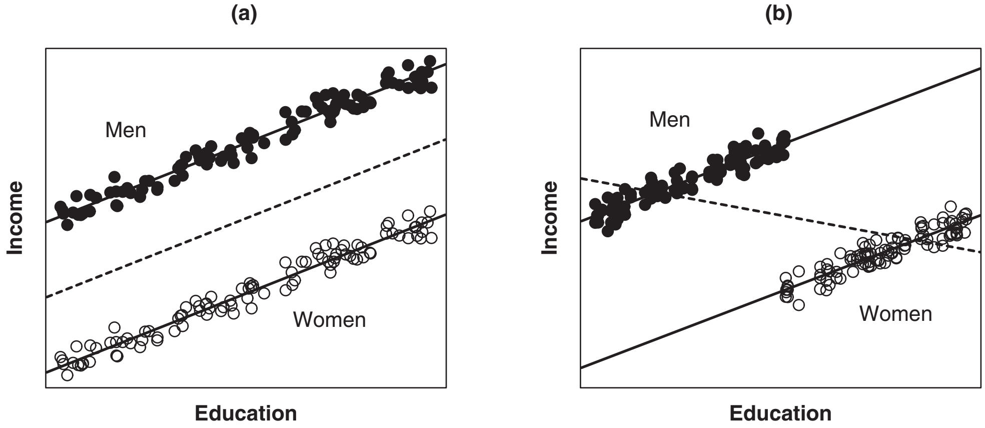

Figure 7.1 Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both (a) and (b), the within-gender (i.e., partial) regressions (solid lines) are parallel. In each graph, the overall (i.e., marginal) regression of income on education (ignoring gender) is given by the broken line.

regressions. Because women have lower incomes than men of equal education, however, by ignoring gender we inflate the size of the errors.

The situation depicted in Figure 7.1(b) is importantly different. Here, gender and education are related, and therefore if we regress income on education alone, we arrive at a biased assessment of the effect of education on income: Because women have a higher average level of education than men, and because—for a given level of education—women's incomes are lower, on average, than men's, the overall regression of income on education has a *negative* slope even though the within-gender regressions have a *positive* slope.[3]

In light of these considerations, we might proceed to partition our sample by gender and perform separate regressions for women and men. This approach is reasonable, but it has its limitations: Fitting separate regressions makes it difficult to estimate and test for gender differences in income. Furthermore, if we can reasonably assume parallel regressions for women and men, we can more efficiently estimate the common education slope by pooling sample data drawn from both groups. In particular, if the usual assumptions of the regression model hold, then it is desirable to fit the common-slope model by least squares.

One way of formulating the common-slope model is

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \tag{7.1}$$

where D, called a *dummy-variable regressor* or an *indicator variable*, is coded 1 for men and 0 for women:

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

[3] That marginal and partial relationships can differ in sign is called *Simpson's paradox* (Simpson, 1951). Here, the marginal relationship between income and education is negative, while the partial relationship, controlling for gender, is positive.

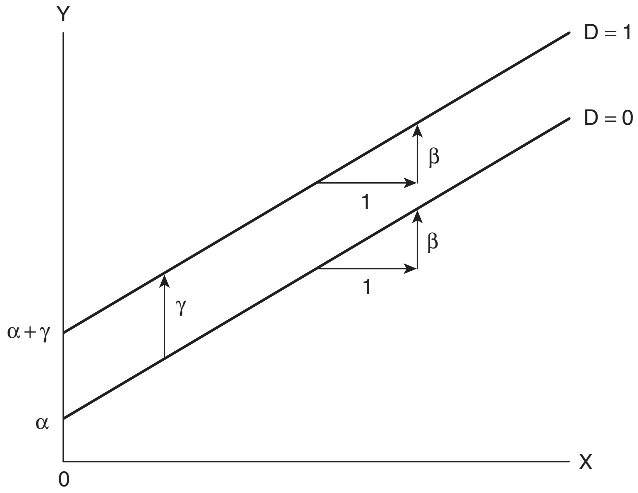

Figure 7.2 The additive dummy-variable regression model. The line labeled $D = 1$ is for men; the line labeled $D = 0$ is for women.

Thus, for women the model becomes

$$Y_i = \alpha + \beta X_i + \gamma(0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

and for men

$$Y_i = \alpha + \beta X_i + \gamma(1) + \varepsilon_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

These regression equations are graphed in Figure 7.2.

This is our initial encounter with an idea that is fundamental to many linear models: the distinction between *explanatory variables* and *regressors*. Here, *gender* is a qualitative explanatory variable (i.e., a factor), with categories *male* and *female*. The dummy variable D is a regressor, representing the factor gender. In contrast, the quantitative explanatory variable *education* and the regressor X are one and the same. Were we to transform education, however, prior to entering it into the regression equation—say, by taking logs—then there would be a distinction between the explanatory variable (education) and the regressor (log education). In subsequent sections of this chapter, it will transpire that an explanatory variable can give rise to several regressors and that some regressors are functions of more than one explanatory variable.

Returning to Equation 7.1 and Figure 7.2, the coefficient γ for the dummy regressor gives the difference in intercepts for the two regression lines. Moreover, because the within-gender regression lines are parallel, γ also represents the constant vertical separation between the lines, and it may, therefore, be interpreted as the expected income advantage accruing to men when education is held constant. If men were *dis*advantaged relative to women with the same level of education, then γ would be *negative*. The coefficient α gives the intercept for women, for whom $D = 0$; and β is the common within-gender education slope.

Figure 7.3 reveals the fundamental geometric "trick" underlying the coding of a dummy regressor: We are, in fact, fitting a regression plane to the data, but the dummy regressor D is defined only at the values 0 and 1. The regression plane intersects the planes $\{X, Y|D = 0\}$ and $\{X, Y|D = 1\}$ in two lines, each with slope β. Because the difference between $D = 0$ and $D = 1$ is one unit, the difference in the Y-intercepts of these two lines is the slope of the plane in the D direction,

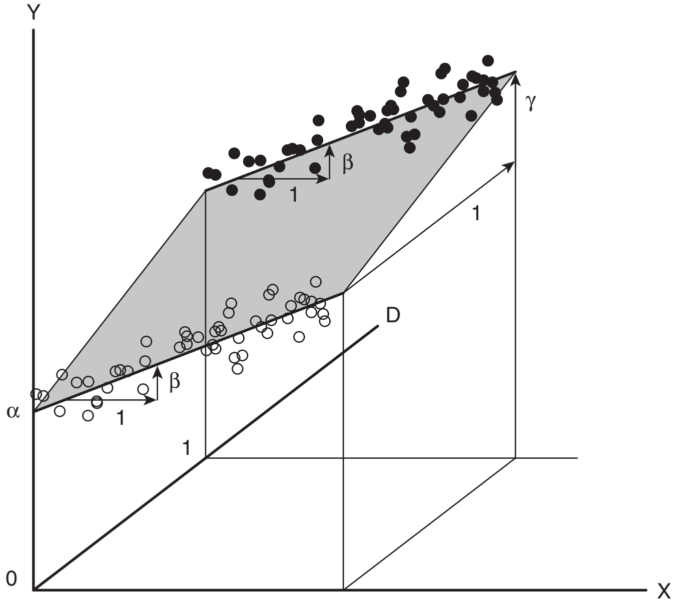

Figure 7.3 The geometric "trick" underlying dummy regression: The linear regression plane is defined only at $D = 0$ and $D = 1$, producing two regression lines with slope β and vertical separation γ. The hollow circles represent women, for whom $D = 0$, and the solid circles men, for whom $D = 1$.

that is γ. Indeed, Figure 7.2 is simply the projection of the two regression lines onto the $\{X, Y\}$ plane.

Essentially similar results are obtained if we instead code D equal to 0 for men and 1 for women, making men the *baseline* (or *reference*) category (see Figure 7.4): The *sign* of γ is reversed, because it now represents the difference in intercepts between women and men (rather than vice versa), but its *magnitude* remains the same. The coefficient α now gives the income intercept for men. It is therefore immaterial which group is coded 1 and which is coded 0, as long as we are careful to interpret the coefficients of the model—for example, the sign of γ—in a manner consistent with the coding scheme that is employed.

To determine whether gender affects income, controlling for education, we can test H_0: $\gamma = 0$, either by a t-test, dividing the estimate of γ by its standard error, or, equivalently, by dropping D from the regression model and formulating an incremental F-test. In either event, the statistical-inference procedures of the previous chapter apply.

Although I have developed dummy-variable regression for a single quantitative regressor, the method can be applied to any number of quantitative explanatory variables, as long as we are willing to assume that the slopes are the same in the two categories of the factor—that is, that the regression surfaces are parallel in the two groups. In general, if we fit the model

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

then, for $D = 0$, we have

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

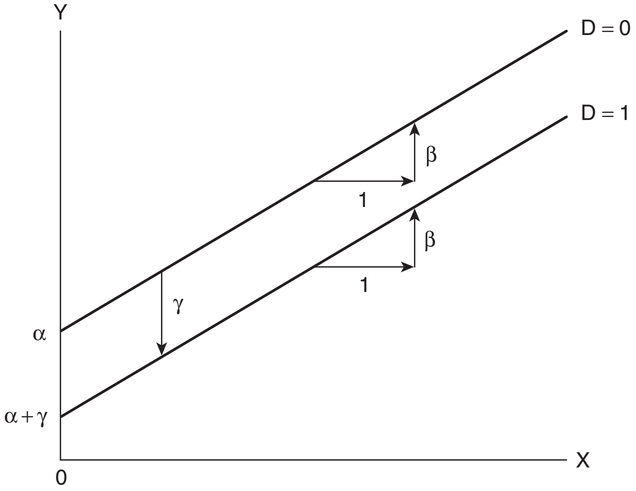

Figure 7.4 The additive dummy-regression model coding $D = 0$ for men and $D = 1$ for women (cf., Figure 7.2).

and, for $D = 1$,

$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

> A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the factor and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant vertical separation between the surfaces given by the coefficient of the dummy regressor.

7.2 Polytomous Factors

The coding method of the previous section generalizes straightforwardly to polytomous factors. By way of illustration, recall (from the previous chapter) the Canadian occupational prestige data. I have classified the occupations into three rough categories: (1) professional and managerial occupations, (2) "white-collar" occupations, and (3) "blue-collar" occupations.[4]

Figure 7.5 shows conditioning plots for the relationship between prestige and each of income and education within occupational types.[5] The partial relationships between prestige and the explanatory variables appear reasonably linear, although there seems to be evidence that the income slope varies across the categories of type of occupation (a possibility that I will pursue in the next section of the chapter). Indeed, this change in slope is an explanation of the nonlinearity in the relationship between prestige and income that we noticed in Chapter 4. These conditioning

[4] Although there are 102 occupations in the full data set, several are difficult to classify and consequently were dropped from the analysis. The omitted occupations are athletes, babysitters, farmers, and "newsboys," leaving us with 98 observations.

[5] In the preceding chapter, I also included the gender composition of the occupations as an explanatory variable, but I omit that variable here. Conditioning plots are described in Section 3.3.4.

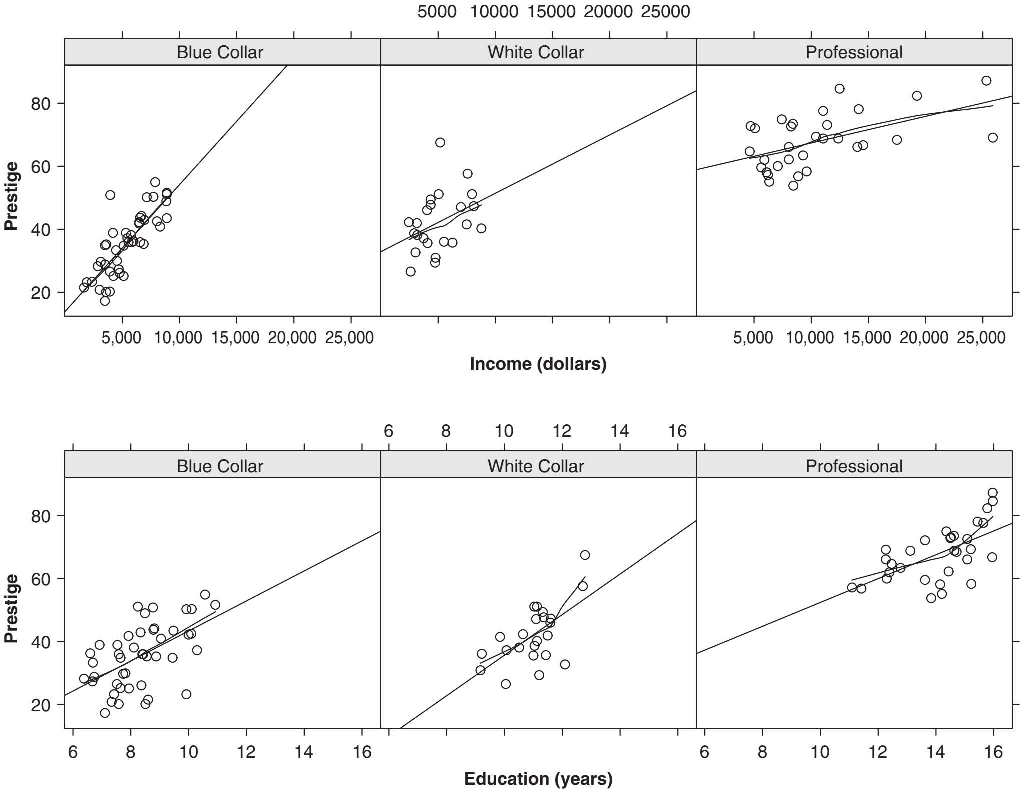

Figure 7.5 Condtioning plots for the relationships between prestige and each of income (top panel) and education (bottom panel) by type of occupation, for the Canadian occupational prestige data. Each panel shows the linear least-squares fit and a lowess smooth with a span of 0.9. The graphs labeled "Professional" are for professional and managerial occupations.

plots do not tell the whole story, however, because the income and education levels of the occupations are correlated, but they give us a reasonable initial look at the data. Conditioning the plot for income by level of education (and vice versa) is out of the question here because of the small size of the data set.

The *three*-category occupational-type factor can be represented in the regression equation by introducing *two* dummy regressors, employing the following coding scheme:

Category	D_1	D_2
Professional and managerial	1	0
White collar	0	1
Blue collar	0	0

(7.2)

A model for the regression of prestige on income, education, and type of occupation is then

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i \tag{7.3}$$

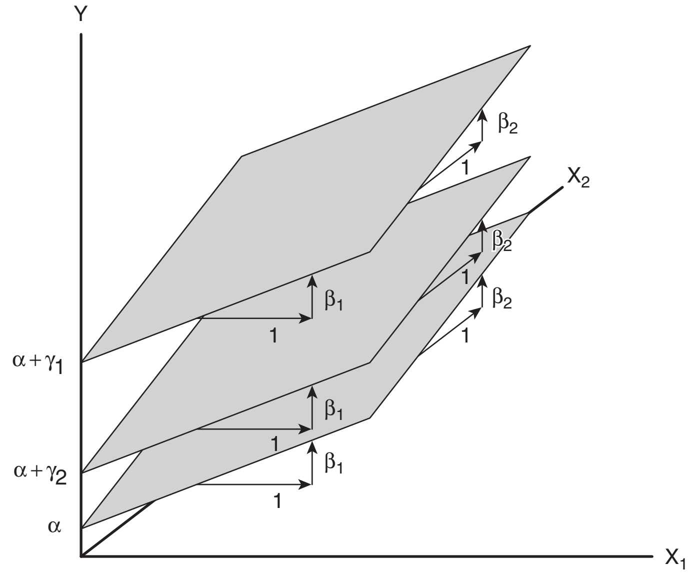

Figure 7.6 The additive dummy-regression model with two quantitative explanatory variables X_1 and X_2 represents parallel planes with potentially different intercepts in the $\{X_1,X_2,Y\}$ space.

where X_1 is income and X_2 is education. This model describes three parallel regression planes, which can differ in their intercepts:

$$\begin{array}{ll}\text{Professional:} & Y_i = (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{White collar:} & Y_i = (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ \text{Blue collar:} & Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i\end{array}$$

The coefficient α, therefore, gives the intercept for blue-collar occupations; γ_1 represents the constant vertical difference between the parallel regression planes for professional and blue-collar occupations (fixing the values of education and income); and γ_2 represents the constant vertical distance between the regression planes for white-collar and blue-collar occupations (again, fixing education and income). Assuming, for simplicity, that all coefficients are positive, and that $\gamma_1 > \gamma_2$, the geometry of the model in Equation 7.3 is illustrated in Figure 7.6.

Because blue-collar occupations are coded 0 for both dummy regressors, “blue collar” implicitly serves as the baseline category to which the other occupational-type categories are compared. The choice of a baseline category is essentially arbitrary, for we would fit precisely the same three regression planes regardless of which of the three occupational-type categories is selected for this role. The values (and meaning) of the individual dummy-variable coefficients γ_1 and γ_2 depend, however, on which category is chosen as the baseline.

It is sometimes natural to select a particular category as a basis for comparison—an experiment that includes a “control group” comes immediately to mind. In this instance, the individual dummy-variable coefficients are of interest, because they reflect differences between the “experimental” groups and the control group, holding other explanatory variables constant.

In most applications, however, the choice of a baseline category is entirely arbitrary, as it is for the occupational prestige regression. We are, therefore, most interested in testing the null hypothesis of no effect of occupational type, controlling for education and income,

$$H_0\colon \gamma_1 = \gamma_2 = 0 \tag{7.4}$$

but the individual hypotheses H_0: $\gamma_1 = 0$ and H_0: $\gamma_2 = 0$—which test, respectively, for differences between professional and blue-collar occupations and between white-collar and blue-collar occupations—are of less intrinsic interest.[6] The null hypothesis in Equation 7.4 can be tested by the incremental-sum-of-squares approach, dropping the two dummy variables for type of occupation from the model.

I have demonstrated how to model the effects of a three-category factor by coding two dummy regressors. It may seem more natural to treat the three occupational categories symmetrically, coding *three* dummy regressors, rather than arbitrarily selecting one category as the baseline:

Category	D_1	D_2	D_3
Professional and managerial	1	0	0
White collar	0	1	0
Blue collar	0	0	1

(7.5)

Then, for the jth occupational type, we would have

$$Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

The problem with this procedure is that there are too many parameters: We have used four parameters (α, γ_1, γ_2, γ_3) to represent only three group intercepts. As a consequence, we could not find unique values for these four parameters even if we knew the three population regression lines. Likewise, we cannot calculate unique least-squares estimates for the model because the set of three dummy variables is perfectly collinear; for example, as is apparent from the table in Equation 7.5, $D_3 = 1 - D_1 - D_2$.

In general, then, for a polytomous factor with m categories, we need to code $m - 1$ dummy regressors. One simple scheme is to select the last category as the baseline and to code $D_{ij} = 1$ when observation i falls in category j, and 0 otherwise:

Category	D_1	D_2	$\cdots$	D_{m-1}
1	1	0	$\cdots$	0
2	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$m-1$	0	0	$\cdots$	1
m	0	0	$\cdots$	0

(7.6)

A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The “omitted” category, coded 0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

[6] The essential point here is not that the separate hypotheses are of *no* interest but that they are an arbitrary subset of the pairwise differences among the categories. In the present case, where there are three categories, the individual hypotheses represent two of the three pairwise group comparisons. The third comparison, between professional and white-collar occupations, is not *directly* represented in the model, although it is given indirectly by the difference $\gamma_1 - \gamma_2$. See Section 7.2.1 for an elaboration of this point.

When there is more than one factor, and if we assume that the factors have additive effects, we can simply code a set of dummy regressors for each. To test the hypothesis that the effect of a factor is nil, we delete its dummy regressors from the model and compute an incremental F-test of the hypothesis that all the associated coefficients are 0.

Regressing occupational prestige (Y) on income (X_1) and education (X_2) produces the fitted regression equation

$$\begin{array}{lllll} \widehat{Y} = & -7.621 & +0.001241X_1 & +4.292X_2 & \quad R^2 = .81400 \\ & (3.116) & (0.000219) & (0.336) & \end{array}$$

As is common practice, I have shown the estimated standard error of each regression coefficient in parentheses beneath the coefficient. The three occupational categories differ considerably in their average levels of prestige:

Category	*Number of Cases*	*Mean Prestige*
Professional and managerial	31	67.85
White collar	23	42.24
Blue collar	44	35.53
All occupations	98	47.33

Inserting dummy variables for type of occupation into the regression equation, employing the coding scheme shown in Equation 7.2, produces the following results:

$$\begin{array}{llllll} \widehat{Y} = & -0.6229 & +0.001013X_1 & +3.673X_2 & +6.039D_1 & -2.737D_2 \\ & (5.2275) & (0.000221) & (0.641) & (3.867) & (2.514) \\ R^2 = .83486 & & & & & \end{array} \tag{7.7}$$

The three fitted regression equations are, therefore,

$$\begin{array}{ll} \text{Professional:} & \widehat{Y} = 5.416 + 0.001013X_1 + 3.673X_2 \\ \text{White collar:} & \widehat{Y} = -3.360 + 0.001013X_1 + 3.673X_2 \\ \text{Blue collar:} & \widehat{Y} = -0.623 + 0.001013X_1 + 3.673X_2 \end{array}$$

Note that the coefficients for both income and education become slightly smaller when type of occupation is controlled. As well, the dummy-variable coefficients (or, equivalently, the category intercepts) reveal that when education and income levels are held constant statistically, the difference in average prestige between professional and blue-collar occupations declines greatly, from $67.85 - 35.53 = 32.\,32$ points to 6.04 points. The difference between white-collar and blue-collar occupations is reversed when income and education are held constant, changing from $42.24 - 35.53 = +6.\,71$ points to -2.74 points. That is, the greater prestige of professional occupations compared with blue-collar occupations appears to be due mostly to differences in education and income between these two classes of occupations. While white-collar occupations have greater prestige, on average, than blue-collar occupations, they have lower prestige than blue-collar occupations of the same educational and income levels.[7]

To test the null hypothesis of no partial effect of type of occupation,

$$H_0\colon \gamma_1 = \gamma_2 = 0$$

[7] These conclusions presuppose that the additive model that we have fit to the data is adequate, which, as we will see in Section 7.3.5, is not the case.

we can calculate the incremental F-statistic

$$F_0 = \frac{n-k-1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \qquad (7.8)$$
$$= \frac{98-4-1}{2} \times \frac{.83486 - .81400}{1 - .83486} = 5.874$$

with 2 and 93 degrees of freedom, for which $p = .0040$. The occupational-type effect is therefore statistically significant but (examining the coefficient standard errors) not very precisely estimated. The education and income coefficients are several times their respective standard errors, and hence are highly statistically significant.

7.2.1 Coefficient Quasi-Variances*

Consider a dummy-regression model with p quantitative explanatory variables and an m-category factor:

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{m-1} D_{i,m-1} + \varepsilon_i$$

The dummy-variable coefficients $\gamma_1, \gamma_2, \ldots, \gamma_{m-1}$ represent differences (or *contrasts*) between each of the other categories of the factor and the reference category m, holding constant $X_1, \ldots, X_p$. If we are interested in a comparison between any other two categories, we can simply take the difference in their dummy-regressor coefficients. Thus, in the preceding example (letting $C_1 \equiv \widehat{\gamma}_1$ and $C_2 \equiv \widehat{\gamma}_2$),

$$C_1 - C_2 = 5.416 - (-3.360) = 8.776$$

is the estimated average difference in prestige between professional and white-collar occupations of equal income and education.

Suppose, however, that we want to know the standard error of $C_1 - C_2$. The standard errors of C_1 and C_2 are available directly in the regression "output" (Equation 7.7), but to compute the standard error of $C_1 - C_2$, we need in addition the estimated sampling covariance of these two coefficients. That is,[8]

$$\text{SE}(C_1 - C_2) = \sqrt{\widehat{V}(C_1) + \widehat{V}(C_2) - 2 \times \widehat{C}(C_1, C_2)}$$

where $\widehat{V}(C_j) = \left[\text{SE}(C_j)\right]^2$ is the estimated sampling variance of coefficient C_j, and $\widehat{C}(C_1, C_2)$ is the estimated sampling covariance of C_1 and C_2. For the occupational prestige regression, $\widehat{C}(C_1, C_2) = 6.797$, and so

$$\text{SE}(C_1 - C_2) = \sqrt{3.867^2 + 2.514^2 - 2 \times 6.797} = 2.771$$

We can use this standard error in the normal manner for a t-test of the difference between C_1 and C_2.[9] For example, noting that the difference exceeds twice its standard error suggests that it is statistically significant.

[8] See Appendix D on probability and estimation. The computation of regression-coefficient covariances is taken up in Chapter 9.

[9] Testing all differences between pairs of factor categories raises an issue of simultaneous inference, however. See the discussion of Scheffé confidence intervals in Section 9.4.4.

Although computer programs for regression analysis typically report the covariance matrix of the regression coefficients if asked to do so, it is not common to include coefficient covariances in published research along with estimated coefficients and standard errors, because with $k+1$ coefficients in the model, there are $k(k+1)/2$ variances and covariances among them—a potentially large number. Readers of a research report are therefore put at a disadvantage by the arbitrary choice of a reference category in dummy regression, because they are unable to calculate the standard errors of the differences between all pairs of categories of a factor.

Quasi-variances of dummy-regression coefficients (Firth, 2003; Firth & De Menezes, 2004) speak to this problem. Let $\widetilde{V}(C_j)$ denote the quasi-variance of dummy coefficient C_j. Then,

$$\text{SE}(C_j - C_{j'}) \approx \sqrt{\widetilde{V}(C_j) + \widetilde{V}(C_{j'})}$$

The squared relative error of this approximation for the contrast $C_j - C_{j'}$ is

$$\text{RE}_{jj'} \equiv \frac{\widetilde{V}(C_j - C_{j'})}{\widehat{V}(C_j - C_{j'})} = \frac{\widetilde{V}(C_j) + \widetilde{V}(C_{j'})}{\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})}$$

The approximation is accurate for this contrast when $\text{RE}_{jj'}$ is close to 1, or, equivalently, when

$$\log(\text{RE}_{jj'}) = \log\left[\widetilde{V}(C_j) + \widetilde{V}(C_{j'})\right] - \log\left[\widehat{V}(C_j) + \widehat{V}(C_{j'}) - 2 \times \widehat{C}(C_j, C_{j'})\right]$$

is close to 0. The quasi-variances $\widetilde{V}(C_j)$ are therefore selected to minimize the sum of squared log relative errors of approximation over all pairwise contrasts, $\sum_{j<j'}\left[\log(\text{RE}_{jj'})\right]^2$. The resulting errors of approximation are typically very small (Firth, 2003; Firth & De Menezes, 2004).

The following table gives dummy-variable coefficients, standard errors, and quasi-variances for type of occupation in the Canadian occupational prestige regression:

Category	C_j	SE(C_j)	$\widetilde{V}(C_j)$
Professional	6.039	3.867	8.155
White collar	−2.737	2.514	−0.4772
Blue collar	0	0	6.797

I have set to 0 the coefficient (and its standard error) for the baseline category, blue collar. The negative quasi-variance for the white-collar coefficient is at first blush disconcerting (after all, ordinary variances cannot be negative), but it is not wrong: The quasi-variances are computed to provide accurate variance approximations for coefficient *differences*; they do not apply directly to the coefficients themselves. For the contrast between professional and white-collar occupations, we have

$$\text{SE}(C_1 - C_2) \approx \sqrt{8.155 - 0.4772} = 2.771$$

Likewise, for the contrast between professional and blue-collar occupations,

$$C_1 - C_3 = 6.039 - 0 = 6.039$$
$$\text{SE}(C_1 - C_3) \approx \sqrt{8.155 + 6.797} = 3.867$$

Note that in this application, the quasi-variance "approximation" to the standard error proves to be exact, and indeed this is necessarily the case when there are just three factor categories, because there are then just three pairwise differences among the categories to capture.[10]

[10] For the details of the computation of quasi-variances, see Chapter 15, Exercise 15.11.

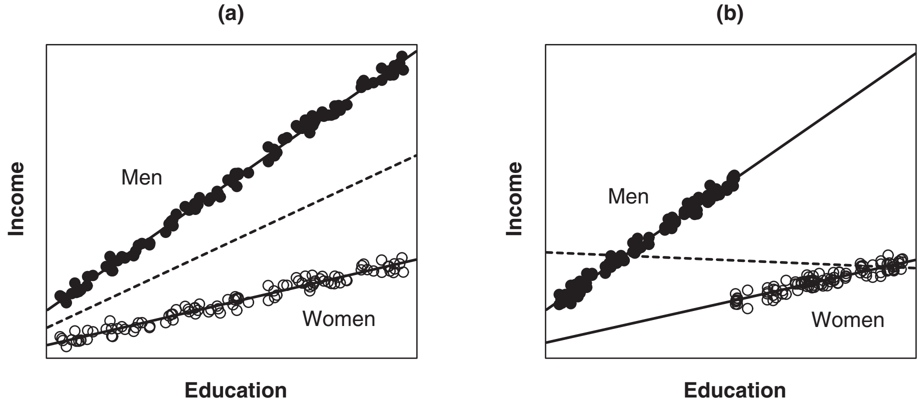

Figure 7.7 Idealized data representing the relationship between income and education for populations of men (filled circles) and women (open circles). In (a), there is no relationship between education and gender; in (b), women have a higher average level of education than men. In both cases, the within-gender regressions (solid lines) are not parallel—the slope for men is greater than the slope for women—and, consequently, education and gender interact in affecting income. In each graph, the overall regression of income on education (ignoring gender) is given by the broken line.

7.3 Modeling Interactions

Two explanatory variables are said to *interact* in determining a response variable when the partial effect of one depends on the value of the other. The additive models that we have considered thus far therefore specify the *absence* of interactions. In this section, I will explain how the dummy-variable regression model can be modified to accommodate interactions between factors and quantitative explanatory variables.[11]

The treatment of dummy-variable regression in the preceding two sections has assumed parallel regressions across the several categories of a factor. If these regressions are *not* parallel, then the factor interacts with one or more of the quantitative explanatory variables. The dummy-regression model can easily be modified to reflect these interactions.

For simplicity, I return to the contrived example of Section 7.1, examining the regression of income on gender and education. Consider the hypothetical data shown in Figure 7.7 (and contrast these examples with those shown in Figure 7.1 on page 121, where the effects of gender and education are additive). In Figure 7.7(a) [as in Figure 7.1(a)], gender and education are independent, because women and men have identical education distributions; in Figure 7.7(b) [as in Figure 7.1(b)], gender and education are related, because women, on average, have higher levels of education than men.

It is apparent in both Figure 7.7(a) and Figure 7.7(b), however, that the within-gender regressions of income on education are not parallel: In both cases, the slope for men is larger than the slope for women. Because the effect of education varies by gender, education and gender interact in affecting income.

It is also the case, incidentally, that the effect of gender varies by education. Because the regressions are not parallel, the relative income advantage of men changes (indeed, grows) with

[11] Interactions between factors are taken up in the next chapter on analysis of variance; interactions between quantitative explanatory variables are discussed in Section 17.1 on polynomial regression.

education. Interaction, then, is a symmetric concept—that the effect of education varies by gender implies that the effect of gender varies by education (and, of course, vice versa).

The simple examples in Figures 7.1 and 7.7 illustrate an important and frequently misunderstood point: *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves.

> Interaction and correlation of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact whether or not they are related to one another statistically. Interaction refers to the manner in which explanatory variables combine to affect a response variable, not to the relationship between the explanatory variables themselves.

7.3.1 Constructing Interaction Regressors

We could model the data in Figure 7.7 by fitting separate regressions of income on education for women and men. As before, however, it is more convenient to fit a combined model, primarily because a combined model facilitates a test of the gender-by-education interaction. Moreover, a properly formulated unified model that permits different intercepts and slopes in the two groups produces the same fit to the data as separate regressions: The full sample is composed of the two groups, and, consequently, the residual sum of squares for the full sample is minimized when the residual sum of squares is minimized in each group.[12]

The following model accommodates different intercepts and slopes for women and men:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i \tag{7.9}$$

Along with the quantitative regressor X for education and the dummy regressor D for gender, I have introduced the *interaction regressor* XD into the regression equation. The interaction regressor is the *product* of the other two regressors; although XD is therefore a function of X and D, it is not a *linear* function, and perfect collinearity is avoided.[13]

For women, model (7.9) becomes

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \end{aligned}$$

and for men

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta)X_i + \varepsilon_i \end{aligned}$$

[12] See Exercise 7.4.

[13] If this procedure seems illegitimate, then think of the interaction regressor as a new variable, say $Z \equiv XD$. The model is linear in X, D, and Z. The "trick" of introducing an interaction regressor is similar to the trick of formulating dummy regressors to capture the effect of a factor: In both cases, there is a distinction between explanatory variables and regressors. Unlike a dummy regressor, however, the interaction regressor is a function of *both* explanatory variables.

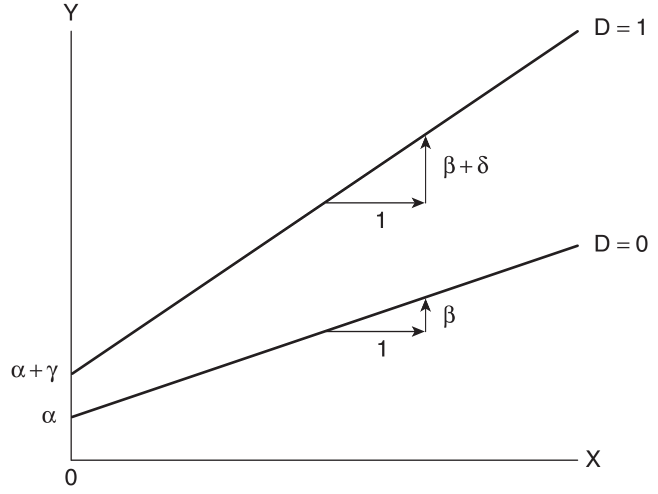

Figure 7.8 The dummy-variable regression model with an interaction regressor. The line labeled $D = 1$ is for men; the line labeled $D = 0$ is for women.

These regression equations are graphed in Figure 7.8: The parameters α and β are, respectively, the intercept and slope for the regression of income on education among women (the baseline category for gender); γ gives the difference in intercepts between the male and female groups; and δ gives the difference in slopes between the two groups. To test for interaction, therefore, we may simply test the hypothesis H_0: $\delta = 0$.

> Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The resulting model permits different slopes in different groups—that is, regression surfaces that are not parallel.

In the additive, no-interaction model of Equation 7.1 and Figure 7.2, the dummy-regressor coefficient γ represents the *unique* partial effect of gender (i.e., the expected income difference between men and women of equal education, regardless of the value at which education is fixed), while the slope β represents the *unique* partial effect of education (i.e., the within-gender expected increment in income for a one-unit increase in education, for both women and men). In the interaction model of Equation 7.9 and Figure 7.8, in contrast, γ is no longer interpretable as the unqualified income difference between men and women of equal education.

Because the within-gender regressions are not parallel, the separation between the regression lines changes; here, γ is simply the separation at $X = 0$—that is, above the origin. It is generally no more important to assess the expected income difference between men and women of 0 education than at other educational levels, and therefore the difference-in-intercepts parameter γ is not of special interest in the interaction model. Indeed, in many instances (although not here), the value $X = 0$ may not occur in the data or may be impossible (as, for example, if X is weight). In such cases, γ has no literal interpretation in the interaction model (see Figure 7.9).

Likewise, in the interaction model, β is not the unqualified partial effect of education, but rather the effect of education among women. Although this coefficient *is* of interest, it is not necessarily

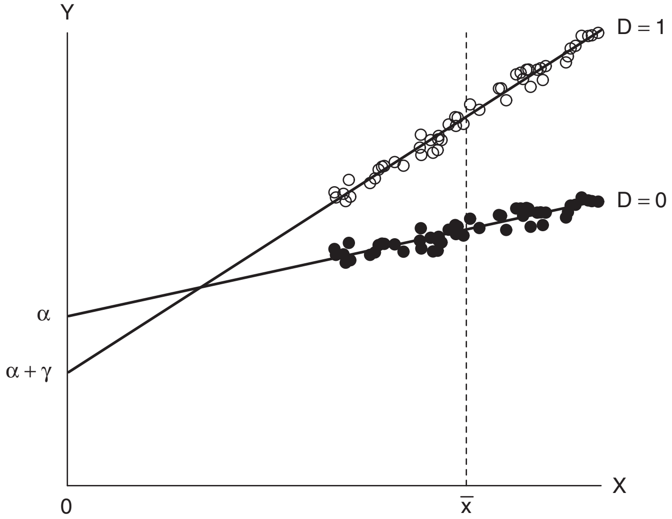

Figure 7.9 Why the difference in intercepts does not represent a meaningful partial effect for a factor when there is interaction: The difference-in-intercepts parameter γ is *negative* even though, within the range of the data, the regression line for the group coded $D = 1$ is *above* the line for the group coded $D = 0$.

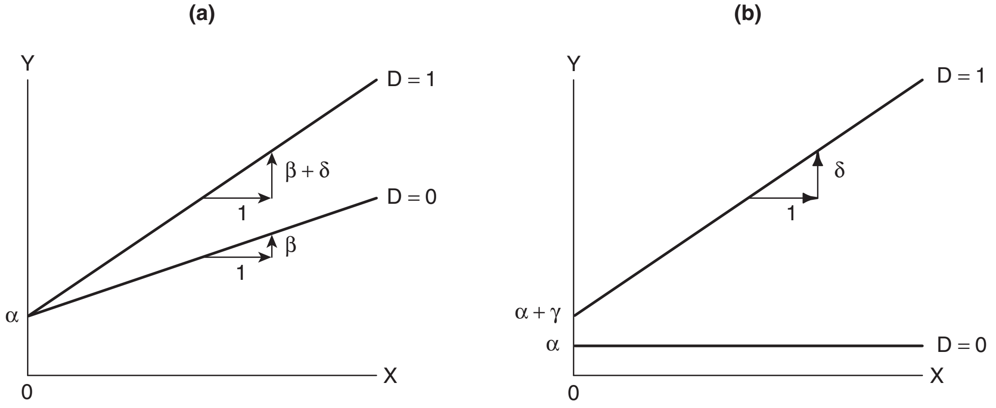

Figure 7.10 Two models that violate the principle of marginality: In (a), the dummy regressor D is omitted from the model $E(Y) = \alpha + \beta X + \delta\ (XD)$; in (b), the quantitative explanatory variable X is omitted from the model $E(Y) = \alpha + \gamma D + \delta(XD)$. These models violate the principle of marginality because they include the term XD, which is a higher-order relative of both X and D (one of which is omitted from each model).

more important than the effect of education among men ($\beta + \delta$), which does not appear *directly* in the model.

7.3.2 The Principle of Marginality

Following Nelder (1977), we say that the separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction. In general, we neither test nor interpret the main effects of explanatory variables that interact. If, however, we can rule out interaction either on theoretical or on empirical grounds, then we can proceed to test, estimate, and interpret the main effects.

As a corollary to this principle, it does not generally make sense to specify and fit models that include interaction regressors but that omit main effects that are marginal to them. This is not to say that such models—which violate the *principle of marginality*—are uninterpretable: They are, rather, not broadly applicable.

> The principle of marginality specifies that a model including a *high-order term* (such as an interaction) should normally also include the "lower-order relatives" of that term (the main effects that "compose" the interaction).

Suppose, for example, that we fit the model

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

which omits the dummy regressor D, but includes its "higher-order relative" XD. As shown in Figure 7.10(a), this model describes regression lines for women and men that have the same intercept but (potentially) different slopes, a specification that is peculiar and of no substantive interest. Similarly, the model

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

graphed in Figure 7.10(b), constrains the slope for women to 0, which is needlessly restrictive.

7.3.3 Interactions With Polytomous Factors

The method for modeling interactions by forming product regressors is easily extended to polytomous factors, to several factors, and to several quantitative explanatory variables. I will use the Canadian occupational prestige regression to illustrate the application of the method, entertaining the possibility that occupational type interacts both with income (X_1) and with education (X_2):

$$\begin{aligned} Y_i = \alpha &+ \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ &+ \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i \end{aligned} \tag{7.10}$$

Note that we require one interaction regressor for each product of a dummy regressor with a quantitative explanatory variable. The regressors $X_1 D_1$ and $X_1 D_2$ capture the interaction between income and occupational type; $X_2 D_1$ and $X_2 D_2$ capture the interaction between education and occupational type. The model therefore permits different intercepts and slopes for the three types of occupations:

$$\begin{array}{llcl} \text{Professional:} & Y_i = (\alpha+\gamma_1) + (\beta_1+\delta_{11})X_{i1} + (\beta_2+\delta_{21})X_{i2} + \varepsilon_i \\ \text{White collar:} & Y_i = (\alpha+\gamma_2) + (\beta_1+\delta_{12})X_{i1} + (\beta_2+\delta_{22})X_{i2} + \varepsilon_i \\ \text{Blue collar:} & Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{array} \tag{7.11}$$

Blue-collar occupations, which are coded 0 for both dummy regressors, serve as the baseline for the intercepts and slopes of the other occupational types. As in the no-interaction model, the choice of baseline category is generally arbitrary, as it is here, and is inconsequential. Fitting the model in Equation 7.10 to the prestige data produces the following results:

$$\begin{aligned} \widehat{Y}_i = \underset{(7.057)}{2.276} + \underset{(0.000556)}{0.003522X_1} + \underset{(0.927)}{1.713X_2} + \underset{(13.72)}{15.35D_1} - \underset{(17.54)}{33.54D_2} \\ - \underset{(0.000599)}{0.002903X_1D_1} - \underset{(0.000894)}{0.002072X_1D_2} \\ + \underset{(1.289)}{1.388X_2D_1} + \underset{(1.757)}{4.291X_2D_2} \\ R^2 = .8747 \end{aligned} \tag{7.12}$$

This example is discussed further in the following section.

7.3.4 Interpreting Dummy-Regression Models With Interactions

It is difficult in dummy-regression models with interactions (and in other complex statistical models) to understand what the model is saying about the data simply by examining the regression coefficients. One approach to interpretation, which works reasonably well in a relatively straightforward model such as Equation 7.12, is to write out the implied regression equation for each group (using Equation 7.11):

$$\begin{array}{lll} \text{Professional:} & \widehat{\text{Prestige}} = & 17.63 + 0.000619 \times \text{Income} + 3.101 \times \text{Education} \\ \text{White collar:} & \widehat{\text{Prestige}} = & -31.26 + 0.001450 \times \text{Income} + 6.004 \times \text{Education} \\ \text{Blue collar:} & \widehat{\text{Prestige}} = & 2.276 + 0.003522 \times \text{Income} + 1.713 \times \text{Education} \end{array} \tag{7.13}$$

From these equations, we can see, for example, that income appears to make much more difference to prestige in blue-collar occupations than in white-collar occupations, and has even less impact on prestige in professional and managerial occupations. Education, in contrast, has the largest impact on prestige among white-collar occupations, and has the smallest effect in blue-collar occupations.

An alternative approach (from Fox, 1987, 2003; Fox & Andersen, 2006) that generalizes readily to more complex models is to examine the high-order terms of the model. In the illustration, the high-order terms are the interactions between income and type and between education and type.

- Focusing in turn on each high-order term, we allow the variables in the term to range over their combinations of values in the data, fixing other variables to typical values. For example, for the interaction between type and income, we let type of occupation take on successively the categories blue collar, white collar, and professional (for which the dummy regressors

D_1 and D_2 are set to the corresponding values given in Equation 7.6), in combination with income values between \$1500 and \$26,000 (the approximate range of income in the Canadian occupational prestige data set); education is fixed to its average value in the data, $\overline{X}_2 = 10.79$.

- We next compute the fitted value of prestige at each combination of values of income and type of occupation. These fitted values are graphed in the "effect display" shown in the upper panel of Figure 7.11; the lower panel of this figure shows a similar effect display for the interaction between education and type of occupation, holding income at its average value. The broken lines in Figure 7.11 give ±2 standard errors around the fitted values—that is, approximate 95% pointwise confidence intervals for the effects.[14] The nature of the interactions between income and type and between education and type is readily discerned from these graphs.

7.3.5 Hypothesis Tests for Main Effects and Interactions

To test the null hypothesis of no interaction between income and type, H_0: $\delta_{11} = \delta_{12} = 0$, we need to delete the interaction regressors X_1D_1 and X_1D_2 from the full model (Equation 7.10) and calculate an incremental F-test; likewise, to test the null hypothesis of no interaction between education and type, H_0: $\delta_{21} = \delta_{22} = 0$, we delete the interaction regressors X_2D_1 and X_2D_2 from the full model. These tests, and tests for the main effects of income, education, and occupational type, are detailed in Tables 7.1 and 7.2: Table 7.1 gives the regression sums of squares for several models, which, along with the residual sum of squares for the full model, $\text{RSS}_1 = 3553$, are the building blocks of the incremental F-tests shown in Table 7.2. Table 7.3 shows the hypothesis tested by each of the incremental F-statistics in Table 7.2.

Although the analysis-of-variance table (Table 7.2) conventionally shows the tests for the main effects of education, income, and type before the education-by-type and income-by-type interactions, the structure of the model makes it sensible to examine the interactions first: Conforming to the principle of marginality, the test for each main effect is computed assuming that the interactions that are higher-order relatives of the main effect are 0 (as shown in Table 7.3). Thus, for example, the test for the income main effect assumes that the income-by-type interaction is absent (i.e., that $\delta_{11} = \delta_{12} = 0$), but not that the education-by-type interaction is absent ($\delta_{21} = \delta_{22} = 0$).[15]

> The principle of marginality serves as a guide to constructing incremental F-tests for the terms in a model that includes interactions.

In this case, then, there is weak evidence of an interaction between education and type of occupation, and much stronger evidence of an income-by-type interaction. Considering the small number of cases, we are squeezing the data quite hard, and it is apparent from the coefficient standard errors (in Equation 7.12) and from the effect displays in Figure 7.11 that the interactions are not precisely estimated. The tests for the main effects of income, education, and type, computed assuming that the higher-order relatives of each such term are absent, are all highly statistically

[14] For standard errors of fitted values, see Exercise 9.14.

[15] Tests constructed to conform to the principle of marginality are sometimes called "type-II" tests, terminology introduced by the SAS statistical software package. This terminology, and alternative tests, are described in the next chapter.

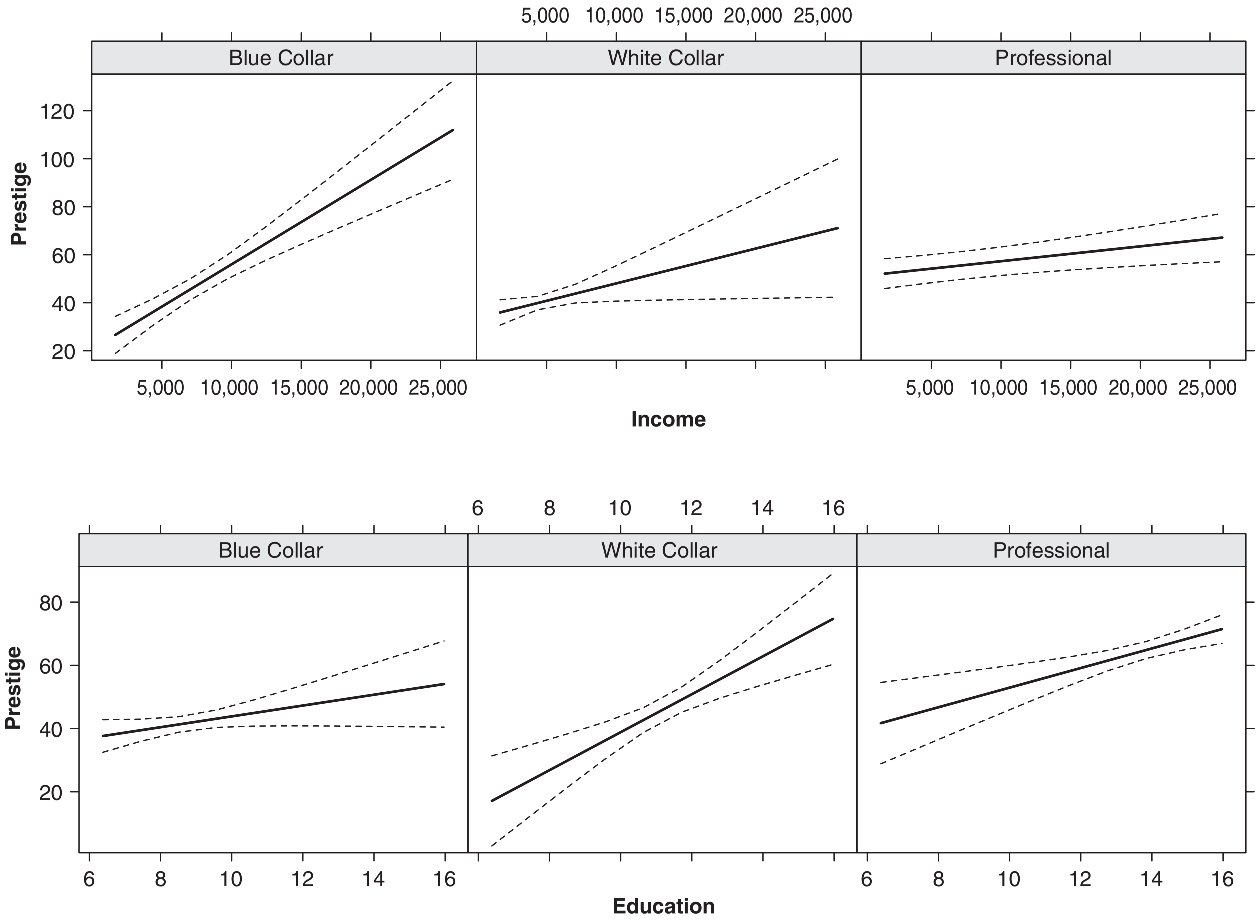

Figure 7.11 Income-by-type (upper panel) and education-by-type (lower panel) “effect displays” for the regression of prestige on income, education, and type of occupation. The solid lines give fitted values under the model, while the broken lines give 95% pointwise confidence intervals around the fit. To compute fitted values in the upper panel, education is set to its average value in the data; in the lower panel, income is set to its average value.

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* ×*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*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ δ_{11}, δ_{12}	24,556.	6
3	*I,E,T,E* ×*T*	$\alpha, \beta_1, \beta_2, \gamma_1, \gamma_2,$ δ_{21}, δ_{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*	$\alpha, \beta_1, \gamma_1, \gamma_2,$ δ_{11}, δ_{12}	23,488.	5
7	*E,T,E* ×*T*	$\alpha, \beta_2, \gamma_1, \gamma_2,$ δ_{21}, δ_{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

Source	*Models* *Contrasted*	*Null Hypothesis*
Income	3–7	$\beta_1 = 0 \mid \delta_{11} = \delta_{12} = 0$
Education	2–6	$\beta_2 = 0 \mid \delta_{21} = \delta_{22} = 0$
Type	4–5	$\gamma_1 = \gamma_2 = 0 \mid \delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$
Income ×Type	1–3	$\delta_{11} = \delta_{12} = 0$
Education × Type	1–2	$\delta_{21} = \delta_{22} = 0$

significant. In light of the strong evidence for an interaction between income and type, however, the income and type main effects are not really of interest.[16]

The degrees of freedom for the several sources of variation add to the total degrees of freedom, but—because the regressors in different sets are correlated—the sums of squares do not add to the total sum of squares.[17] What is important here (and more generally) is that sensible hypotheses are tested, not that the sums of squares add to the total sum of squares.

7.4 A Caution Concerning Standardized Coefficients

In Chapter 5, I explained the use—and limitations—of standardized regression coefficients. It is appropriate to sound another cautionary note here: Inexperienced researchers sometimes report standardized coefficients for dummy regressors. As I have explained, an *unstandardized* coefficient for a dummy regressor is interpretable as the expected response-variable difference between a particular category and the baseline category for the dummy-regressor set (controlling, of course, for the other explanatory variables in the model).

If a dummy-regressor coefficient is standardized, then this straightforward interpretation is lost. Furthermore, because a 0/1 dummy regressor cannot be increased by one standard deviation, the usual interpretation of a standardized regression coefficient also does not apply. Standardization is a linear transformation, so many characteristics of the regression model—the value of R^2, for example—do not change, but the standardized coefficient itself is not directly interpretable. These difficulties can be avoided by standardizing only the response variable and *quantitative* explanatory variables in a regression, leaving dummy regressors in 0/1 form.

A similar point applies to interaction regressors. We may legitimately standardize a quantitative explanatory variable *prior* to taking its product with a dummy regressor, but to standardize the interaction regressor itself is not sensible: The interaction regressor cannot change independently of the main-effect regressors that compose it and are marginal to it.

> It is not sensible to standardize dummy regressors or interaction regressors.

Exercises

Exercise 7.1. Suppose that the values -1 and 1 are used for the dummy regressor D in Equation 7.1 instead of 0 and 1. Write out the regression equations for men and women, and explain how the parameters of the model are to be interpreted. Does this alternative coding of the

[16] We tested the occupational type main effect in Section 7.2 (Equation 7.8 on page 129), but using an estimate of error variance based on Model 4, which does not contain the interactions. In Table 7.2, the estimated error variance is based on the full model, Model 1. Sound general practice is to use the largest model fit to the data to estimate the error variance even when, as is frequently the case, this model includes effects that are not statistically significant. The largest model necessarily has the smallest residual sum of squares, but it also has the fewest residual degrees of freedom. These two factors tend to offset one another, and it usually makes little difference whether the estimated error variance is based on the full model or on a model that deletes nonsignificant terms. Nevertheless, using the full model ensures an unbiased estimate of the error variance.

[17] See Section 10.2 for a detailed explanation of this phenomenon.

dummy regressor adequately capture the effect of gender? Is it fair to conclude that the dummy-regression model will "work" properly as long as two distinct values of the dummy regressor are employed, one each for women and men? Is there a reason to prefer one coding to another?

Exercise 7.2. Adjusted means (based on Section 7.2): Let $\overline{Y}_1$ represent the ("unadjusted") mean prestige score of professional occupations in the Canadian occupational prestige data, $\overline{Y}_2$ that of white-collar occupations, and $\overline{Y}_3$ that of blue-collar occupations. Differences among the $\overline{Y}_j$ may partly reflect differences among occupational types in their income and education levels. In the dummy-variable regression in Equation 7.7, type-of-occupation differences are "controlled" for income and education, producing the fitted regression equation

$$\widehat{Y} = A + B_1X_1 + B_2X_2 + C_1D_1 + C_2D_2$$

Consequently, if we fix income and education at particular values—say, $X_1 = x_1$ and $X_2 = x_2$—then the fitted prestige scores for the several occupation types are given by (treating "blue collar" as the baseline type):

$$\begin{aligned} \widehat{Y}_1 &= (A + C_1) + B_1x_1 + B_2x_2 \\ \widehat{Y}_2 &= (A + C_2) + B_1x_1 + B_2x_2 \\ \widehat{Y}_3 &= \quad A \quad + B_1x_1 + B_2x_2 \end{aligned}$$

(a) Note that the *differences* among the $\widehat{Y}_j$ depend only on the dummy-variable coefficients C_1 and C_2 and not on the values of x_1 and x_2. Why is this so?
(b) When $x_1 = \overline{X}_1$ and $x_2 = \overline{X}_2$, the $\widehat{Y}_j$ are called *adjusted means* and are denoted $\widetilde{Y}_j$. How can the adjusted means $\widetilde{Y}_j$ be interpreted? In what sense is $\widetilde{Y}_j$ an "adjusted" mean?
(c) Locate the "unadjusted" and adjusted means for women and men in each of Figures 7.1(a) and (b) (on page 121). Construct a similar figure in which the difference between adjusted means is *smaller* than the difference in unadjusted means.
(d) Using the results in the text, along with the mean income and education values for the three occupational types, compute adjusted mean prestige scores for each of the three types, controlling for income and education. Compare the adjusted with the unadjusted means for the three types of occupations and comment on the differences, if any, between them.

Exercise 7.3. Can the concept of an adjusted mean, introduced in Exercise 7.2, be extended to a model that includes interactions? If so, show how adjusted means can be found for the data in Figure 7.7(a) and (b) (on page 131).

Exercise 7.4. Verify that the regression equations for each occupational type given in Equation 7.13 (page 136) are identical to the results obtained by regressing prestige on income and education *separately* for each of the three types of occupations. Explain why this is the case.

Summary

- A dichotomous factor can be entered into a regression equation by formulating a dummy regressor, coded 1 for one category of the variable and 0 for the other category. A model incorporating a dummy regressor represents parallel regression surfaces, with the constant separation between the surfaces given by the coefficient of the dummy regressor.
- A polytomous factor can be entered into a regression by coding a set of 0/1 dummy regressors, one fewer than the number of categories of the factor. The "omitted" category, coded

0 for all dummy regressors in the set, serves as a baseline to which the other categories are compared. The model represents parallel regression surfaces, one for each category of the factor.

- Interactions can be incorporated by coding interaction regressors, taking products of dummy regressors with quantitative explanatory variables. The model permits different slopes in different groups—that is, regression surfaces that are not parallel.
- *Interaction* and *correlation* of explanatory variables are empirically and logically distinct phenomena. Two explanatory variables can interact *whether or not* they are related to one another statistically. Interaction refers to the manner in which explanatory variables *combine* to affect a response variable, not to the relationship *between* the explanatory variables themselves
- The principle of marginality specifies that a model including a high-order term (such as an interaction) should normally also include the lower-order relatives of that term (the main effects that "compose" the interaction). The principle of marginality also serves as a guide to constructing incremental F-tests for the terms in a model that includes interactions, and for examining the effects of explanatory variables.
- It is not sensible to standardize dummy regressors or interaction regressors.