5 Nominal Independent Variables

Excerpts from "Multiple Regression and Causal Analysis" by McKee J. McClendon. 1994. F.E. Peacock

In the preceding chapters, we have assumed that the variables used in a regression equation are measured at least at the *interval* level or, if *ordinal* variables are used, that scores can be assigned to the ordered categories that approximate the intervals between the categories (see Chapter 1 for definitions of the levels of measurement). In order to compute sums, deviation scores, and the amount by which Y changes per unit increase in X, it is necessary to have measures of the variables that reflect the amount by which one case differs from another case on each variable. *Nominal* variables do not provide information about the intervals between the categories of a variable nor do they rank-order the categories. Nominal variables are simply classification schemes that group units into mutually exclusive categories, where each category is defined by some attribute shared by all members of the category. Examples of nominal variables are sex (male or female), race (e.g., black, white, American Indian, or Asian), marital status (married, separated, divorced, widowed, or never-married), and employment status (employed full-time, employed part-time, not employed, or retired). For variables such as these, there is no measurement system that specifies how high or low members of one category are in comparison to members of other categories. Although it is common to assign numeric codes to the categories of nominal variables (e.g., full-time = 1, part-time = 2, not employed = 3, retired = 4), these numbers are purely arbitrary and are used only for purposes of record-keeping and data processing.

Nevertheless, nominal variables may still contain information that is useful for explaining variation in Y. The fact that these variables do not provide ordinal and interval information does not mean that there are no meaningful differences between the categories in terms of the dependent variable under investigation. Before we can use regression analysis to extract such potentially valuable information, however, it is necessary to create one or more appropriately coded variables from the nominal variable. These new variables are then included in the regression equation to represent the nominal variable. Although not just any coding scheme for these variables will work, there are several alternative schemes that will provide meaningful interpretations and tests of the effects of the nominal variable on Y. We will examine *dummy variable coding* (perhaps the most widely used coding) as well as *effects coding* and *contrast coding*.

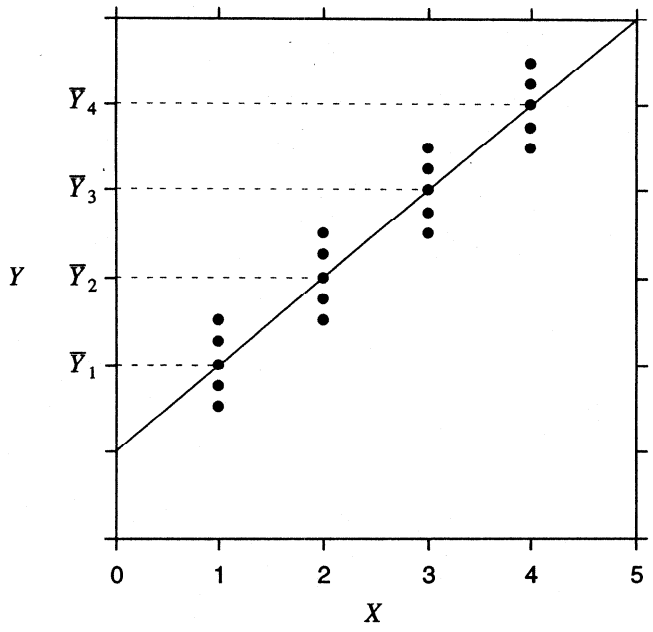
Dummy Variable Coding

Dichotomous Nominal Variables

We will begin with the simplest type of nominal variable, one that contains only two categories (i.e., a dichotomous variable). We will also begin with bivariate regression before moving on to the multivariate case. When using regression to analyze a dependent variable and a nominal independent variable, the focus is on whether the categories of the nominal variable (whether there are two or more than two) have different means on Y. For example, do blacks and whites (the nominal variable) have different income means ($\overline{Y}$'s), or do Protestants, Catholics, Jews, and atheists (the nominal variable) have different mean frequencies of church or temple attendance ($\overline{Y}$'s)? When there are only two groups/categories, a t test for the difference of means may be used to decide if the mean difference is significant; when there are two or more categories of the nominal variable, the F test provided by a one-way analysis of variance (ANOVA) may be used to determine whether there are any mean differences in Y between the groups defined by the nominal variable. Regression analysis can also be used to extract the same information and more.

In order to understand how regression analysis can be used to test for differences in means, it is useful to remember that the linear regression of Y on an interval level X is also a test for differences in means. Figure 5.1 shows a regression line fit to a scatterplot of a perfectly linear relationship between X and Y. For each value of X, there is a distribution of Y scores (a conditional distribution of Y). For each value of X, the regression line passes through the conditional mean of Y ($\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$, $\overline{Y}_4$), that is, through the mean of Y for all cases with the same value of X. The discussion of the sum of squares in Chapter 2 pointed out that *the mean is a least-squares statistic*; that is, the sum of squared deviations around the mean of a variable is smaller than the sum of squared deviations around any other value of that variable. As a result, the prediction

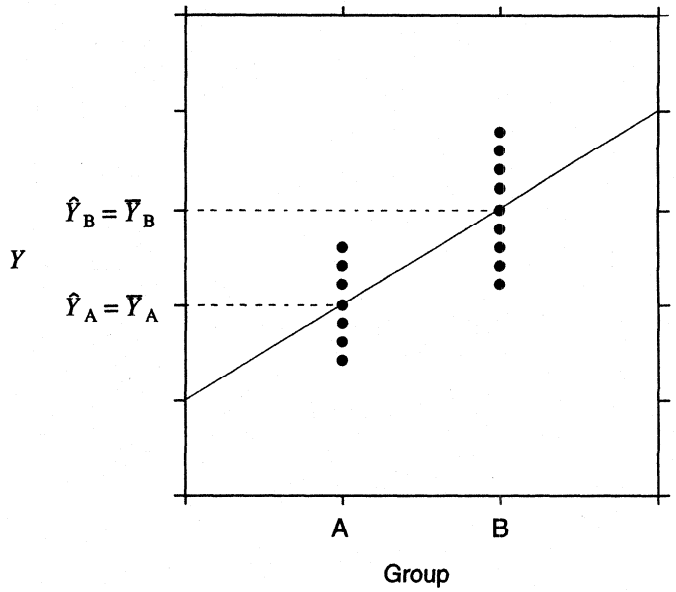
FIGURE 5.1 Scatterplot and Linear Regression Line for X and Y



of Y that minimizes the sum of squared residuals (the least-squares criterion) for all cases with the same value of X is the mean of Y for those cases (i.e., the conditional mean of Y). Thus, a perfectly linear relationship between X and Y is one in which the conditional means of Y all fall on a straight line. The test for the regression slope is actually a test for a linear trend in the conditional means of Y.

When X is a dichotomous nominal variable, there will be a conditional distribution of Y for each of the two groups defined by X. When using X to predict Y, the least-squares criterion would lead to the prediction of the conditional mean of Y for each group/category defined by X. This is shown in Figure 5.2 for the two categories of the nominal variable, which are labeled A and B. The prediction of $\bar{Y}_A$ for all cases in group A would minimize the sum of squared residuals for group A, i.e., $\Sigma (Y_A - \bar{Y}_A)^2$, and the prediction of $\bar{Y}_B$ for all cases in group B would minimize $\Sigma (Y_B - \bar{Y}_B)^2$. Thus, the least-squares regression line would pass through the two conditional means in order to minimize the total sum of squared residuals, as shown in Figure 5.2. Notice that no numeric values are shown for groups A and B. In terms of finding the best-fitting line, it is arbitrary how we score A and B. Whatever scores we assign to A and B (which

FIGURE 5.2 Least-Squares Line for Y and a Dichotomous Nominal X



we can refer to as X_A and X_B), the least-squares line will pass through the coordinates $(X_B, \overline{Y}_B)$ and $(X_A, \overline{Y}_A)$. The slope of the line will be

$$b_{YX} = \frac{\overline{Y}_B - \overline{Y}_A}{X_B - X_A} = \frac{\overline{Y}_A - \overline{Y}_B}{X_A - X_B} \tag{5.1}$$

Equation 5.1 indicates that regardless of which group is given the higher score and what the difference in the scores is, we can find a line that gives the best predictions of Y simply by dividing the difference between the means of Y for groups A and B by the difference between whatever scores we assign to groups A and B. We will see, however, that some scores that we can assign to X will give particularly meaningful interpretations of the slope.

To illustrate the above points we will examine some data from the 1982 Akron Area Survey, a random-digit-dialed telephone survey of residents of Summit County, Ohio. The dependent variable Y is a life satisfaction scale ranging from one to seven, with $Y = 1$ meaning *dissatisfied* and $Y = 7$ meaning *satisfied*. The independent variable is race, with $X = 1$ being the arbitrary code assigned to blacks and $X = 2$ being the code assigned to whites by the Akron Area Survey. Table 5.1 summarizes the values of Y for each race.

Nonsense Coding. To illustrate the arbitrariness of the coding of the race variable, *nonsense coding* will be used instead of the codes provided by the

TABLE 5.1 Life Satisfaction (Y) by Race (X)

Label	*n*	*Mean*	*St. Dev.*	*Variance*
Black	63	4.7619	2.1153	4.4747
White	656	5.5168	1.3546	1.8348
Total	719	5.4506	1.4512	2.1058

Akron Area Survey data set, which were RACE = 1 for blacks and RACE = 2 for whites. The nonsense codes that have been assigned are 74 for blacks and 19 for whites. The SPSS commands for transforming the race variable to the nonsense codes and using it in a regression run are shown below.

```
RECODE RACE (1 = 74) (2 = 19)
REGRESSION VARS = SATLIFE RACE/
  DEP = SATLIFE/ ENTER RACE
```

The output of the regression run will give a value for the least-squares slope that will be equal to the following value computed by Equation 5.1:

$$b_{YX} = \frac{\overline{Y}_B - \overline{Y}_W}{X_B - X_W} = \frac{4.7619 - 5.5168}{74 - 19} = \frac{-.7549}{55} = -.01373$$

The intercept for the equation is computed according to the usual formula. First, the mean for the nonsense-coded X must be calculated:

$$\overline{X} = \frac{n_W X_W + n_B X_B}{n} = \frac{656(19) + 63(74)}{719} = 23.8192$$

The intercept can now be computed as follows:

$$a = \overline{Y} - b\overline{X} = 5.4506 - (-.01373)23.8192 = 5.7776$$

The least-squares regression equation is thus

$$\hat{Y} = 5.7776 - .01373X$$

This equation has a negative slope, indicating that race has a negative effect on life satisfaction. The negative slope has meaning only when we remember that blacks were arbitrarily given a higher score on X than whites. Thus, the negative b reflects the fact that the group with the higher score on RACE (blacks) has lower life satisfaction. If we interpret the slope literally as the change in Y per unit increase in X, we would say that as race increases by one unit, satisfaction decreases by .01373 units (on a seven-point scale). This literal interpretation doesn't make any sense in terms of the groups that are represented by X because a one-unit difference on X doesn't correspond to the

difference between the scores of blacks and whites on X; the difference between blacks and whites equals 74 − 19 = 55, not 1. Thus, the nonsense coding leads to a nonsense interpretation of the slope. Furthermore, the intercept is meaningless because no group has a score of 0 on X.

The equation, however, does lead to meaningul predictions of Y:

$$\hat{Y}_W = 5.7776 - .01373(19) = 5.5167$$

$$\hat{Y}_B = 5.7776 - .01373(74) = 4.7616$$

The predicted score for whites equals the mean for whites, and the predicted score for blacks equals the mean for blacks (see Table 5.1). These predictions can be used to compute the coefficient of determination. In order to compute this statistic we need the sum of squared residuals for the predicted values of Y. Since the predicted score for each race equals its mean on Y, the variance of Y for each race shown in Table 5.1 equals the variance of the residuals for each race. If we multiply the variance for each race by the number of respondents of each race minus 1, we will get the sum of squared residuals for each race:

$$\sum (Y_W - \hat{Y}_W)^2 = (n_W - 1)s^2_{\hat{Y}_W} = 655(1.8348) = 1201.794$$

$$\sum (Y_B - \hat{Y}_B)^2 = (n_B - 1)s^2_{\hat{Y}_B} = 62(4.4747) = 277.4314$$

Adding the above equations together gives the total sum of squared residuals for all respondents:

$$\sum (Y - \hat{Y})^2 = 1201.7940 + 277.4314 = 1479.2254$$

The sum of squares for Y can be computed by multiplying $n - 1$ times the variance of Y shown in Table 5.1 for all respondents:

$$\sum (Y - \bar{Y})^2 = (n - 1)s^2_Y = 718(2.1058) = 1511.9644$$

The coefficient of determination, or r^2, equals

$$r^2_{YX} = \frac{\sum (Y - \bar{Y})^2 - \sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{1511.9644 - 1479.2254}{1511.9644} = .02165$$

This indicates that race explains about 2 percent of the variance in life satisfaction. This isn't a very high proportion, which indicates that there is a lot of variance in satisfaction within each race. The racial difference, however, is not trivial; blacks and whites differ by about .75 points on a seven-point scale. This difference is equivalent to about half of a standard deviation in Y. The reason r^2 is low is that blacks comprise less than ten percent of the sample; thus, there is very little variance in X with which to explain Y. This fact, however, does not

mean that the racial difference is unimportant. This is a case in which you must pay attention to factors in addition to measures of the strength of association.

Nonsense coding will also give valid significance tests (t and/or F) for the effect of the dichotomous X on Y. We can use r^2 to calculate the F statistic for the null hypothesis that $b_{YX} = 0$:

$$F = \frac{r^2_{YX}}{(1 - r^2_{YX})/n - 2} = \frac{.02165}{(1 - .02165)/717} = 15.8666$$

For $F = 15.8666$, $df_1 = 1$, and $df_2 = 718$, $p < .001$. Thus, we would reject the null hypothesis that blacks and whites are equally satisfied with life and conclude that life satisfaction is higher for whites. Since $t^2 = F$ for b, $t = \sqrt{F} = \sqrt{15.8666} = -3.9832$. (Note that although the square root of a number is always positive, the proper sign of t is negative because the slope is negative.)

We have seen that nonsense coding can be validly used with a dichotomous nominal variable; any codes will work. The coding will not affect the measures of strength of association, such as r, r^2, and B; this will be illustrated later. The regression slope will have the correct sign, which is easily interpreted if you keep in mind which group is given the highest score on X. You can make sense of the value for b if you use the difference between the scores of the two groups on X; if you multiply the slope times the difference in scores you will get the difference between the means for the two groups, i.e., $\bar{Y}_B - \bar{Y}_W = b_{YX}(X_B - X_W) = -.01373(74 - 19) = -.7552$. A final important point that will be illustrated below is that no matter what scores you give to the two groups, your decision will not affect the values of t and F in the test for the significance of the slope; t and F are invariant with respect to the scoring of X.

Dummy Coding. *Dummy variable* coding is often used as a method that provides more readily interpretable slopes and intercepts for nominal independent variables. In dummy coding, one group is given a score of 1 and the other group is given a score of 0. In this example, blacks will be given a score of 1 and whites a score of 0.

The regression slope can again be calculated with Equation 5.1, since the best prediction for each group will still be each group's mean on Y:

$$b_{YX} = \frac{\bar{Y}_B - \bar{Y}_W}{X_B - X_W} = \frac{4.7619 - 5.5168}{1 - 0} = \frac{-.7549}{1} = -.7549$$

As can be seen, *the slope for a dummy variable will equal the mean of Y for the group coded 1 on X minus the mean of Y for the group coded 0 on X.* This results because the difference between their scores on X is unity. The sign of b indicates whether the group scored 1 has a higher mean (+ slope) or a lower mean (− slope) than the group scored 0. The slope for race indicates that the mean life satisfaction for blacks is .7549 less than that for whites. The slope can also be interpreted in the usual manner; when there is a positive unit difference

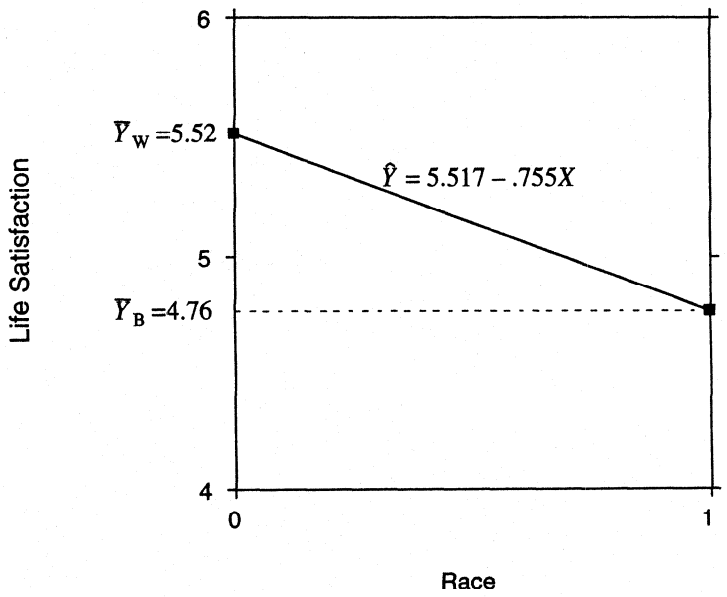
in X (which corresponds to the racial difference in values of X) there will be a difference in Y equal to b (i.e., $-.7549$).

Since the least-squares equation predicts the conditional mean of Y for each of the two groups represented by a dummy variable X, the line will pass through the coordinates $(0, \overline{Y}_W)$ and $(1, \overline{Y}_B)$. The first coordinate corresponds to $X = 0$; thus, the predicted value when $X = 0$ is $\overline{Y}_W$, which by definition is the Y-intercept. *The intercept in a dummy variable regression is therefore the mean of the group coded 0 on the dummy variable* (in this case, whites). Thus, the intercept equals 5.5168. This makes the value of the intercept a meaningful quantity. Having determined the slope and intercept, we can write the dummy variable equation for race and life satisfaction as

$$\hat{Y} = 5.5168 - .7549X$$

Figure 5.3 shows the least-squares line for life satisfaction regressed on the dummy race variable. The line crosses the Y-axis at a value equal to the mean for whites, the group coded 0 on X. The figure also shows that the slope of the line equals the mean of the group scored 1 (blacks) minus the mean of the group scored 0 (whites).

FIGURE 5.3 Life Satisfaction (Y) Regressed on Dummy-Coded Race (X)



We can verify that the intercept equals the mean of the group scored 0 by using the usual formula for the intercept. First, we need to compute the mean of X, which is

$$\bar{X} = \frac{\sum X}{n} = \frac{1 + 1 + \cdots + 1 + 0 + 0 + \cdots + 0}{n} = \frac{n_B}{n} = \frac{63}{719} = .0876 = p$$

Since the scores on X are either 1's or 0's, the sum of X will equal the number of cases scored 1 on X. In this case, that will be the number of blacks. When the sum is divided by n, it is shown that the mean equals a proportion, i.e., the proportion of all cases with a score of 1. In the above formula the mean is .0876, which indicates that .0876 of the respondents are black. In general, the mean of a dummy variable equals the proportion of cases that are in the group coded 1. Without proof, the standard deviation of a dummy variable can be computed with the following formula:

$$s_X = \sqrt{p(1 - p)} = \sqrt{(.0876)(.9124)} = \sqrt{.0799} = .2827$$

Thus, the standard deviation of the dummy race variable is .2827, and the variance (the number under the radical) is .0799; the variance equals $p(1 - p)$. The intercept of the least-squares line can now be computed as

$$a = \bar{Y} - b\bar{X} = 5.4506 - (-.7549)(.0876) = 5.5167$$

This verifies that the intercept equals the mean of the group scored 0. The same value for the intercept is reported in the SPSS output shown in Table 5.2.

TABLE 5.2 Regression of Life Satisfaction on Dummy-Coded Race

```
RECODE RACE (1 = 1) (2 = 0)
REGRESSION VARS = SATLIFE RACE/
  DEP = SATLIFE/ ENTER RACE

                    ****MULTIPLE REGRESSION****

Multiple R              .14718  Analysis of Variance
R Square                .02166                  DF  Sum of Squares   Mean Square
Adjusted R Square       .02030  Regression       1       32.75310      32.75310
Standard Error         1.43635  Residual       717     1479.24412       2.06310

                        F = 15.87566            Signif F = .0001

---------------------Variables in the Equation-----------------------------------

Variable            B          SE B          Beta           T          Sig T
RACE             -.75486        .18945       -.14718       -3.984        .0001
(Constant)       5.51677        .05608                     98.373        .0000
```

Table 5.2 gives the SPSS statements that may be used to run a dummy variable regression of life satisfaction on race, along with the output of the SPSS run. The values for the bivariate slope and the intercept shown in Table 5.2 are virtually identical to those calculated previously.

To state in a more general form the characteristics of dummy variable regression that have been covered thus far, let

$$Y_0 = Y \text{ scores when } X = 0$$

$$Y_1 = Y \text{ scores when } X = 1$$

The least-squares solution for a dummy-coded X always gives the following values for the slope and intercept:

$$b_{YX} = \overline{Y}_1 - \overline{Y}_0 \quad \textbf{(5.2)}$$

$$a_{YX} = \overline{Y}_0 \quad \textbf{(5.3)}$$

As was implied in the discussion of nonsense coding, dummy variable coding will give the same values of t, F, r, and r^2 as nonsense coding or any other coding. This is illustrated in the SPSS regression analysis shown in Table 5.2, where we can see that F and r^2 are nearly identical to the values that we calculated earlier for nonsense coding. The dummy variable solution differs from the solutions using other coding only with respect to the slope and intercept of the regression equation.

It should also be noted that the t test for b when X is dummy-coded is equivalent to a t test for the difference of means. This is because the slope for a dummy X equals the difference between the two groups' means on Y (as shown above). Thus, the following hypotheses apply for a bivariate regression with a dummy X:

$$H_0: b = 0 \text{ or } \overline{Y}_1 = \overline{Y}_0$$

$$H_1: b \neq 0 \text{ or } \overline{Y}_1 \neq \overline{Y}_0$$

The equivalence between dummy regression for a dichotomous X and the t test for a difference of two group means is illustrated in Appendix 5A, where the output of the SPSS procedure T-TEST is discussed.

Polychotomous Nominal Variables

When a nominal variable has more than two categories, we can no longer use a single dummy variable to represent all of the categories. We cannot use a single variable coded with three or more scores, either, because a regression program would treat the numeric codes as if they represented an interval variable. The data in Table 5.3 will be used to illustrate the analysis of nominal variables with three or more categories; these data are also from the 1982 Akron Area Survey. The dependent variable is again life satisfaction, and the nominal

TABLE 5.3 Life Satisfaction by Marital Status

Code	*Status*	n	*Mean*
1	Married	411	5.6837
2	Divorced/Separated	106	4.9057
3	Widowed	69	5.4348
4	Never-Married	150	5.1600
	Total	736	5.4410

independent variable is marital status (MARITAL). In Table 5.3, the separated and divorced categories in the original data have been merged because of the small number of separated respondents.

If we were to use a single variable coded as in Table 5.3, regression analysis would assume that never-married are one unit higher than widowed, that widowed are one unit higher than divorced, and that divorced are one unit higher than married. This would be an inappropriate scoring since the variable is nominal. The solution to the scoring problem is to use dummy variables. It is not possible, however, to represent the marital variable with just one dummy variable. But how many dummy variables will be needed? There are $g = 4$ dummy variables that *could* be created, where g equals the number of categories in a nominal variable. One dummy could be created for each marital status category, where each category would have a score of 1 on its dummy variable and all the other categories would have a 0 score on that variable. The coding would look like this:

The $g = 4$ Possible Dummy Variables for Marital Status

Marital Status	X_1 DIVORCED	X_2 WIDOWED	X_3 NEVER	X_4 MARRIED
Divorced/Separated	1	0	0	0
Widowed	0	1	0	0
Never-Married	0	0	1	0
Married	0	0	0	1

Each person that is divorced would have a score of 1 on a dummy variable called DIVORCED; all other persons would have a score of 0 on DIVORCED. Thus, this variable would represent two groups, divorced and not divorced. The DIVORCED variable would make no distinction between the various categories that are not divorced, i.e., the widowed, never-married, and married. The widowed persons would have a score of 1 on WIDOWED, and all of the other groups would have a score of 0. Thus, WIDOWED would represent two

groups, widowed and not widowed. The dummy variables NEVER and MARRIED would be interpreted in an analogous fashion.

Although there are g distinct dummy variables that could be created, all we need for the regression equation is $g - 1$ dummy variables. In fact, the regression coefficients cannot be computed for an equation that includes all of the g dummy variables because there will be perfect multicollinearity among the dummy variables in that case. This is because of the following identity:

$$\text{DIVORCED} + \text{WIDOWED} + \text{NEVER} + \text{MARRIED} = 1$$

Since each person has a score of 1 on one and only one of the dummies, and a score of 0 on all of the other dummies, the sum of the four dummies will equal 1 for each person. If we rearrange the above equation by leaving the MARRIED variable, for example, to the left of the equal sign and moving all of the other variables to the right, we have

$$\text{MARRIED} = 1 - \text{DIVORCED} - \text{WIDOWED} - \text{NEVER}$$

where

$$0 = 1 - 1 - 0 - 0 \quad \text{(for divorced persons)}$$
$$0 = 1 - 0 - 1 - 0 \quad \text{(for widowed persons)}$$
$$0 = 1 - 0 - 0 - 1 \quad \text{(for never-married)}$$
$$1 = 1 - 0 - 0 - 0 \quad \text{(for married persons)}$$

The above shows that the MARRIED dummy variable is a perfectly linear function of the other three dummy variables. Thus, $R^2 = 1$ for MARRIED as the dependent variable and the other three dummies as predictors. There would be perfect multicollinearity among the four variables, and thus it would not be possible to estimate their separate effects on Y.[1]

The above discussion leads to the conclusion that one of the g dummies must be left out of the equation. In general, only $g - 1$ dummy variables can be used. The group represented by the dummy that will be omitted will not, however, be left out of the analysis. It will serve as a reference group, as we shall see. The dummy variable to be omitted depends on which group should theoretically serve as a reference group. In the following analysis the married respondents have been chosen to be the reference group, and thus the MARRIED dummy will not be included in the equation. This decision is based on the premise that the married status has the greatest normative support, and thus it would be of interest to compare each of the deviant marital statuses with the married group.

The omission of the MARRIED dummy leaves the following three dummies for entry into the regression equation:

---

1. The cross-product matrix $\mathbf{x'x}$ used in the matrix solution for the multiple regression equation would be singular and thus could not be inverted.

The $g - 1 = 3$ Dummy Variables Selected for the Regression Equation

Marital Status	X_1 DIVORCED	X_2 WIDOWED	X_3 NEVER
Divorced/Separated	1	0	0
Widowed	0	1	0
Never-Married	0	0	1
Married	0	0	0

Notice that married respondents will have a 0 on each of the three dummy variables. The regression equation to be estimated is

$$\hat{Y} = a + b_1\text{DIVORCED} + b_2\text{WIDOWED} + b_3\text{NEVER}$$

Analogously to bivariate dummy regression, the solution that minimizes the sum of squared residuals of Y will be to select coefficients such that no matter which marital status a person is in, including the currently married group, the equation will predict the mean of Y for the group to which the person belongs. The following symbols will be used for these conditional means of Y:

$$\overline{Y}_M = \text{Mean for Married Group}$$
$$\overline{Y}_D = \text{Mean for Divorced Group}$$
$$\overline{Y}_W = \text{Mean for Widowed Group}$$
$$\overline{Y}_N = \text{Mean for Never-Married Group}$$

If the least-squares solution predicts the conditional mean of Y for the group to which each person belongs, then we can interpret each of the regression coefficients by substituting the scores for each group into the equation and simplifying as follows:

$$\hat{Y}_M = a + b_1(0) + b_2(0) + b_3(0) = a = \overline{Y}_M$$
$$\hat{Y}_D = a + b_1(1) + b_2(0) + b_3(0) = a + b_1 = \overline{Y}_D$$
$$\hat{Y}_W = a + b_1(0) + b_2(1) + b_3(0) = a + b_2 = \overline{Y}_W$$
$$\hat{Y}_N = a + b_1(0) + b_2(0) + b_3(1) = a + b_3 = \overline{Y}_N$$

Since the married group has a 0 score for each dummy variable, the above results show that when 0 is substituted for each variable, the predicted score for the married group is a, the intercept. Since the best prediction for the married group is the mean of Y for all married persons, we can see that the intercept equals the mean of the married group, which is the reference group. With respect to the divorced group, that has a score of 1 on the first dummy variable and 0 on the others, the predicted score is the intercept plus the slope for the

DIVORCE dummy variable; thus, $a + b_1$ is equal to the mean for the divorced group. The predicted scores for widowed and never-married are derived in analogous fashion.

We can use the above predicted scores and the means from Table 5.3 to specify the *meaning* of each regression coefficient, and their numeric values, as follows:

$$\overline{Y}_M = a = 5.6837 \quad \textbf{(5.4)}$$

$$\overline{Y}_D - \overline{Y}_M = (a + b_1) - a = b_1 = 4.9057 - 5.6837 = -.778 \quad \textbf{(5.5)}$$

$$\overline{Y}_W - \overline{Y}_M = (a + b_2) - a = b_2 = 5.4348 - 5.6837 = -.2489 \quad \textbf{(5.6)}$$

$$\overline{Y}_N - \overline{Y}_M = (a + b_3) - a = b_3 = 5.1600 - 5.6837 = -.5237 \quad \textbf{(5.7)}$$

Equation 5.5 shows that the slope for the DIVORCE variable (b_1) equals the difference between the mean for the divorced group and the mean for the married group. The value of the slope indicates that divorced persons are .778 less satisfied with their lives than married persons. Equation 5.6 indicates that the slope for WIDOWED (b_2) equals the widowed mean minus the married mean and indicates that widowed persons are .2489 less satisfied than married persons. Finally, according to Equation 5.7, the slope for NEVER (b_3) equals the never-married mean minus the married mean and indicates that never-married persons are .5237 less satisfied than married persons.

The least-squares regression equation has thus been shown to be equal to

$$\hat{Y} = 5.6837 - .778 \cdot \text{DIVORCED} - .2489 \cdot \text{WIDOWED} - .5237 \cdot \text{NEVER}$$

Although each dummy variable represents the group coded 1 versus all other groups that are coded 0, the slope for the dummy doesn't represent the mean difference between the group coded 1 and all persons coded 0 as it does in the bivariate case. These slopes are partial slopes, so they indicate the change in Y per unit change in the dummy variable, holding all other dummy variables constant. Thus, for example, when DIVORCE scores change by one unit but WIDOWED and NEVER are held constant, the unit change in DIVORCE represents the difference between married persons and divorced persons. When WIDOWED changes by one unit and DIVORCED and NEVER are held constant, the change in WIDOWED represents the difference between widowed and married persons. Thus, *each partial slope equals the mean of Y for the group coded 1 on that dummy variable minus the mean of Y for the group coded 0 on all of the dummies*; the latter group is referred to as the reference group, which in this case is the married group. *The intercept a equals the mean of Y for the reference group.*

In general, if the regression equation for a set of $g - 1$ dummy variables is $\hat{Y} = a + \Sigma b_i X_i$ ($i = 1, \ldots, g - 1$), where X_i is the dummy variable for the ith

TABLE 5.4 Life Satisfaction Regressed on the Dummy-Coded Marital-Status Variables

****MULTIPLE REGRESSION****

Multiple R	.20640	Analysis of Variance			
R Square	.04260		DF	Sum of Squares	Mean Square
Adjusted R Square	.03868	Regression	3	66.43387	22.14462
Standard Error	1.42818	Residual	732	1493.05390	2.03969

F = 10.85685 Signif F = .0000

Variables in the Equation

Variable	B	SE B	Beta	T	Sig T
NEVER	−.52370	.13624	−.14493	−3.844	.0001
WIDOWED	−.24892	.18581	−.04984	−1.340	.1808
DIVORCED	−.77804	.15558	−.18767	−5.001	.0000
(Constant)	5.68370	.07040		80.681	.0000

group (i.e., the code of the ith group equals unity on X_i), then the intercept and slopes for the dummy variable regression equation are

$$a = \overline{Y}_0 \tag{5.8}$$

$$b_i = \overline{Y}_i - \overline{Y}_0 \tag{5.9}$$

where $\overline{Y}_0$ equals the mean of the reference group (i.e., the group coded 0 on each X_i).

The following SPSS commands may be used to transform the marital-status nominal variable, which has four categories, into the three dummy variables that have been chosen for entry into the regression equation.

```
IF MARITAL EQ 1 MARRIED = 1
IF MARITAL NE 1 MARRIED = 0
IF MARITAL EQ 2 DIVORCED = 1
IF MARITAL NE 2 DIVORCED = 0
IF MARITAL EQ 3 WIDOWED = 1
IF MARITAL NE 3 WIDOWED = 0
IF MARITAL EQ 4 NEVER = 1
IF MARITAL NE 4 NEVER = 0
REGRESSION VARS = SATLIFE DIVORCED WIDOWED NEVER/
  DEP = SATLIFE/ ENTER DIVORCED WIDOWED NEVER
```

The output from the SPSS regression run is given in Table 5.4. Notice that the partial slopes and the intercept given in Table 5.4 are exactly equal to those values computed from the group means using Equations 5.1 to 5.4.

The hypotheses that are tested by the t and F statistics in Table 5.4 are

$$F \text{ for } R^2 \qquad H_0: b_1 = b_2 = b_3 = 0 \text{ or } \overline{Y}_M = \overline{Y}_D = \overline{Y}_W = \overline{Y}_N$$

$$t \text{ for } b_i \qquad H_0: b_i = 0 \text{ or } \overline{Y}_i = \overline{Y}_M$$

The F value and associated level of significance for the test that all slopes are equal to 0 or that the means are equal for all groups is rejected at $p = .0000$. Therefore, at least one of the marital statuses has a different mean satisfaction level than the others. The squared multiple correlation indicates that we can account for about 4.3% of the variance in life satisfaction simply by knowing whether a person is married, divorced/separated, widowed, or never-married. The t tests for the individual slopes indicates that the never-married and the divorced are significantly less satisfied with their lives than the married group, but there is not a significant difference between the widowed and married persons.

When dummy variables are used to represent a nominal independent variable with three or more categories, the F test for R^2 in the regression equation is equivalent to the F test for group differences in means performed by a one-way analysis of variance (ANOVA). See Appendix 5B for a description of an equivalent analysis of variance.

Inclusion of Another *X* with the Dummy Variables

If we add another X to the equation containing the dummy variables, the interpretation changes somewhat from that given above. For purposes of this discussion, the new X might be an ordinal, interval, or ratio variable, or it might even be a dummy variable representing an additional nominal variable. In the new equation, the intercept equals the predicted value of Y when all the variables are equal to zero, including the new X that has been added. The value of a no longer equals the mean for the reference group; instead, it is the predicted value for members of the reference group who have a score of zero on the new X. The slopes no longer equal the difference between the mean of the group coded 1 and the mean of the reference group, since the new X is now being held constant when the groups are contrasted; the partial slope now equals the predicted difference between the group coded 1 and the reference group for persons who do not differ on the new X.

Table 5.5 shows the output of an SPSS run that adds the variable AGE to the dummy variables already in the equation. The intercept is not a meaningful value in this equation because it represents the predicted score for married persons who are zero years of age. The partial slopes for each dummy variable now represent the difference between a particular marital status and the married group, holding age constant. The partial slopes, however, still represent a comparison between the group coded 1 on the dummy variable and the ref-

TABLE 5.5 Life Satisfaction Regressed on the Dummy-Coded Marital Status Variables and Age

****MULTIPLE REGRESSION****

Multiple R	.23153	Analysis of Variance			
R Square	.05361		DF	Sum of Squares	Mean Square
Adjusted R Square	.04831	Regression	4	81.91767	20.47942
Standard Error	1.42319	Residual	714	1446.17691	2.02546

F = 10.11101 Signif F = .0000

----------Variables in the Equation----------

Variable	B	SE B	Beta	T	Sig T
NEVER	−.35075	.15074	−.09728	−2.327	.0203
WIDOWED	−.47926	.21054	−.09556	−2.276	.0231
DIVORCED	−.75238	.15611	−.18226	−4.820	.0000
AGE	.01197	.00406	.13499	2.949	.0033
(Constant)	5.18099	.18729		27.663	.0000

erence group. Thus, the slope for DIVORCED indicates that when any age difference between divorced/separated persons and married persons is held constant, divorce or separation reduces life satisfaction by about .75 points. Although the DIVORCED and NEVER dummy variables are still significant after age is controlled, the negative effect of having never been married is reduced from −.52 to −.35. Apparently part of the original difference between never-married persons and married persons was due to the fact that people who have never been married are younger, on the average, than married individuals. We also see that after controlling for age, WIDOWED now also has a significant negative effect on life satisfaction; the fact that the average age of widowed persons is greater than the age of married persons had been *suppressing* the true negative effect of widowhood on life satisfaction. In sum, after controlling for age we can now infer that if the three nonmarried groups did not differ in age from the married group, being divorced, widowed, or never-married would each reduce satisfaction with life.

It was noted above when a single nominal variable is converted to dummy variables for inclusion in a regression equation, the F test for the regression equation is equal to the F test in a one-way analysis of variance for the same nominal variable. We now note that the equivalence between regression analysis and analysis of variance continues when an interval X (or X's) is added to the regression equation containing the dummy variables; the F tests in the regression analysis are now equal to the F tests in an analysis of covariance (ANCOVA). An interval variable such as age is called a covariate in analysis of variance. The equivalence of regression analysis and analysis of covariance is shown in Appendix 5B.

Effects Coding

Dummy variable coding requires that one category of the nominal variable be specified as a reference group against which each of the other categories are tested. This is often a very satisfactory procedure for testing differences between groups. There may be times, however, when it is difficult to choose one group with which to compare each of the others. There might be two or more categories that would serve equally well from a theoretical standpoint as reference groups, or there may not be any theoretical reasons for specifying any of the groups as a reference group, as in an exploratory study, for example. In both of these cases the analyst may want to consider *effects coding*.

An example of effects coding for the marital status variable is given below.

Effects Coding

Marital Status	X_1 DIVORCED	X_2 WIDOWED	X_3 NEVER
Divorced/Separated	1	0	0
Widowed	0	1	0
Never-Married	0	0	1
Married	−1	−1	−1

As can be seen above, the only difference between effects coding and dummy variable coding is that one of the groups is coded −1 on each variable instead of 0. Just as with dummy coding, there will be $g - 1$ effects-coded variables.

Effects coding, however, changes the meaning of the regression coefficients. In effects coding the intercept will equal the mean of the group means on the dependent variable,

$$a = \bar{\bar{Y}} = \frac{\bar{Y}_1 + \bar{Y}_2 + \cdots + \bar{Y}_g}{g} \tag{5.10}$$

If there are g groups, the intercept will equal the mean of the g group means. This mean of means is indicated by $\bar{\bar{Y}}$ and is often called the *unweighted grand mean*. In the case of marital status and life satisfaction, the grand mean of the $g = 4$ group means given in Table 5.3 is

$$a = \bar{\bar{Y}} = \frac{\bar{Y}_M + \bar{Y}_D + \bar{Y}_W + \bar{Y}_N}{4} = \frac{5.6837 + 4.9057 + 5.4348 + 5.1589}{4} = 5.2958$$

Notice that the grand mean is smaller than the mean given in Table 5.3, which is the regular mean of Y computed as the arithmetic average of all 737 observations. The arithmetic mean of Y is influenced greatly by the higher mean of the married persons who comprise the majority of the respondents. The married persons, however, do not carry any more weight than the unmarried persons

in the computation of the grand mean. Thus, the grand mean is lower because the smaller unmarried groups all have lower life satisfaction than the married group.

In effects coding, *the unstandardized regression slope for an effect-coded variable equals the difference between the mean of the group coded 1 on the variable and the grand mean of all groups.* Thus, the unstandardized slope for group i equals

$$b_i = \overline{Y}_i - \overline{\overline{Y}} \tag{5.11}$$

Instead of comparing each group to a single reference group, effects-coded variables compare each group to the mean of all groups. In the case of marital status and satisfaction, we have

$$b_D = \overline{Y}_D - \overline{\overline{Y}} = 4.9057 - 5.2958 = -.3901 \quad \text{[divorced and separated]}$$
$$b_W = \overline{Y}_W - \overline{\overline{Y}} = 5.4348 - 5.2958 = .1390 \quad \text{[widowed]}$$
$$b_N = \overline{Y}_N - \overline{\overline{Y}} = 5.1589 - 5.2958 = -.1369 \quad \text{[never-married]}$$

In sum, the regression equation with effects-coded variables for each unmarried group is

$$\hat{Y} = 5.2958 - .3901\text{DIVORCED} + .1390\text{WIDOWED} - .1369\text{NEVER}$$

If an effects-coded variable had been included for married respondents in place of one of the unmarried groups, the slope would be

$$b_M = \overline{Y}_M - \overline{\overline{Y}} = 5.6837 - 5.2958 = .3879 \quad \text{[married]}$$

But just as with dummy-coded variables, only $g - 1$ variables can be included in the equation to represent the g categories of the nominal variable. The fact that one group is not coded 1 on any of the variables (and is coded -1 on all of the variables) does not mean it is not influencing the regression equation coefficients. The married group, for example, influences the grand mean and consequently influences the intercept and each of the unstandardized slopes.

The null hypothesis to be tested by either t or F for each effects-coded variable is

$$H_0: b_i = 0; \quad \text{or } \overline{Y}_i = \overline{\overline{Y}}$$

However, since the mean of group i is included in the grand mean, the only way that the null hypothesis can be true is for the mean of group i to be equal to the mean of all the other groups, excluding group i. Thus, the null hypothesis can also be stated as

$$H_0: \overline{Y}_i = \overline{\overline{Y}}_{(-i)}$$

The subscript $(-i)$ means that it is the mean of the means of all groups except

TABLE 5.6 Life Satisfaction Regressed on the Effects-Coded Marital Status Variables

****MULTIPLE REGRESSION****

Multiple R	.20640	Analysis of Variance			
R Square	.04260		DF	Sum of Squares	Mean Square
Adjusted R Square	.03868	Regression	3	66.43387	22.14462
Standard Error	1.42818	Residual	732	1493.05390	2.03967

F = 10.85685 Signif F = .0000

----------Variables in the Equation----------

Variable	B	SE B	Beta	T	Sig T
NEVER	−.13604	.10493	−.07456	−1.297	.1952
WIDOWED	.13875	.13781	.06296	1.007	.3144
DIVORCED	−.39038	.11761	−.19537	−3.319	.0009
(Constant)	5.29604	.06489		81.620	.0000

group *i*. Thus, for each effects-coded variable, we are testing whether the mean of one of the groups differs from the mean of the means of all other groups.

SPSS statements that may be used to create the effects-coded variables are shown below.

```
IF MARITAL EQ 2 DIVORCED = 1
IF MARITAL NE 2 DIVORCED = 0
IF MARITAL EQ 3 WIDOWED = 1
IF MARITAL NE 3 WIDOWED = 0
IF MARITAL EQ 4 NEVER = 1
IF MARITAL NE 4 NEVER = 0
IF MARITAL EQ 1 DIVORCED = -1
IF MARITAL EQ 1 WIDOWED = -1
IF MARITAL EQ 1 NEVER = -1
```

The output from the SPSS regression program is shown in Table 5.6. The results for the multiple correlation statistics, the standard error of estimate, and the analysis of variance for the entire equation are the same as for the dummy variable regression in Table 5.4. Whether dummy coding or effects coding is used, the predicted values of the dependent variable will be identical because in each case three variables are being used to represent the conditional means of the four marital status groups. Thus, the *F* test for the entire equation indicates that the null hypothesis that all of the conditional means are equal should be rejected.

The *t* tests for the three effects-coded independent variables are not identical to those for the dummy-coded variables. The test for NEVER in Table 5.4 indi-

cated that the never-married respondents had significantly lower life satisfaction than the married respondents. But the effects-coded NEVER is not significant at $p < .05$; the mean for never-married respondents is not less than the mean of the means for the other three groups (the "ever" married respondents). Table 5.4 showed that the widowed respondents did not differ significantly from the married respondents in satisfaction with their lives, and the effects coding indicates that the widowed do not differ significantly from the mean of the means of the never-married, divorced, and married respondents. Both dummy and effects coding indicate that the slope for the DIVORCED variable is significantly different from zero; the divorced group has the lowest life satisfaction. If we had chosen to include an effects-coded variable for married respondents in place of any of the other effects-coded variables, it would probably be significant at the .05 level since the married respondents have the highest life satisfaction. This indicates that the number of variables for which the null hypothesis is rejected may depend on which group is used as the reference group. If the mean of the group coded -1 on each of the $g - 1$ variables is near the middle of the group means, there may be more significant variables than if the group coded -1 has one of the highest or lowest means. The number of significant results, of course, may also depend on which category is selected as the reference group in dummy variable coding.

Contrast Coding

Effects coding is atheoretical in that the investigator does not specify which groups are to be tested for differences of means but instead simply compares each group with all of the others together. Dummy coding gives the investigator more control over which groups are to be compared, but all tests for differences of means are between a single reference group and each of the other groups. The final coding scheme to be considered—*contrast coding*—provides the researcher with the most control and flexibility in specifying the group comparisons or contrasts that are to be tested.

As with effects coding and dummy coding, contrast coding involves creating $g - 1$ variables to represent the g categories of the nominal variable. For each of these variables the investigator specifies one subset of groups that is to be contrasted with a second subset of groups. Each subset may consist of only one group. Each of these variables is referred to as a *contrast*. Thus, for example, three contrasts would be specified or created for the four categories of marital status.

For each of the $g - 1$ contrast variables a set of codes c_i ($i = 1, 2, \ldots, g$) must be selected for the g groups. There are two sets of restrictions on the c_i. First, the codes for each contrast variable must sum to zero:

$$\sum_{i=1}^{g} c_i = 0$$

This means, of course, that some of the groups must have positive codes and some must have negative codes. Some may also be zero. Furthermore, the codes for the subset of groups with positive codes must all be equal, and the codes for the subset of groups with negative codes must all be equal.[2] The consequence of these restrictions, in combination with the second type of restriction to be specified below, is that *contrast coding implements a contrast between the subset of groups with positive codes and the subset of groups with negative codes.*

The second restriction is that

$$\sum_{i=1}^{g} c_{ij}c_{ik} = 0, \qquad j \neq k$$

This restriction means that the sum of products of the c_i for each pair of contrast-coded variables must equal zero. If the sum of products of the codes equals zero, the codes are said to be *orthogonal*, or uncorrelated. This does not mean that the coded variables themselves will be orthogonal; that is, when the correlation between the variables is calculated across all n observations, it need not be zero. Only when each group has an equal number of cases will orthogonal coding produce orthogonal variables.

An example of contrast coding for the marital status variable follows.

Contrast Coding

Marital Status	X_1 LOSS	X_2 WIDOWED	X_3 NEVER
Divorced/Separated	$+\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{4}$
Widowed	$+\frac{1}{3}$	$+\frac{1}{2}$	$-\frac{1}{4}$
Never-Married	0	0	$+\frac{3}{4}$
Married	$-\frac{2}{3}$	0	$-\frac{1}{4}$
Sum	0	0	0

As indicated above, the sum of the c_i for each variable equals zero. There are three pairs of variables, and the sums of cross-products of the codes for these variables are

2. This restriction applies only when contrasts are being specified for a purely nominal variable. When contrasts are specified for ordinal/interval categorical variables, sometimes referred to as *orthogonal polynomial contrasts*, only the restriction that the c's must sum to zero holds.

$$\sum_{i=1}^{4} c_{i1}c_{i2} = \left(\frac{1}{3}\right)\left(-\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) + (0)(0) + \left(-\frac{2}{3}\right)(0)$$

$$= -\frac{1}{6} + \frac{1}{6} + 0 + 0 = 0$$

$$\sum_{i=1}^{4} c_{i1}c_{i3} = \left(\frac{1}{3}\right)\left(-\frac{1}{4}\right) + \left(\frac{1}{3}\right)\left(-\frac{1}{4}\right) + (0)\left(\frac{3}{4}\right) + \left(-\frac{2}{3}\right)\left(-\frac{1}{4}\right)$$

$$= -\frac{1}{12} - \frac{1}{12} + 0 + \frac{2}{12} = 0$$

$$\sum_{i=1}^{4} c_{i2}c_{i3} = \left(-\frac{1}{2}\right)\left(-\frac{1}{4}\right) + \left(\frac{1}{2}\right)\left(-\frac{1}{4}\right) + (0)\left(\frac{3}{4}\right) + (0)\left(-\frac{1}{4}\right)$$

$$= \frac{1}{8} - \frac{1}{8} + 0 + 0 = 0$$

Each sum of cross-products equals zero, and thus the codes are orthogonal. Since the restrictions for contrast coding are satisfied, these variables are contrasts.

The first variable contrasts divorced and widowed persons (+ codes) with married persons. The variables have been named after the groups with positive codes. Since the two groups with positive codes on the first variable have both lost spouses, this variable has been named LOSS. Thus, this contrast is specified in order to estimate the effect of having lost a spouse. The second variable is WIDOWED, and it contrasts widowed persons (+ codes) with divorced persons (− codes) to examine the effect of the two different ways of losing a spouse. The third variable (NEVER) contrasts never-married persons (+ code) with married, widowed, and divorced persons (− codes) in order to estimate the effect of never having been married.

Let $g_{(+)}$ be the number of groups in the positive subset for a contrast variable, and let $g_{(-)}$ be the number of groups in the negative subset. The values of the positive codes c^+ and the negative codes c^- were chosen as follows:

$$c^+ = +\frac{g_{(-)}}{g_{(+)} + g_{(-)}}$$

$$c^- = -\frac{g_{(+)}}{g_{(+)} + g_{(-)}}$$

For example, for the NEVER variable the codes are

$$c^+ = +\frac{3}{1 + 3} = +\frac{3}{4}$$

$$c^- = -\frac{1}{1+3} = -\frac{1}{4}$$

These rules satisfy the restriction that the sum of the c_i's must equal zero. They also provide for the following interpretation of the unstandardized regression coefficients:

$$b_i = \bar{\bar{Y}}_{(+)} - \bar{\bar{Y}}_{(-)} \tag{5.12}$$

Equation 5.12 indicates that the regression slope for each contrast-coded variable will equal the mean of the means for the groups in the positive subset minus the mean of the means for the groups in the negative subset:

$$b_1 = \frac{\bar{Y}_D + \bar{Y}_W}{g_{(+)}} - \bar{Y}_M = \frac{4.9057 + 5.4348}{2} - 5.6887 = 5.1703 - 5.6887 = -.5185$$

$$b_2 = \bar{Y}_W - \bar{Y}_D = 5.4348 - 4.9057 = .5291$$

$$b_3 = \bar{Y}_N - \frac{\bar{Y}_M + \bar{Y}_D + \bar{Y}_W}{g_{(-)}} = 5.1589 - \frac{5.6837 + 4.9057 + 5.4348}{3}$$

$$= 5.1589 - 5.3414 = -.1825$$

This interpretation of the slopes comes from the fact that the difference between the positive code and the negative code for each variable is unity. Thus, an increase of one unit on the independent variable is equivalent to changing from a level represented by the code of the negative subset to a level represented by the code of the positive subset. It is not necessary to assign codes in this manner (as long as the sum of the codes equals zero and the sum of cross-products of codes equals zero), but it provides an easy and meaningful interpretation of the slopes.

The regression intercept for a contrast-coded set of variables is equal to the unweighted mean of means, i.e., the grand mean, just as it was in effects coding:

$$a = \bar{\bar{Y}} = \frac{\sum_{i=1}^{g} \bar{Y}_i}{g} \tag{5.13}$$

Thus, $a = 5.2958$, just as in effects coding. The regression equation for the above contrast-coded version of the marital status variable is

$$\hat{Y} = 5.2958 - .5185\text{LOSS} + .5291\text{WIDOWED} - .1825\text{NEVER}$$

The null hypothesis to be tested by the t or F test for the slope is

$$H_0: b_i = 0; \quad \text{or } \bar{\bar{Y}}_{(+)} = \bar{\bar{Y}}_{(-)}$$

The output from the SPSS regression program is shown in Table 5.7. The

TABLE 5.7 Life Satisfaction Regressed on Contrast-Coded Marital Status Variables

****MULTIPLE REGRESSION****

Multiple R	.20640	Analysis of Variance			
R Square	.04260		DF	Sum of Squares	Mean Square
Adjusted R Square	.03868	Regression	3	66.43387	22.14462
Standard Error	1.42818	Residual	732	1493.05390	2.03967

F = 10.85685 Signif F = .0000

----------Variables in the Equation----------

Variable	B	SE B	Beta	T	Sig T
NEVER	−.18138	.13990	−.05019	−1.297	.1952
WIDOWED	.52912	.22091	.08815	2.395	.0169
LOSS	−.51348	.13101	−.15325	−3.919	.0000
(Constant)	5.29604	.06489		81.620	.0000

multiple correlation statistics, the standard error of estimate, and the analysis of variance results for the entire equation are identical to those for dummy coding and effects coding. The tests for the individual contrasts show that the never-married persons are not significantly less satisfied with their lives than the ever-married persons (married, divorced, and widowed). Widowed respondents, however, are significantly more satisfied than divorced respondents, and those who have lost their spouses (divorced and widowed) are less satisfied with their lives than currently married individuals.

An example of an alternate set of contrast codes for marital status is given below.

Contrast Coding

Marital Status	X_1 MARRIED	X_2 NEVER	X_3 WIDOWED
Divorced/Separated	$-\frac{1}{4}$	$-\frac{1}{3}$	$-\frac{1}{2}$
Widowed	$-\frac{1}{4}$	$-\frac{1}{3}$	$+\frac{1}{2}$
Never-Married	$-\frac{1}{4}$	$+\frac{2}{3}$	0
Married	$+\frac{3}{4}$	0	0
Sum	0	0	0

Verify that the sum of cross-products of the codes for each pair of variables equals zero. The first variable (MARRIED) contrasts currently married persons with those who are not currently married (divorced, widowed, and never-married). The second variable (NEVER) contrasts those who have never been married with those who have lost a spouse (divorced and widowed). Finally, the third variable (WIDOWED) contrasts widowed respondents with divorced respondents, just as in the first set of contrast codes. Compute the slopes for the three variables in the above contrast coding of marital status.

Summary

In general, nominal variables cannot be entered directly into regression equations. New variables must be created, with appropriate coding, to represent the nominal variable in an regression equation. If the nominal variable has g categories, then $g - 1$ specially coded variables must be used to represent fully the nominal variable. Dummy variable coding, effects coding, and contrast coding are three useful ways of scoring the new variables.

In dummy coding the investigator must specify one category of the nominal variable as a reference group and create a dummy variable for each of the other groups. The slope for each dummy variable equals the mean for the group coded 1 on that dummy variable minus the mean of the reference group. The consequence of this coding is that the reference group is compared to each of the other groups. Thus, it is important to pick a reference group that provides the most interesting or important theoretical comparisons.

Effects coding appears more important for exploratory studies or for other situations where it is not desirable to single out one category as a reference group. In effects coding, $g - 1$ variables are created, just as in dummy coding. Thus, there will be an effects-coded variable for each group except one. The slope for each effects-coded variable represents the mean of the group represented by that variable minus the grand mean for all of the other groups (the unweighted mean of means).

Contrast coding is the most flexible type of coding, and thus it gives the investigator the most control over the various types of group comparisons that may be investigated. In contrast coding, each variable contrasts the mean of means of one subset of groups with the mean of means of another subset of groups. Each subset, however, may consist of only one group. Just as in dummy coding and effects coding, only $g - 1$ contrast variables can be used.

Finally, it is important to remember that each method of coding leads to the same predicted scores; that is, for each method, the predicted score for each group is the mean of that group on the dependent variable. The consequence is that the squared multiple correlation is the same for each type of coding; dummy, effects, and contrast coding all explain the same proportion of variance in the dependent variable. Thus, if the objective is to determine how much

variance is explained by a nominal variable, or simply to control for the nominal variable while focusing on the effects of other variables, then one coding method is as good as another.

Reference

Blalock, Hubert M., Jr. 1979. *Social Statistics*. New York: McGraw Hill.