8

Categorical or Nominal Independent Variables

8.1 INTRODUCTION

8.1.1 Categories as a Set of Independent Variables

In this chapter, we introduce the options for treating categorical variables (nominal or qualitative scales) as independent variables in MRC. These IVs, such as religion or experimental treatments, may be represented by sets of IVs known as code variables. Each code variable represents a different aspect of the nominal variable. Taken together, the set of code variables represents the full information available in the original categories. As we will see, several different methods of selecting code variables are available, with the best method being determined by the researchers' central questions.

Our presentation in this initial chapter on categorical IVs is limited to simple regression models that do not involve interactions or nonlinear relationships. However, the central idea of this chapter—the representation of nominal variables as a set of code variables—sets the stage for the consideration of more complex regression models that include nominal IVs. Chapter 9 then considers more complex models, including curvilinear relationships in experiments, interactions between nominal variables, and interactions between nominal and continuous variables. In each case, the basic ideas presented in the present chapter provide the foundation for the interpretation of unstandardized *B*s in even the most complex regression models involving nominal variables.

8.1.2 The Representation of Categories or Nominal Scales

Nominal scales or categories are those that make qualitative distinctions among the objects they describe such as religion, treatment groups in experiments, region of country, ethnic group, occupation, diagnosis, or marital status. Each such research factor G categorizes the participants into one of g groups, where $g \geq 2$. The g groups are mutually exclusive and exhaustive: No participant is in more than one category, and all participants are categorized into one of the groups. Again, in the general case the g groups are not necessarily ordered from "lower" to "higher." To use nominal variables as IVs in MRC, it is necessary to represent them quantitatively, that is, as numbers. This problem can be summarized as "How do you score religion?"

There are several frequently used coding systems among the many possible systems used to accomplish this scoring. As we will see in this chapter, each system, *taken as a set*, produces identical results for the overall effect of the nominal variable (R, R^2, and the F test of significance of the IV). However, the regression coefficients produced for each of the variables in the set when considered simultaneously answer different questions, depending on the coding system. All coding systems use $g - 1$ different IVs (code variables) to represent the g groups, each representing *one* aspect of the distinctions among the g groups. As we will see, there are several different rationales for choosing the code variable set, because each of the alternatives will put the focus on different specific aspects of the differences among the groups. An important characteristic of alternative coding systems to be noted by the novice is that the numerical value and the interpretation of a regression coefficient for any given variable in the set may change as a function of the coding system that is chosen.

Taken as a set, the $g - 1$ code variables that comprise the coding system, however selected, represent all the information contained in the nominal scale or categories. Thus, if our nominal scale of religion is a classification into the $g = 4$ categories Catholic, Protestant, Jewish, and Other (in which we have decided to include "None" so that all respondents are categorized[1]), it will take $g - 1 = 3$ code variables $[C_1, C_2, C_3]$ to fully represent the information in this classification. One might think that it would take $g = 4$ code variables $[C_1, C_2, C_3, C_4]$ to do so, but in our MRC model, the C_4 variable would be fully redundant with (predictable from) the other three and thus C_4 provides no additional information beyond what is contained in the set $[C_1, C_2, C_3]$. This redundancy is most easily seen if we consider a nominal variable with $g = 2$ categories such as gender. Once we have a code variable to identify all females, anyone left over is a male; the distinction on sex has been completely made. A second code variable identifying males would provide no new information about the distinction on sex. Indeed, the gth code variable is not only unnecessary, but mathematically mischievous: Its redundancy renders the regression equation not (uniquely) solvable (see Section 10.5.1 on exact collinearity).

Given that religion can be represented by a set of three code variables, what are they? Among the more popular choices are dummy variables, unweighted effects, weighted effects, and planned contrasts. Each of these coding systems leads to a different interpretation of the meaning of the results for the individual code variables. Researchers should choose the coding system that provides information that most directly addresses their substantive research questions.

8.2 DUMMY-VARIABLE CODING

8.2.1 Coding the Groups

Consider the example of the four groups (Catholic, Protestant, Jewish, and Other) that comprise our nominal variable of religion. Table 8.2.1 presents alternative dummy-variable coding schemes that could be used for our numerical example. In dummy coding, one group (in Table 8.2.1B, Protestant) is designated as the reference group and is assigned a value of 0 for every code variable. The choice of the reference group is statistically but not substantively arbitrary. Hardy (1993) has suggested three practical considerations that should guide this choice. First, the reference group should serve as a useful comparison (e.g., a control group;

[1]This decision is substantive rather than statistical, possibly made because the numbers in both groups are too small to produce reliable estimates, and the expected differences from other groups too heterogeneous to investigate in either group.

TABLE 8.2.1
Illustration of Dummy-Variable Coding Systems: Religious Groups

A. Catholic as reference group.

Religion	C_1	C_2	C_3
Catholic	0	0	0
Protestant	1	0	0
Jewish	0	1	0
Other	0	0	1

C. Jewish as reference group.

Religion	C_1	C_2	C_3
Catholic	1	0	0
Protestant	0	1	0
Jewish	0	0	0
Other	0	0	1

B. Protestant as reference group (used in chapter).

Religion	C_1	C_2	C_3
Catholic	1	0	0
Protestant	0	0	0
Jewish	0	1	0
Other	0	0	1

D. Other as reference group

Religion	C_1	C_2	C_3
Catholic	1	0	0
Protestant	0	1	0
Jewish	0	0	1
Other	0	0	0

the group expected to score highest or lowest on Y; a standard treatment). Second, for clarity of interpretation of the results, the reference group should be well defined and not a "wastebasket" category (e.g., "Other" for religion). Third, the reference group should *not* have a very small sample size relative to the other groups. This consideration enhances the likelihood of replication of individual effects in future research. We chose Protestant as the reference category for the example we will use in this chapter. Protestant is a well-defined category, and Protestants represent the largest religious group in the United States population. We also chose Protestants for a purely pedagogical reason (contrary to Hardy's first recommendation). Protestants are expected to have a mean somewhere in the middle of the scale for the DV in this (fictitious) example, attitude toward abortion (ATA). This choice permits us to illustrate the interpretation of outcomes in which other groups have higher as well as lower means than our reference group.

Having chosen Protestant as our reference group, each of the other groups is given a value of 1 on the dummy-coded variable that will contrast it with the reference group in the regression analysis and a value of 0 on the other dummy-coded variables. As is illustrated in Table 8.2.1B, C_1 contrasts Catholic with Protestant, C_2 contrasts Jewish with Protestant, and C_3 contrasts Other with Protestant in the regression equation. All $g - 1$ code variables (here, 3) must be included in the regression equation to represent the overall effect of religion. Each code variable contributes 1 *df* to the overall $g - 1$ *df* that comprise the nominal variable. If some of the code variables are omitted, the interpretation of each of the effects can dramatically change (see Serlin & Levin, 1985). Indeed, we will see later in this section that the interpretation of zero-order (simple Pearson) correlations or *B*s of dummy variables with Y is strikingly different from the interpretation of regression coefficients.

The other sections in Table 8.2.1 show alternative dummy codes in which each of the categories of religion are taken as the reference group. Table 8.2.1A shows that coding system taking Catholic as the reference group, Table 8.2.1C shows the coding system taking Jewish as the reference group, and Table 8.2.1D shows the coding system taking Other as the reference group. In each case, the group whose row entries are coded [0 0 0] will be the reference group when the variables are considered as a set. The specific group being contrasted with the reference group in the regression analysis by the specific code variable (e.g., C_1) is represented by a 1.

Table 8.2.2 displays other examples of dummy coding of nominal variables. In Part A, four regions of the United States are represented by $g - 1 = 3$ dummy codes. In this example, South was chosen as the reference group. In Table 8.2.2B, three treatment conditions in a randomized experiment are represented by 2 dummy codes with Control as the reference group. In Table 8.2.2C, sex is represented by a single dummy code with male as the reference group. The coding systems presented in this chapter are completely general and can be applied to any nominal variable with 2 or more categories. Whether the nominal variable represents a natural category like religion or experimental treatment groups has no consequence for the analysis.

Table 8.2.3 presents hypothetical data for our example of the four religious groups for $n = 36$ cases. The sample sizes of these groups are Catholic ($n_1 = 9$), Protestant ($n_2 = 13$), Jewish ($n_3 = 6$), and Other ($n_4 = 8$). Throughout this chapter we will use these four groups and the dependent variable of attitude toward abortion (ATA) as our illustration. On this measure, higher scores represent more favorable attitudes. The sample sizes are deliberately unequal because this is the more general case. Where equal sample sizes simplify the interpretation of

TABLE 8.2.2
Other Illustrations of Dummy-Variable Coding

A. Dummy coding of regions of the United States ($g = 4$).

	Code variables		
Region	C_1	C_2	C_3
Northeast	1	0	0
Midwest	0	1	0
West	0	0	1
South	0	0	0

B. Dummy coding of experimental treatment groups ($g = 3$).

	Code variables	
Experimental group	C_1	C_2
Treatment 1	1	0
Treatment 2	0	1
Control	0	0

C. Dummy coding of sex ($g = 2$).

	Code variable
Sex	C_1
Female	1
Male	0

Note: The reference group for each dummy variable coding scheme is in boldface type.

CH08EX01

TABLE 8.2.3
Illustrative Data for Dummy-Variable Coding for Religion and Attitude Toward Abortion ($g = 4$)

Case no.	Group	DV	C_1	C_2	C_3
1	*C*	61	1	0	0
2	*O*	78	0	0	1
3	*P*	47	0	0	0
4	*C*	65	1	0	0
5	*C*	45	1	0	0
6	*O*	106	0	0	1
7	*P*	120	0	0	0
8	*C*	49	1	0	0
9	*O*	45	0	0	1
10	*O*	62	0	0	1
11	*C*	79	1	0	0
12	*O*	54	0	0	1
13	*P*	140	0	0	0
14	*C*	52	1	0	0
15	*P*	88	0	0	0
16	*C*	70	1	0	0
17	*C*	56	1	0	0
18	*J*	124	0	1	0
19	*O*	98	0	0	1
20	*C*	69	1	0	0
21	*P*	56	0	0	0
22	*J*	135	0	1	0
23	*P*	64	0	0	0
24	*P*	130	0	0	0
25	*J*	74	0	1	0
26	*O*	58	0	0	1
27	*P*	116	0	0	0
28	*O*	60	0	0	1
29	*J*	84	0	1	0
30	*P*	68	0	0	0
31	*P*	90	0	0	0
32	*P*	112	0	0	0
33	*J*	94	0	1	0
34	*P*	80	0	0	0
35	*J*	110	0	1	0
36	*P*	102	0	0	0

Group means: $M_C = 60.67$; $M_P = 93.31$; $M_J = 103.50$; $M_O = 70.13$.

Note: DV is attitude toward abortion. Higher scores represent more favorable attitudes. For religious group, C = Catholic; P = Protestant; J = Jewish; O = Other.

the results, this will be pointed out. Table 8.2.3 also includes the set of dummy code variables C_1, C_2, and C_3 presented in the coding system in Part B of Table 8.3.1. Thus, cases 3, 7, 13, 15, and 21 (among others) are Protestant. As members of the reference group, their C_1, C_2, C_3 scores are [0 0 0]. Similarly, cases 1, 4, and 5 (among others) are Catholic and are scored [1 0 0]; cases 18, 22, and 25 (among others) are Jewish and are scored [0 1 0]; and Cases 2, 6 and 9 (among others) are members of other religions and are scored [0 0 1].

TABLE 8.2.4
Correlations, Means, and Standard Deviations of the Illustrative Data for Dummy-Variable Coding

		r ATA	C_1	C_2	C_3	r^2_{Yi}	$t_i(df = 34)$
ATA		1.000	−.442	.355	−.225	—	
Catholic	C_1	−.442	1.000	−.258	−.309	.1954	−3.214*
Jewish	C_2	.355	−.258	1.000	−.239	.1260	0.881
Other	C_3	−.225	−.309	−.239	1.000	.0506	−2.203*
	M	81.69	.250	.167	.222		
	sd	27.88	.439	.378	.422		

$R^2_{Y.123} = .3549$; $F = 5.869^*$ ($df = 3, 32$).
$\tilde{R}^2_{Y.123} = .2945$.

Note: C_1 is the dummy code for Catholic, C_2 is the dummy code for Jewish, and C_3 is the dummy code for Other religions (see Table 8.2.1B). C_1 represents Catholic vs. non-Catholic, C_2 represents Jewish vs. non-Jewish, and C_3 represents Other vs. non-Other, if each code variable is taken separately. ATA = attitude toward abortion. $^*P < .05$.

Through the use of dummy codes, the categorical information has been rendered in quantitative form. We can now fully and meaningfully exploit the data in Table 8.2.3 through the use of MRC: Statistics on individual variables and combinations of variables can be computed, bounded by confidence limits, and tested for statistical significance to provide projections to the population. The MRC results for our illustrative example are given in Tables 8.2.4 and 8.2.5.

TABLE 8.2.5
Analysis of Illustrative Data: Attitude Toward Abortion

A. Dummy-variable coding: partial and semipartial correlations and regression coefficients.

C_i	pr_i	pr_i^2	sr_i	sr_i^2	β_i	B_i	SE_{B_i}	t_i
C_1	−.494	.2441	−.456	.2083	−.5141	−32.64	10.16	−3.214*
C_2	.154	.0237	.125	.0157	.1382	10.19	11.56	0.882
C_3	−.363	.1317	−.313	.0978	−.3506	−23.18	10.52	−2.203*

B. Predicted values in groups.

$$\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$$
$$= -32.64C_1 + 10.19C_2 - 23.18C_3 + 93.31.$$

Catholic:	$\hat{Y}_1 = -32.64(1) + 10.19(0) - 23.18(0) + 93.31 = 60.67 = \bar{Y}_1$
Protestant:	$\hat{Y}_2 = -32.64(0) + 10.19(0) - 23.18(0) + 93.31 = 93.31 = \bar{Y}_2$
Jewish:	$\hat{Y}_3 = -32.64(0) + 10.19(1) - 23.18(0) + 93.31 = 103.50 = \bar{Y}_3$
Other:	$\hat{Y}_4 = -32.64(0) + 10.19(0) - 23.18(1) + 93.31 = 70.13 = \bar{Y}_4$

$$sd^2_{Y-\hat{Y}} = sd^2_Y(1 - R^2)\frac{n}{(n-k-1)} = 27.49^2(1 - .3549)\frac{36}{36-3-1} = 548.41.$$

Note: $df = 32$. $^*P < .05$.

8.2.2 Pearson Correlations of Dummy Variables With *Y*

Each dummy-code variable C_i is a dichotomy that expresses one meaningful aspect of group membership. For example, when considered alone, C_1 represents Catholic versus non-Catholic. When we calculate the Pearson correlation between C_1 and ATA, we get the point-biserial correlation (see Section 2.3.3) between Catholic versus non-Catholic and ATA in this sample. These r_{Yi} values are given in the first column of Table 8.2.4. Thus, Catholic versus non-Catholic status in this sample correlates $-.442$ with ATA or equivalently accounts for $(-.442)^2 = .1954$ of the variance in ATA (column labeled r_{Yi}^2). Jewish versus non-Jewish status correlates .355 and accounts for $(.355)^2 = .1260$ of the ATA variance. Other versus non-Other (Catholic, Protestant, and Jewish combined) correlates $-.225$ with ATA and accounts for $(-.225)^2 = .0506$ of the ATA variance.

Table 8.2.4 displays the correlations for the code variables corresponding to Catholic, Jewish, and Other. However, no correlations or proportions of variance accounted for are given for the reference group, here Protestant. How do we get these values? There are two ways. First, we can rerun the analysis using any of the other dummy variable coding systems shown in Table 8.2.1 in which Protestant is *not* the reference group. For example, if we use the dummy codes in Part A of the table, we find that Protestant versus non-Protestant correlates .318 with ATA and the proportion of variance accounted for is $(.318)^2 = .1011$. The r_{Yi} values for Jewish and Other in this second analysis will be identical to those reported in Table 8.2.4. Alternatively, we can calculate r_{Yi} for the reference group from our knowledge of the proportion of subjects in the total sample in each group and the correlations of each of the dummy code variables with the dependent variable. When the proportion of subjects in each group is *not* equal,

$$r_{Yr} = \frac{\sum r_{Yi}\sqrt{P_i(1-P_i)}}{\sqrt{P_r(1-P_r)}}. \tag{8.2.1}$$

In this equation, r_{Yi} represents the correlations of each of the $g-1$ dummy codes with the dependent variable (ATA), r_{Yr} is the correlation of the reference group with the dependent variable, P_i is the proportion of the total sample in the group coded 1 on each dummy variable, and P_r is the proportion of the sample in the reference group. Applying this formula to our present example,

$$r_{Yr} = -\frac{-0.442\sqrt{(.25)(.75)} + .355\sqrt{(.167)(.833)} + (-.225)\sqrt{(.222)(.778)}}{\sqrt{(.361)(.639)}}$$
$$= .318.$$

When n_i is equal in each of the groups, Eq. (8.2.1) simplifies to $r_{Yr} = -\Sigma r_{Yi}$.

In interpreting the correlations, the sign of r_{Yi} indicates the direction of the relationship. If the group coded 1 has a higher mean than the mean of the other groups combined, then the sign is positive. For example, C_2 codes Jewish versus non-Jewish students. Since Jewish students had a higher mean ATA than non-Jewish students, the sign of r_{Y2} was positive. If the group coded 1 has a lower mean than the mean of the other groups combined, then the sign is negative (e.g., r_{Y1} for Catholic). The proportion of variance in Y accounted for is as described in Section 2.6, except that the source of the variance is group membership (for example, Catholic versus non-Catholic) rather than a continuous IV.

The interpretation of r_{Yi} and r_{Yi}^2 also requires careful attention to how cases were sampled. The magnitude of r_{Yi} will depend in part on the proportion of the sample that is composed of members of the group that is coded 1. We learned in Chapter 2 that *r*s increase directly with the variability of the variables being correlated. As noted, for dichotomies in which each case has a score of 0 or 1, the sample *sd* is $\sqrt{(P)(1-P)}$. Thus, the *sd* of a dummy variable depends solely on its proportion in the total sample. This value reaches its maximum when $P_i = .50$ for the group and becomes smaller as P_i either increases toward 1.0 or decreases toward 0.0. Because r_{Yi} varies with *sd*, the magnitude of a correlation with a dummy variable will change with the relative size of the group coded 1 in the total sample, reaching its maximum at $P_i = .50$. Therefore, the interpretation of any given r_{Yi} (or r^2_{Yi}) depends on the meaning of P_i in the sample.

To illustrate this point, let us revisit our example of religion and attitude toward abortion. If the 36 cases were randomly sampled from the population of students at a Midwestern university, then the r_{Yi} is a reasonable estimate of ρ_{Yi} in this population. This occurs because the proportion of each group in the sample reflects within sampling error the proportion of each group in this population. For the Catholic group (G_3), $P_i = 9/36 = .25$, $r_{Yi} = -.258$, and $r^2_{Yi} = .0667$ in this Midwestern university sample. On the other hand, other university populations will have different proportions of Catholic students. If the random sample were now taken at another university in which the proportion of Catholic students were closer to .50 (i.e., $.25 < P_i < .75$), then r^2_{Yi} would be larger, all other things being equal.

Another circumstance resulting in different r_{Yi}s is a sampling plan in which equal numbers of Catholics, Protestants, Jews, and Others are sampled from the population and their attitude toward abortion scores are measured. The equal *P*s (.25) in the resulting data do not reflect the actual proportions of each group in the population. The resulting data will yield different r_{Yi}s from those obtained using random sampling. For our hypothetical Midwestern university population, we would expect no change in the correlation for Catholics (original $P_i = .250$), a smaller correlation for Protestants (original $P_i = .361$), a larger correlation for Jews (original $P_i = .167$), and a slightly larger correlation for Other (original $P_i = .222$) under this equal number sampling plan. In comparing the original correlations referenced on random sampling with the correlations referenced on equal sampling, the correlation for any group whose P_i was originally closer to .50 would be expected to decrease, whereas the correlation for any group whose P_i was originally further from .50 would be expected to increase. Once again, these changes in r_{Yi} occur with changes in the proportion of the group in the sample because the associated *sd*s of the dummy variables correspondingly increase or decrease.

These examples illustrate the importance of carefully considering the population and the sampling method when interpreting correlations and squared correlations involving dummy variables. Correlations computed from random or other types of representative samples produce good estimates of the value of the correlation in that particular population. Correlations computed from samples containing an equal number of cases in each group permit generalization to a hypothetical population in which each group occurs in an equal proportion.[2] Such sampling plans are useful for two different reasons. First, there may be an interest in within-group effects in groups having a low proportion in the population. For example, researchers studying ethnic minority groups often oversample such groups in order to have sufficient numbers to study within group relationships with reasonable statistical power. Second, it may be that the theory generalizes not to a particular population but rather to abstract properties of subjects within a population, and subjects are therefore either selected or manipulated to allow a comparison of these properties with maximal statistical power (e.g., Pitts & West, 2001). The interpretation of the proportion of variance accounted for is then appropriate only for the hypothetical population with those relative group frequencies. In interpreting correlations,

---

[2]When the proportions of each group in the population are known and groups have been oversampled or undersampled relative to their respective population proportions, weighted regression techniques can be used to provide good estimates of the population values (Winship & Radbill, 1994). Weighted regression techniques were introduced in Section 4.5.4 in another context.

considerable care must be taken because of the effects of the population and the sampling plan on the obtained values.

Confidence Intervals and Significance Tests for Bivariate r

In Section 2.8.2, we considered the confidence interval for r. The calculation of a confidence interval (CI) for a correlation between a dichotomous (e.g., dummy-coded) and a continuous variable follows the identical procedure.

To briefly review, the upper and lower limits of the CI do not fall at equal distances from r. We use the Fisher z' transformation of r to bypass this problem. Values of the r to z' transform are given in Appendix Table B. For our Midwestern University example, $-.44$ (rounded down) is the correlation between the dummy code corresponding to Catholic and favorable attitude toward abortion. This r converts to a z' of .472. z' has an approximately normal distribution and a standard error as applied to Eq. (2.8.3) of

$$SE_{z'} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{36-3}} = .174.$$

For the 95% confidence interval, we then find the upper and lower limits for z'. Recall that for the normal distribution, 1.96 is the multiplier for the standard error to construct the confidence limits (see Appendix Table C). Thus, the confidence interval is $.472 \pm (1.96)(.174)$. This gives the 95% limits for z' as .131 and .813. We then convert these values back to r, again using Appendix Table B, and restore the original negative sign for the correlation. This results in a 95% confidence interval for r from $-.13$ to $-.67$.

For researchers who prefer significance tests, we can alternatively test the obtained sample correlation against a population value of $\rho = 0$ as discussed in Section 2.8.3. We use Eq. (2.8.10) reproduced below:

$$(2.8.10) \qquad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{with} \quad df = n-2.$$

For our Midwestern University example, $r = -.442$ for the correlation between Catholic and favorable attitude towards abortion referenced on our 36 cases. Substituting into the formula, we find

$$t = \frac{-.442\sqrt{36-2}}{\sqrt{1-(-.442)^2}} = -2.87.$$

For $df = n - 2 = 34$, this value of t (ignoring the negative sign) easily exceeds the $p < .05$ significance criterion of 2.032. We therefore reject H_0 and conclude there is a negative correlation in the Midwestern University population between being Catholic and having a favorable attitude toward abortion.

This is exactly the same t value we would obtain if we used the familiar t test for the means between two independent groups,

$$(8.2.2) \qquad t = \frac{M_{Y_1} - M_{Y_0}}{\sqrt{SD^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

to test the difference between the means of ATA for Catholic and non-Catholic groups. The two significance tests represented by Eqs. (2.8.10) and (8.2.2) are algebraically identical when two groups are being compared.[3] The chief advantage of the use of the t test of r rather the t test of the

[3]Eq. (2.3.7), $r_{pb} = (M_{Y_1} - M_{Y_0})\sqrt{PQ}/sd_Y$, reminds us that the point biserial correlation can be expressed directly in terms of the mean difference.

mean difference between the two groups is that it gives us the proportion of variance accounted for (r^2) directly. The correlational context also helps remind us of the critical importance of the sampling plan in interpreting our results. The population and sampling issues discussed in the previous section are equally important in interpreting group differences in observational studies and in experiments, but these issues are often ignored.

8.2.3 Correlations Among Dummy-Coded Variables

The definition of nominal variables requires that each case be classified into one and only one group. For such mutually exclusive categories, the correlation between pairs of categories must be negative. If a person who is Protestant is necessarily non-Catholic, and if Catholic, necessarily non-Protestant. However, this correlation is never -1.00 when there are more than two categories: a person who is non-Protestant, may be either Catholic or non-Catholic, because there are other groups (Jewish, Other) as well. Because these correlations are between dichotomies (e.g., Catholic vs. non-Catholic; Jewish vs. non-Jewish), they are phi coefficients (see Section 2.3.4) and can be computed by hand or computer using the usual Pearson product-moment correlation formula. The correlation between two dummy codes (C_i, C_j) can be calculated using the following formula:

$$r_{ij} = -\sqrt{\frac{n_i n_j}{(n-n_i)(n-n_j)}} = -\sqrt{\frac{P_i P_j}{(1-P_i)(1-P_j)}}.$$

For example, the correlation between the dummy codes representing Catholic and Jewish in our running example is

$$r_{12} = -\sqrt{\frac{(.250)(.167)}{(1-.250)(1-.167)}} = -.259.$$

Thus, we conclude that dummy codes, which represent separate aspects of the nominal variable G, will necessarily be partly redundant (correlated with each other). This conclusion has two important implications that will be developed more fully in Sections 8.2.5 and 8.2.6. First, the unstandardized regression coefficients will usually be the focus of interpretation. They compare the unique effect of the group of interest (e.g., Catholic) with the effects of other groups that comprise the nominal variable G held constant. Second, we cannot find the proportion of variance in Y due to G simply by summing the separate r^2_{Yi} for the $g-1$ dummy variables. We must necessarily use the squared multiple correlation which takes into account the redundancy (correlation) between the set of dummy codes that comprise G. To see this, readers may compare the sum of the values in the r^2_{Yi} column of Table 8.2.4 with the value of $R^2_{Y.123}$. The value of $R^2_{Y.123}$ is less than the sum of the r^2_{Yi} values because of the partial redundancy among the dummy codes.

8.2.4 Multiple Correlation of the Dummy-Variable Set With Y

Returning to our running example of the effects of religion on ATA, we can write a regression equation that specifies the influence of the set of dummy-coded variables in the usual way,

$$\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0. \tag{8.2.3}$$

When we run an MRC analysis using this equation and the religion and attitude toward abortion data in Table 8.2.3, we find that $R^2_{Y.123} = .3549$. Since the three dummy codes C_1, C_2, and

C_3 as a set comprise the group (here, religion), this also means that $R^2_{Y.G} = .3549$. As noted earlier, any of the dummy-variable coding systems shown in Table 8.2.1 will yield this same value for $R^2_{Y.G}$. We thus can state that 35.5% of the variance in ATA scores is associated with religion in this sample or that the R of ATA and religion is .596. Note that R^2 depends on the distribution of the n_i of the four groups; a change in the relative sizes holding the M_{Y_i}s constant would in general change R^2. This dependence on the P_i is characteristic of R, as it is of any kind of correlation, and must be kept in mind in interpreting the results.

We can construct a confidence interval for R^2 as shown in Section 3.6.2. In the present example, substituting into Eg. 3.6.2, we find

$$SE^2_{R^2} = \frac{4(.3549)(1 - .3549)(36 - 3 - 1)^2}{(36^2 - 1)(36 + 3)} = .0186.$$

The square root of this variance, SE_{R^2}, is the standard error of R^2, which is used in the calculation of the confidence interval, $SE_{R^2} = \sqrt{.0186} = .136$. From Appendix Table A, the critical value of t for $df = 32$ and $\alpha = .05$ is 2.037. The approximate 95% confidence interval is calculated as $R^2 \pm t(SE_{R^2}) = .3549 \pm (2.037)(.136)$. Thus, the 95% confidence interval for R^2 ranges from .0773 to .6325. This confidence interval is only approximate and should be interpreted cautiously given the relatively small n (see Olkin & Finn, 1995).

Alternatively, for researchers who prefer significance tests, we can use Eq. (3.6.5) (reproduced here) to test the significance of R^2 as compared to the null (nil) hypothesis:

(3.6.5) $$F = \frac{R^2(n - k - 1)}{(1 - R^2)k} = \frac{R^2(n - g)}{(1 - R^2)(g - 1)}.$$

Substituting in our present values, we find

$$F = \frac{.3549(36 - 4)}{(1 - .3549)(3)} = 5.869.$$

For $df = 3,\ 32$, the F required for significance at $\alpha = .05$ is 2.90 (see Appendix Table D.2), hence our obtained F is statistically significant. We reject the null hypothesis that religion accounts for no variance in ATA scores in the population that was sampled.

We may also wish to report the adjusted (or shrunken) R^2. We saw in Chapter 3 that R^2 provides an accurate value for the sample but overestimates the proportion of variance accounted for in the population. For a better estimate of the Y variance accounted for by religion in the population, we use Eq. (3.7.4) to estimate the shrunken R^2 as

$$\tilde{R}^2 = 1 - (1 - .3549)\,\frac{35}{32} = .2945.$$

Our best estimate of the proportion of ATA variance accounted for by religion in the population is 29.4%. Here again it is important to keep in mind how the sampling was carried out, because the population to which we may generalize is the one implicit in the sampling procedure.

8.2.5 Regression Coefficients for Dummy Variables

Let us consider the meaning of each of the unstandardized regression coefficients in the equation predicting Y in more depth. That equation (Eq. 8.2.3) as we have seen will be $\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$. In our running example, $\hat{Y}$ is the predicted value of ATA, B_0 is the intercept, B_1 is unstandardized regression coefficient for the first dummy code, B_2 is the unstandardized regression coefficient for the second dummy code, and B_3 is the unstandardized regression coefficient for the third dummy code. If we use the dummy coding system presented in Table 8.2.1B in which Protestant is the reference group, then C_1 corresponds to Catholic, C_2 corresponds to Jewish, and C_3 corresponds to Other. Substituting the values of the dummy codes corresponding to each religion into the equation, we find

$$\begin{aligned}
\text{Catholic:}\quad & \hat{Y} = B_1(1) + B_2(0) + B_3(0) + B_0 = B_1 + B_0 = M_{\text{Catholic}};\\
\text{Protestant:}\quad & \hat{Y} = B_1(0) + B_2(0) + B_3(0) + B_0 = B_0 = M_{\text{Protestant}} \text{ (reference group)};\\
\text{Jewish:}\quad & \hat{Y} = B_1(0) + B_2(0) + B_3(0) + B_0 = B_2 + B_0 = M_{\text{Jewish}};\\
\text{Other:}\quad & \hat{Y} = B_1(0) + B_2(0) + B_3(1) + B_0 = B_3 + B_0 = M_{\text{Other}}.
\end{aligned}$$

Thus, in Eq. (8.2.3) when $C_1 = 0$, $C_2 = 0$, and $C_3 = 0$ for the reference group (Protestant), the predicted value of Y equals B_0, the regression intercept, which also equals the mean of the reference group. The same value $\hat{Y} = B_0$ is predicted for all subjects in the reference group. A 1-unit change on C_1 (i.e., a change from $C_1 = 0$ to $C_1 = 1$) represents the difference in the value of the Group 1 (Catholic) mean from the reference group (Protestant) mean on the DV. A 1-unit change on C_2 represents the difference in the value of the Group 2 (Jewish) mean from the reference group mean on the DV. Finally, a 1-unit change in the value on C_3 represents the difference in the value of the Group 3 (Other) mean from the reference group mean on the DV. Thus, each regression coefficient and its significance test is a comparison of the mean of one of the groups with the mean of the reference group, here Protestant.

To illustrate these relationships for our numerical example,

$$\hat{Y} = -32.64C_1 + 10.19C_2 - 23.18C_3 + 93.31.$$

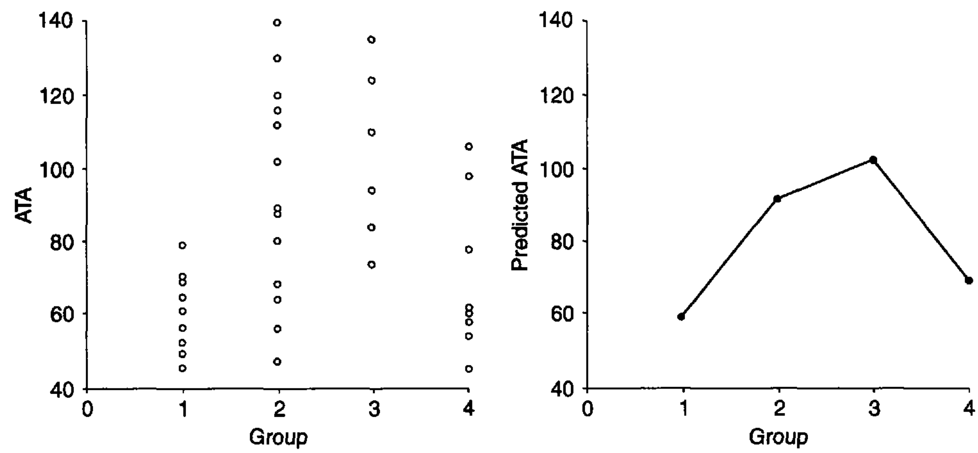
Table 8.2.5 shows the results of substituting the dummy codes corresponding to each religious group into this equation. For each group, $\hat{Y}$ is identical to the mean for the group shown in Table 8.2.3.

Graphical Depiction

Figure 8.2.1 depicts the results of the regression analysis graphically. Figure 8.2.1(A) shows the scatterplot of the raw data for the four religious groups, which have been numbered 1 = Catholic, 2 = Protestant, 3 = Jewish, and 4 = Other on the x axis. Over these values the plot displays the distribution of scores on ATA corresponding to each religious group.[4] Figure 8.2.1(B) shows the predicted values (group means) on ATA for each of the religious groups, again plotted over the numbers identifying each religious group. Figure 8.2.1(C) shows the residuals $Y_i - \hat{Y}_i$ from predicting each case's score on ATA. The values on the x axis are the predicted values, $\hat{Y}$, for each group. The residuals for each group are now displayed above their respective group means. Thus, the residuals for Catholic are displayed above 60.7 (the mean for the Catholic group; see Table 8.2.3), the residuals for Other are displayed above 70.1, the residuals for Protestant are displayed above 93.3, and the residuals for Jewish are displayed above 103.5. This plot allows us to examine whether the variability around the predicted values differs greatly across the groups (heteroscedasticity). Finally, Fig. 8.2.1(D) shows a q-q plot of

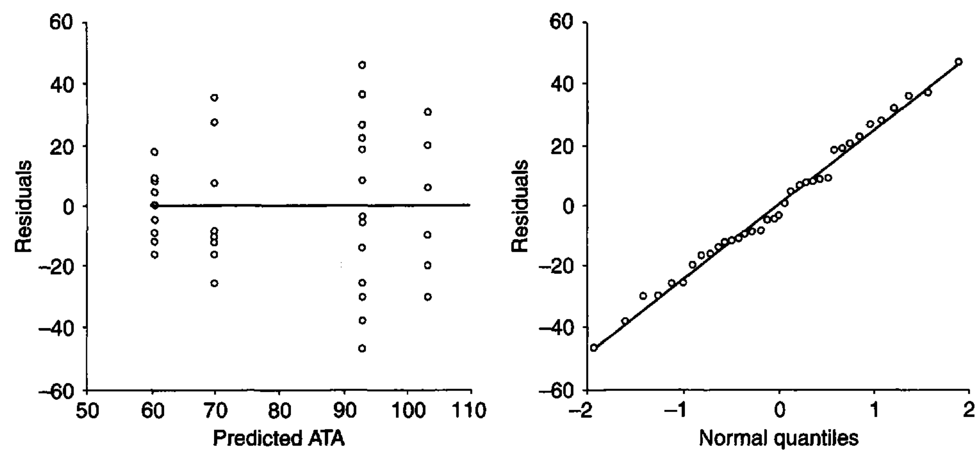
[4]In the present example, the simple scatterplot clearly presents the data because each person in a group has a different value on ATA. In larger data sets more than one person in the group (often many) may have the same value on the dependent variable. Each of these points will then be plotted on top of each other (overplotting) and cannot be distinguished. The graphical option of "jittering" points presented in Section 4.2.2 will help. Other graphs that allow for the comparisons of the distributions across the groups avoid these problems (see Cleveland, 1994).

(A) Scatterplot of raw data. (B) Scatterplot of predicted ATA vs. group.



Note: In Parts (A) and (B), the values of religion are as follows: 1 = Catholic, 2 = Protestant, 3 = Jewish, 4 = Other. In Part (B) the predicted ATA values for each group are plotted over the group number. These predicted values equal the means for each group.

(C) Scatterplot of residuals vs. fit values. (D) q-q plot of residuals against normal distribution.



Note: In Part (C) the horizontal line indicates where the residuals = 0. In Part (D) the q-q plot of the residuals closely follows a straight line, indicating the residuals are normally distributed.

FIGURE 8.2.1 Results of regression analysis for religious groups and ATA.

the residuals against a normal distribution (see Section 4.4.6). The q-q plot indicates that the residuals approximate a straight line, so the assumption of normality of the residuals is met.

Confidence Intervals and Significance Tests for Regression Coefficients

Once again, we can directly use the methods for constructing confidence intervals and for performing significance tests for the unstandardized regression coefficients that were presented in Chapter 3. Each of the unstandardized regression coefficients has a t distribution with

$df = n - k - 1$, where k is the number of code variables ($= g - 1$). For our running example of religion, $df = 36 - 3 - 1 = 32$. Thus, to construct a confidence interval, we simply take $B_i \pm tSE_{B_i}$. In general, SE_{B_i} will differ for each dummy variable's regression coefficient.

To illustrate we construct 95% confidence intervals for each of the regression coefficients in our running example. For C_1, the dummy code for Catholic, $B_1 = -32.64$ and the corresponding SE is 10.15. From Appendix Table A, the critical value of t for $df = 32$ and $\alpha = .05$ is 2.037. Then the confidence interval (CI) is $-32.64 \pm (2.037)(10.15)$ which ranges from -53.31 to -11.96. The CI for each B is

$$
\begin{aligned}
B_1:&\quad -32.64 \pm (2.037)(10.15) = -53.31 \text{ to } -11.96;\\
B_2:&\quad 10.19 \pm (2.037)(11.56) = -13.36 \text{ to } 33.74;\\
B_3:&\quad -23.18 \pm (2.037)(10.52) = -44.61 \text{ to } -1.75;\\
B_0:&\quad 93.31 \pm (2.037)(6.50) = 80.07 \text{ to } 106.55.
\end{aligned}
$$

indicating that we are 95% confident that the population difference between Protestants and Catholics on ATA is between -12 and -53, the difference between Protestants and Jews is between -13 and 34, the difference between Protestants and Others who are neither Catholics nor Jews is between -2 and -45, and the population mean for the reference group falls between 80 and 107. Of course, these estimates hold only for the population represented by the study sample (presumably students at a Midwestern University).

The significance tests of the null hypothesis that unstandardized regression coefficient $B_i = 0$ is similarly straightforward, t equaling the coefficient divided by its SE. In our running example, the null (nil) hypothesis significance test of the B_1 coefficient for Catholic is $t = -32.64/10.15 = -3.21$, which exceeds the magnitude of the critical $t = 2.037$ for $\alpha = .05$ and $df = 32$, leading us to the conclusion that the mean ATA is higher for Protestants than for Catholics. Researchers having a null hypothesis that the value of B_i is equal to some specific value c (e.g., H_0: $B_1 = -10$ in the population) can test this hypothesis by $t = (B_1 - c)/SE_{B_1}$ with $df = n - k - 1$. In this equation, c is the constant representing the new "null" hypothesis. In this case $[-32.64 - (-10)]/10.15 = 2.23$, which again exceeds the critical $t = 2.037$. Thus, the population mean for Protestants exceeds the population mean for Catholics by more than 10 points (a statement that might also have been based on the CI for this difference).

Finally, researchers may sometimes be interested in comparisons of two groups in which neither of these groups is the reference group. For example, the researcher may be interested in comparing the mean of the Catholic with the mean of the Jewish group. The easiest way to accomplish this is to rerun the analysis after the data have been recoded using another dummy coding system in which one of the groups being compared is the reference group. Rerunning the analysis with the dummy coding system from Table 8.2.1A, in which Catholic is the reference group, will yield $B_2 = 42.83$, indicating that the mean of ATA in the Jewish group is 42.83 larger than in the reference Catholic group. The analysis also produces $SE_{B_2} = 12.34$ and $t = 3.47$, $df = 32$, $p < .05$. Alternatively, this test may be performed by hand to compare B_i to B_j by

$$t = \frac{B_i - B_j}{\sqrt{SE_{Y-\hat{Y}}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, \tag{8.2.4}$$

where $SE_{Y-\hat{Y}}$ is the standard error of estimate available from the computer printout. Recall from Sections 2.6 and 3.6 that the standard error of estimate represents the standard deviation of the residuals about the regression line. Its square can also be expressed as

$$sd^2_{Y-\hat{Y}} = \frac{\sum (Y - \hat{Y})^2}{n - k - 1} = sd^2_Y(1 - R^2)\,\frac{n}{n - k - 1}$$

Applying this equation to the comparison of the Jewish (B_2) with the Catholic (B_1) groups, we find

$$t = \frac{10.19 - (-32.64)}{\sqrt{(23.42)^2\left(\frac{1}{6}+\frac{1}{9}\right)}} = \frac{42.83}{12.34} = 3.47$$

which is significant at $p < .05$. Any other possible comparisons of group means not involving the reference group may also be carried out using these procedures.

At the same time, caution must be exercised in conducting such comparisons, particularly when there is not a strong a priori prediction that two specific groups will differ. Although it is possible to compare all pairs of group means, such procedures are not advisable because they increase the level of α for the study beyond the stated level, typically $\alpha = .05$. Kirk (1995, Chapter 4) and Toothaker (1991) present a full discussion of procedures for controlling the Type I error rate when multiple groups are compared.

Standardized Regression Coefficients

The standardized regression coefficient (β_i) is far less useful for nominal than for continuous variables. Unlike continuous variables, the variability of a dichotomy cannot be changed without also changing its mean (that is, the proportion of the sample coded 1). In the previous section we saw that unstandardized regression coefficients for dummy variables represent differences between means. These differences do not depend on the relative sizes of the groups. In general, the standardized β_i for coded nominal scales will vary with changes in the relative n_is, decreasing their general usefulness. When βs are reported, they must always be carefully interpreted in light of the population and sampling procedures, which will affect their magnitude. We have presented the standardized β_is in Table 8.2.5 for the sake of completeness.

8.2.6 Partial and Semipartial Correlations for Dummy Variables

Partial Correlation

The partial correlation is the correlation of X_i with that part of Y that is independent of the other IVs in the equation. In the specific context of dummy variables, holding the other IVs constant means retaining *only* the distinction between the *i*th group and the reference group. Concretely, pr_1 ($= -.494$; see Table 8.2.5) in our example is the correlation between ATA and Catholic versus non-Catholic, holding Jewish versus non-Jewish and Other versus non-Other constant. Consequently, $pr_1 = -.494$ is an expression in correlational terms of the difference between the Catholic group and Protestant group in ATA scores. Similarly, from Table 8.2.5, $pr_2 = .154$ relates ATA to Jewish versus Protestant (holding constant Catholic and Other) and $pr_3 = -.363$ relates ATA to Other versus Protestant (holding constant Catholic and Jewish). Otherwise stated, the pr_i can be viewed as a representation of B_i in a correlational rather than raw score metric. As with other measures of correlation, the interpretation of a given pr_i must take into account the population and sampling plan.

Semipartial Correlation

Recall that the squared semipartial correlation sr_i^2 is the amount by which $R^2_{Y.123...k}$ would be reduced if X_i were omitted from the IVs. That is, $sr_i^2 = R^2_{Y.123...k} - R^2_{Y.123...(i)...k}$. Here, the ($i$) in the subscript symbolizes the omission of X_i. With dummy-variable coding, the omission of X_i is equivalent to collapsing group i in with the reference group. Consider what happens in our example if C_1 is omitted. Both Catholic and Protestant are coded $C_2 = 0$, $C_3 = 0$ and are therefore not distinguished. The result is that we have reduced our four original religious groups to three: Catholic/Protestant (combined), Jewish, and Other. We see in Table 8.2.5 that $sr_1^2 = .2083$. This means that the loss of the Catholic-Protestant distinction would result in the loss of 20.8% of the ATA variance accounted for by religion. Equivalently, our R^2 would drop from .3549 to .1468. Thus, sr_i^2 in dummy-variable regression provides a measure in terms of the proportion of total Y variance of the importance of distinguishing group i from the reference group. Thus, the Jewish-Protestant distinction (sr_2^2) accounts for only 1.6%, whereas the Other-Protestant distinction (sr_3^2) accounts for 9.8% of the ATA variance. We note again that these values, as all correlations involving dummy variables, are dependent on the proportions of the cases in each group.

In a previous section, we presented a significance test for the null (nil) hypothesis that $B_i = 0$ in the population. As noted in Chapter 3, t tests of partial coefficients including B_i, standardized β_i, pr_i, and sr_i for any given IV, including a dummy variable C_i will yield identical results. As we saw in Section 3.6, these equivalent t tests of the null (nil) hypothesis for partial relationships can be written in several different ways. One simple one in terms of the semipartial correlation and R^2 is given below:

$$t = sr_i \sqrt{\frac{n-k-1}{1-R^2}} \quad \text{with} \quad df = (n-k-1).$$

Applying this equation to our example to test sr_1, we find

$$t = -.456 \sqrt{\frac{36-3-1}{1-.3549}} = 3.214 \quad \text{with} \quad df = 32.$$

The results of this test are identical to the test of the unstandardized regression coefficient reported in Table 8.2.5. This value of $t = 3.214$ exceeds the required value of $t = 2.037$ for $\alpha = .05$, so we can reject the null hypothesis. The result of this test may be interpreted equivalently in terms of B_i, standardized β_i, pr_i, or sr_i^2. Interpreting the test for B_1, it shows that Catholics have a lower mean ATA score than Protestants (reference group). In terms of the partial correlation pr_1, if we consider only Catholics and Protestants in the population, there is a negative point-biserial correlation between this religious dichotomy and ATA. Finally, in terms of the squared semipartial correlation, sr_1^2, the test shows that if we dropped the distinction between Catholics and Protestants, R^2 would drop from its value obtained when religion is categorized into four groups.

8.2.7 Dummy-Variable Multiple Regression/Correlation and One-Way Analysis of Variance

Readers familiar with analysis of variance (ANOVA) may wonder about its relationship to MRC. Both are applications of the same general linear model, and when the independent variables are nominal scales they are identical. This equivalence has been obscured by differences in terminology between MRC and ANOVA, and divergent traditions that have linked observational designs with MRC and experimental designs with ANOVA (see Chapter 1; also Aiken & West, 1991, Chapter 9; Tatsuoka, 1975).

Viewed from the perspective of ANOVA, the problem addressed in this chapter is one way analysis of variance. We have considered g levels of a factor, with n_i observations on Y in each group. For the illustrative data of Table 8.2.3 we would assemble the Y values into the g-designated groups and proceed to find three sums of squares: total (SS_{TOTAL}), between groups (SS_{BG}), and within groups (SS_{WG}). These sums of squares are defined as follows:

$$SS_{TOTAL} = \sum (Y_j - M_G)^2 \qquad df_{TOTAL} = n - 1;$$
$$SS_{BG} = \sum n_i(M_i - M_G)^2 \qquad df_{BG} = g - 1;$$
$$SS_{WG} = SS_{TOTAL} - SS_{BG} \qquad df_{WG} = n - g.$$

In these equations Y_j is the score on Y for subject j, M_i is the mean of the ith group, n_i is the number of subjects in the ith group, M_G is the grand mean (the unweighted mean of the group means), g is the number of groups, and n is the total number of subjects in the entire sample. As is shown in Table 8.2.6 for our running example of attitude toward abortion, each SS is divided by its corresponding df, yielding three mean square (MS) values. The mean square between groups (MS_{BG}) is then divided by the mean square within groups (MS_{WG}) to yield an F statistic. This F statistic tests the null (nil) hypothesis that there are no differences among the means of the groups on Y in the population represented by the sample.

When we examine Table 8.2.6, we note that the F from the ANOVA is *identical* to the F computed earlier as a test of R^2 (see Table 8.2.4) using these data. This can be understood conceptually in that the null hypothesis of the ANOVA, equality of the g population means, is mathematically equivalent to the null hypothesis of the MRC analysis, which is that no variance in Y is accounted for by group membership. Clearly, if the population means are all equal, Y variance is not reduced by assigning to the members of the population their respective identical group means. These group means are necessarily also equal to the grand mean of the combined populations. Each null hypothesis implies the other; they differ only verbally.

The two F ratios are, in fact, algebraically identical:

$$SS_{BG} = R^2 SS_{TOTAL};$$
$$SS_{WG} = (1 - R^2) SS_{TOTAL}.$$

TABLE 8.2.6
Analysis of Variance of Attitude Toward Abortion Data

A. Analysis of variance summary table.

Source	SS	df	MS	F
Total	27,205.64	35	—	
Between groups	9,656.49	3	3,218.83	5.869
Within groups	17,549.15	32	548.41	

B. Cell means for attitude toward abortion.

	G_1	G_2	G_3	G_4
M_{Y_i}	60.67	93.31	103.50	70.13
n_i	9	13	6	8

Note: G_1 is Catholic; G_2 is Protestant; G_3 is Jewish; G_4 is Other.

C. Calculations.

$$F = \frac{SS_{BG}/(g-1)}{SS_{WG}/(n-g)} = \frac{MS_{BG}}{MS_{WG}} = \frac{R^2/(g-1)}{(1-R^2)(n-g)} = 5.869 \qquad df = 3,\ 32.$$
$$\eta^2 = \frac{SS_{BG}}{SS_{TOTAL}} = \frac{9,656.49}{27,205.64} = .3549 = R^2.$$

Substituting these in the ANOVA formula for F, we find

$$F = \frac{SS_{\mathrm{BG}}/(g-1)}{SS_{\mathrm{WG}}/(n-g)} = \frac{R^2 SS_{\mathrm{TOTAL}}/(g-1)}{(1-R^2)SS_{\mathrm{TOTAL}}/(n-g)} = \frac{R^2/(g-1)}{(1-R^2)/(n-g)}.$$

The final value is the F ratio for the significance test of R^2 in MRC.

Many modern ANOVA texts also present the formula for the proportion of variance in Y accounted for by the G factor. This statistic is known as η^2 (eta squared) and is written as follows:

(8.2.5)
$$\eta^2 = \frac{SS_{\mathrm{BG}}}{SS_{\mathrm{TOTAL}}}.$$

Note in Table 8.2.6 that the application of this formula to our running example gives $\eta^2 = .354944$, the same value as found for $R^2_{Y.123}$. In general, $\eta^2 = R^2_{Y...123...(g-1)}$.

Further, just as the shrunken (or adjusted) R^2 of Eq. (3.5.5) yields an improved estimate of the proportion of variance of Y accounted for in the population, the same improved estimate in ANOVA is known as ϵ^2 (epsilon squared), and it is readily proved that $\epsilon^2 = \tilde{R}^2_{Y...123...(g-1)}$. Thus, we see that the yield of an MRC analysis of coded nominal group membership includes the information that a one-way ANOVA yields. In addition, we also directly obtain various useful correlational measures such as simple, partial, semipartial, and multiple correlations. These measures are typically not reported in standard ANOVA statistical packages. In subsequent sections of this chapter we will see that other coding systems bring out still other identities between ANOVA and MRC. Later in this chapter we will see how analysis of covariance can be duplicated and extended by means of MRC. In Chapter 9, we will consider more complex fixed effects ANOVA models such as factorial ANOVA.

8.2.8 A Cautionary Note: Dummy-Variable-Like Coding Systems

Researchers sometimes use other dummy-variable-like coding systems in which a number other than 0 is assigned to the reference group and numbers other than 1 are used to represent group membership. As one example, some researchers use a coding system in which sex is coded female = 1 and male = 2 . This coding system will yield the same results for the correlations and t tests of regression coefficients as were described here. However, recall that the intercept is the value of $\hat{Y}$ when C has a value = 0. Thus, the intercept will represent the mean of Y for the nonexistent case in which gender = 0 in this coding system. As a second example, a coding system in which female = 2 and male = 0 might be used. Here, the intercept will correctly equal the M_Y in the male reference group. However, B for the dummy variable will have a different meaning than is normally intended. Recall that the unstandardized regression coefficient represents the value of a 1-unit change in the variable. However, in this coding scheme there is a 2-unit difference between the male and female coded values. Because B provides the value of a 1-unit change, it will now equal *one-half* of the difference between the male and female means on Y.

These examples illustrate that the use of dummy-variable-like coding systems lead to complications in the interpretation of the results of the MRC analysis of a single nominal independent variable. The moral here is "keep it simple": use a standard zero-one dummy coding system. In more complex models, the use of nonstandard dummy-variable-like coding systems may not only change the meaning of certain regression coefficients, it may also lead to inappropriate confidence intervals and significance tests for these coefficients. If a data set is encountered with such a nonstandard coding scheme, researchers are advised to transform (recode) the data to permit analysis using conventional dummy codes.

8.2.9 Dummy-Variable Coding When Groups Are Not Mutually Exclusive

As indicated in the beginning of this chapter, dummy coding is intended for the situation in which each subject may belong to only one group in the set. When this is not the case—for example, if the nominal scale is intended to measure ethnicity in a circumstance in which an individual may claim more than one ethnic identity—the interpretation of the coefficients resulting from the inclusion of the $g - 1$ variables in the equation predicting Y will necessarily change. The B coefficients will no longer readily reproduce the original Y means of the groups as in the analyses we have just reviewed. Instead, each B will represent the mean difference between the group coded 1 and the reference group *from each of which has been partialed the effects of the overlap in group membership.* Other coefficients such as partial and semipartial correlations will necessarily be similarly partialed for overlap among the groups. Interpretation of the variables in such cases needs to be done with *extreme care* to avoid erroneous conclusions.

8.3 UNWEIGHTED EFFECTS CODING

8.3.1 Introduction: Unweighted and Weighted Effects Coding

In Section 8.2 on dummy coding variables, we showed that all measures of the partial effect of a single code variable—the regression coefficients and the partial and semipartial correlation coefficients—are interpreted with respect to a reference group. Such interpretations are very often useful, which contributes to the popularity of dummy codes. However, situations arise in which such interpretations may *not* be optimal, as the following example adapted from Suits (1984) illustrates.

Imagine you have conducted a large study comparing the number of years of schooling attained by residents in different regions of the United States. Dummy-variable codes are used to represent residence in the four regions of the United States as shown in Table 8.2.2A, in which South is designated as the reference group. You are invited to present your findings to a Senate committee.

> If you explain in the usual language that you have "omitted the (code) variable for 'South'," a distinguished Southern senator might well demand indignantly, "Now just a minute here. Let me get this straight. You did what?"
>
> To straighten out the natural confusion, you might explain that you haven't really "left out" the South. On the contrary, you have "established the South as the reference group from which to measure the educational attainment in other regions as deviations." The resulting confusion and consternation among the rest of the committee can well be imagined (adapted with small changes from Suits, 1984, p. 178).

Problems such as those addressed by this educational research project are often better represented by other coding systems. For many research questions, the central issue is how the outcomes in each separate group differ from the average (mean) outcome for the entire sample. To answer such questions, we will use effects coding. In this coding system we again use $g - 1$ codes to represent the g groups. However, as we will see, the B_is now represent the deviation of the outcome for each separate group from the mean of the groups rather than from a selected reference group.

When the sample sizes of the groups differ, we need to decide between two possible comparisons with the separate group means. One possibility is the unweighted mean, in which each of the groups count equally. The unweighted mean is represented in the case of four groups and in general as

$$(8.3.1) \qquad M_U = \frac{M_1 + M_2 + M_3 + M_4}{4} = \frac{\sum M_i}{g}.$$

Even if there are 100 cases in group 1, and only 10 cases in group 2, 20 cases in group 3, and 50 cases in group 4, the means of each separate group contribute equally to the overall unweighted mean. This unweighted coding system is particularly useful when the groups represent different experimental treatment groups and differences in sample size are the result of incidental factors such as the cost or difficulty of mounting each experimental treatment. For the example noted, this would mean we wished to compare the mean years of each region of the country to the mean education of the four regions (treated equally).

The second possibility is to use the weighted mean, in which the number of cases in each group are involved in the computation. In the case of four groups and in general, the weighted mean is represented as

$$(8.3.2) \qquad M_W = \frac{n_1M_1 + n_2M_2 + n_3M_3 + n_4M_4}{n_1 + n_2 + n_3 + n_4} = \frac{\sum n_iM_i}{\sum n_i}.$$

Of importance, $\Sigma n_i = n$, the total sample size, so that the weighted mean $M_W = \Sigma Y/n$, the usual sample mean when group membership is ignored. Weighted effects coding will be of particular importance when the relative size of each group in the sample is representative of its proportion in the population. For example, if we have taken a random sample and wish to generalize the results to the population, weighted effects coding would be the approach of choice. In our earlier example, this would imply the total population of the United States as the reference, complete with its unequal population size in the various regions.

To understand this difference more fully, consider a group of researchers who wish to study the average income of adult residents of the southwestern United States. If they select a random sample of adult residents in the four southwestern states of Arizona, California, Nevada, and New Mexico, the great majority of the people included in the sample would be from the state with the largest population, California. California residents also have a substantially higher mean income than the other states. The use of the weighted mean would permit generalization to the income of residents of the southwestern region, which in fact is dominated in both population and income by residents of California. In contrast, the use of the unweighted mean would permit generalization to a hypothetical southwestern population in which each of the states contributed equally, regardless of its population.

In this Section (8.3), we consider unweighted effects coding in which comparisons are made with the unweighted mean. As noted earlier, this is often the most useful choice for analyzing data from experiments. In the next Section (8.4), we consider weighted effects codes, which are most useful when cases have been sampled from some larger population using random or representative sampling. When the sample sizes (n_i) are equal in each group, weighted effects codes simplify and become identical to unweighted effects codes.

8.3.2 Constructing Unweighted Effects Codes

Table 8.3.1 presents four examples of unweighted effects codes. In each case, $g - 1$ code variables will be needed to represent the g groups that comprise the nominal variable. In the case of unweighted effects coding, one of the groups must be chosen as the base for the coding scheme and is assigned a value of -1 on all of the code variables. We will term this group the *base group*. In contrast to dummy-variable coding (Section 8.2.1), the base group is often

TABLE 8.3.1
Illustration of Unweighted Effects Coding

A. Catholic as base group.

	Code variables		
Religion	C_1	C_2	C_3
Catholic	−1	−1	−1
Protestant	1	0	0
Jewish	0	1	0
Other	0	0	1

B. Protestant as base group (used in chapter).

	Code variables		
Religion	C_1	C_2	C_3
Catholic	1	0	0
Protestant	−1	−1	−1
Jewish	0	1	0
Other	0	0	1

C. Unweighted effects coding of experimental treatment groups ($g = 3$).

	Code variables	
Experimental group	C_1	C_2
Treatment 1	1	0
Treatment 2	0	1
Control	−1	−1

D. Dummy coding of sex ($g = 2$).

	Code variable
Sex	C_1
Female	1
Male	−1

selected to be the group for which comparisons with the mean are of *least* interest. This is because the MRC analyses do not *directly* inform us about this group. In unweighted effects coding the regression coefficients represent comparisons of the mean of each group, except the base group, with the unweighted mean.

Table 8.3.1A presents a set of unweighted effects codes using Catholic as the base group, and Table 8.3.1B presents a second set of unweighted effects codes using Protestant as the base group for our running example of four religious groups. The unweighted effects codes presented in Table 8.3.1B will be used throughout this section. Other unweighted effects coding schemes using Jewish or Other as the omitted group could also be constructed. Table 8.3.1C presents a set of unweighted effects codes for comparisons of three experimental treatment groups, and Table 8.3.1D presents the unweighted effect code for comparisons on sex.

To construct unweighted effects codes, we designate one group to serve as the "base group." In Table 8.3.1B, the Protestant group was chosen as the base group to facilitate direct comparison with the results of the dummy-coded analysis presented in Section 8.3. The base group is assigned a value of −1 for each code variable. Each of the other groups is assigned a value of 1 for one code variable and a value of 0 for all other code variables, paralleling the assignment of dummy codes. In Table 8.3.1B, Catholic is assigned [1 0 0], Jewish is assigned [0 1 0], and Other is assigned [0 0 1]. The critical difference between the dummy coding scheme of Table 8.2.1B and the present unweighted effects coding scheme is the set of codes for the Protestant base group [−1 −1 −1].

In this unweighted effects coding system, C_1 contrasts Catholic, C_2 contrasts Jewish, and C_3 contrasts Other with the unweighted mean of the four religious groups in the regression equation. The contrast of the base group with the unweighted mean is not given directly, but can be easily calculated. As with dummy variables, all $g - 1$ code variables

TABLE 8.3.2
Illustrative Data for Unweighted Effects Coding for Religion and Attitude Toward Abortion ($g = 4$)

Case no.	Group	DV	C_1	C_2	C_3
1	1	61	1	0	0
2	4	78	0	0	1
3	2	47	−1	−1	−1
4	1	65	1	0	0
5	1	45	1	0	0
6	4	106	0	0	1
7	2	120	−1	−1	−1
8	1	49	1	0	0
9	4	45	0	0	1
10	4	62	0	0	1
11	1	79	1	0	0
12	4	54	0	0	1
13	2	140	−1	−1	−1
14	1	52	1	0	0
15	2	88	−1	−1	−1
16	1	70	1	0	0
17	1	56	1	0	0
18	3	124	0	1	0
19	4	98	0	0	1
20	1	69	1	0	0
21	2	56	−1	−1	−1
22	3	135	0	1	0
23	2	64	−1	−1	−1
24	2	130	−1	−1	−1
25	3	74	0	1	0
26	4	58	0	0	1
27	2	116	−1	−1	−1
28	4	60	0	0	1
29	3	84	0	1	0
30	2	68	−1	−1	−1
31	2	90	−1	−1	−1
32	2	112	−1	−1	−1
33	3	94	0	1	0
34	2	80	−1	−1	−1
35	3	110	0	1	0
36	2	102	−1	−1	−1

Group means: $M_1 = 60.67; M_2 = 93.31; M_3 = 103.50; M_4 = 70.13.$
$M_U = (60.67 + 93.31 + 103.50 + 70.13)/4 = 81.90 = B_0.$

Note: DV is attitude toward abortion. For religious group, 1 = Catholic, 2 = Protestant, 3 = Jewish, 4 = Other. M_U is unweighted mean of four groups.

(here, 3) must be included in the regression equation to represent the overall effect of religion. Each code variable contributes 1 df to the $g - 1$ df that comprise religion. If any code variables are omitted, the interpretation of the results can dramatically change. Table 8.3.2 presents the data file that would be created using the unweighted effects codes from Table 8.3.1B, for our running example. Tables 8.3.3 and 8.3.4 present the results from the MRC analysis.

TABLE 8.3.3
Correlations, Means, and Standard Deviations of the Illustrative Data for Unweighted Effects Coding

		r					
		ATA	C_1	C_2	C_3	r^2_{Yi}	t_i (df = 34)
ATA		1.000	−.444	−.029	−.328	—	—
Catholic	C_1	−.444	1.000	.629	.595	.1974	−2.892*
Jewish	C_2	−.029	.629	1.000	.636	.0008	−0.170
Other	C_3	−.328	.595	.636	1.000	.1074	−2.022
	M	81.69	−.111	−.194	−.139		
	SD	27.49	.807	.730	.783		

$R^2_{Y.123} = .3549$; $F = 5.869^*$ ($df = 3, 32$).

$\tilde{R}^2_{Y.123} = .2944$.

Note: ATA = attitude toward abortion. $^*p < .05$. M is the unweighted mean of the group means.

8.3.3 The R^2 and the r_{Yi}s for Unweighted Effects Codes

R^2 and $\tilde{R}^2$

We first note that $R^2 = .3549$, $F = 5.899$, and $\tilde{R}^2 = .2945$, the same values as were obtained by dummy-variable coding. We remind readers that the different coding systems are alternative ways of rendering the information as to group membership into quantitative form. Each of these coding systems *taken as a set* carries all the group information and represents the same nominal variable. Given the same Y data they must yield the same R^2 and hence the same

TABLE 8.3.4
Analysis of Illustrative Data: Attitude Toward Abortion

A. Unweighted effects coding: partial and semipartial correlations and regression coefficients.

C_i	pr_i	pr_i^2	sr_i	sr_i^2	β_i	B_i	SE_{B_i}	t_i
C_1	−.481	.2310	−.440	.1937	−.5977	−21.23	6.85	−3.10*
C_2	.436	.1900	.389	.1513	.5500	21.60	7.88	2.74*
C_3	−.281	.0787	−.235	.0551	−.3217	−11.77	7.12	−1.65

Note: $df = 32$. $^*p < .05$.

B. Predicted values in groups.

$$\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$$
$$= -21.23C_1 + 21.60C_2 - 11.77C_3 + 81.90.$$

Catholic: $\hat{Y}_1 = -21.23(1) + 21.60(0) - 11.77(0) + 81.90 \quad = 60.67 \;= \hat{Y}_1.$
Protestant: $\hat{Y}_2 = -21.23(-1) + 21.60(-1) - 11.77(-1) + 81.90 \quad = 93.31 \;= \hat{Y}_2.$
Jewish: $\hat{Y}_3 = -21.23(0) + 21.60(1) - 11.70(0) + 81.90 \quad = 103.50 = \hat{Y}_3.$
Other: $\hat{Y}_4 = -21.23(0) + 21.60(0) - 11.77(1) + 81.90 \quad = 70.13 \;= \hat{Y}_4.$

Note: $B_0 = 81.90$ is unweighted mean of four groups means.

$$SD^2_{Y-\hat{Y}} = SD^2_Y(1 - R^2)\,\frac{n}{(n-k-1)} = 27.49^2(1 - .354944)\,\frac{36}{36-3-1} = 548.41.$$

F statistic. In contrast, the results for the individual code variables (C_is) change with changes in the coding scheme.

Pearson Correlations

The interpretation of the simple Pearson correlations between each of the code variables with Y is less straightforward for unweighted effects coding than for dummy coding. For each code variable C_i there are now three possible values: $+1, 0, -1$. For example, in Table 8.3.1B, the code variable C_1 has the following values:

Group	C_1
Catholic	+1
Protestant	−1
Jewish	0
Other	0

For C_1 the Pearson correlation with Y, R_{Yi}, reflects the contrast between the group coded 1 and the base group coded -1 with the effect of the other groups minimized.[5] If the sample sizes in the group coded $+1$ (here, Catholic, $n_1 = 9$) differs appreciably in size from the group coded -1 in the contrast (here, Protestant, $n_2 = 13$), it is prudent not to interpret the individual r_{Yi}s from unweighted effects codes. In the special case in which the n_is of these two groups are equal, the r_{Yi} for unweighted effects codes will have the same value and interpretation as sr_i from the dummy coding scheme when the reference group is the same as the base group.

As with dummy codes, unweighted effects codes are in general correlated with each other. This means that $R^2_{Y.123}$ will not equal the sum of the three r^2_{Yi}s. When all groups are of the same size, r_{ij} between any two code variables will be .50 regardless of the number of groups. When the groups have unequal ns, intercorrelations will be larger or smaller than .50 depending on the relative sizes of the groups. In Table 8.4.3, the r_{ij}s range from .595 to .636.

8.3.4 Regression Coefficients and Other Partial Effects in Unweighted Code Sets

Once again, we can substitute into our standard regression equation to help understand the meaning of each of the unstandardized regression coefficients. Substituting values of the unweighted effects codes from Table 8.3.1B, we find

$$
\begin{aligned}
&\text{Catholic:} && \hat{Y} = B_1(+1) + B_2(0) + B_3(0) + B_0 = B_1 + B_0 = M_1;\\
&\text{Protestant:} && \hat{Y} = B_1(-1) + B_2(-1) + B_3(-1) + B_0 = -B_1 - B_2 - B_3 + B_0 = M_2;\\
&\text{Jewish:} && \hat{Y} = B_1(0) + B_2(+1) + B_3(0) + B_0 = B_2 + B_0 = M_3;\\
&\text{Other:} && \hat{Y} = B_1(0) + B_2(0) + B_3(+1) + B_0 = B_3 + B_0 = M_4.
\end{aligned}
$$

In unweighted effects coding, B_0 is the unweighted mean of the four groups, $M_U = (M_1 + M_2 + M_3 + M_4)/4$. Each of the unstandardized regression coefficients represents the discrepancy of the corresponding group mean from the unweighted grand mean associated with a 1-unit change on C_i. Thus, B_1 represents the difference between the mean of the group coded 1 on C_1 (Catholic) and the unweighted grand mean of the four religious groups. B_2 represents the

[5]The word *minimized* is used to avoid lengthy discussion of a minor mathematical point. The minimum influence of the 0-coded groups on r_{Yi} is literally nil whenever the n of the group coded $+1$ equals the n of the group coded -1.

difference between the mean of the group coded 1 on C_2 (Jewish) and the unweighted grand mean. And B_3 represents the difference between the group coded 1 on C_3 (Other) and the unweighted grand mean. The difference between the mean of the omitted group (Protestant) and the unweighted grand mean is obtained by subtraction:

$$M_2 = -B_1 - B_2 - B_3 + B_0.$$

To illustrate these relationships, we estimated our standard regression equation (Eq. 8.2.3), $\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$, using the unweighted effects codes shown in Table 8.3.1B with Protestant as the base group. The result is shown in Eq. (8.3.3):

$$\hat{Y} = -21.23C_1 + 21.60C_2 - 11.77C_3 + 81.90. \tag{8.3.3}$$

Table 8.3.4 shows the results of substituting the unweighted effect codes corresponding into our standard regression equation. For each group, $\hat{Y}$ is identical to the mean for the religious group shown in Table 8.3.2. The graphical depiction of the results for unweighted effects codes is identical to that for dummy codes.

Confidence Intervals and Significance Tests for B_i

Procedures for constructing confidence intervals and conducting significance tests for each of the unstandardized regression coefficients are identical to those presented in Section 8.2.3 for dummy variables. We simply take $B_i \pm t_{SE_{B_i}}$ where $df = n - k - 1$. As before, $df = 32$ and $t = 2.037$. Note that the values of each B_i and SE_i have changed (compare Table 8.2.5 with Table 8.3.4) so the actual values of the confidence intervals will change. Substituting in the current values for B_i and SE_{B_i}, we find

$$\begin{aligned}
B_0&: \quad 81.90 \pm (2.037)(4.55) = 72.62 \text{ to } 91.18;\\
B_1&: \quad -21.23 \pm (2.037)(6.85) = -35.18 \text{ to } -7.28;\\
B_2&: \quad 21.60 \pm (2.037)(7.88) = 5.55 \text{ to } 37.65;\\
B_3&: \quad -11.77 \pm (2.037)(7.12) = -26.27 \text{ to } 2.73.
\end{aligned}$$

Significance tests for each of the unstandardized regression coefficients are presented in Table 8.3.4. Again, these values differ from those presented in Table 8.2.5 for the analysis of the dummy-coded variables. These changes reflect the change in the meaning and the value of the unstandardized regression coefficients between dummy coding and unweighted effects coding.

Two other significance tests may be of interest to researchers. First is the test of the difference between the mean of the base group and the unweighted grand mean. This test can be accomplished most simply by using another unweighted effects coding system with a different base group. For example, if we use the coding system depicted in Table 8.3.1A, in which Catholic is the base group, the test of B_1 provides a test of the difference between the mean of the Protestant group and the unweighted grand mean. Alternatively, this test can be performed using information from the original analysis with Protestant as the base group. The equation is as follows:

$$t_b = \frac{-g\sum B_i}{\sqrt{sd^2_{Y-\hat{Y}}\left[\frac{(g-1)^2}{n_b} + \sum\frac{1}{n_i}\right]}} \qquad df = n - g - 1.$$

In this equation, g is the number of groups comprising the nominal variable, k is the number of code variables, i runs from 1 to $g - 1$ (i.e., not including the base group), and n_b is the number of subjects in the base group. Applying this formula to our running example using the results presented in Table 8.3.4, we find

$$t_b = \frac{-4(-21.23 + 21.60 - 11.77)}{\sqrt{548.41\left(\frac{(4-1)^2}{13} + \frac{1}{9} + \frac{1}{6} + \frac{1}{8}\right)}} = \frac{11.41}{6.13} = 1.862 \qquad df = 32.$$

Since the critical value of t with $df = 32$ and $\alpha = .05$ is 2.037, we conclude that the mean of the Protestant base group does not differ from the unweighted mean of the four groups in the population.

Finally, researchers may be interested in testing the difference between the means of two groups i and j in the population. Recall that the analysis using dummy codes directly tests the difference between each group mean and the reference group. Thus, this question can be directly answered by recoding the data so that dummy codes are used as the set of code variables. The reference group for the dummy codes should be chosen so that it is one of the two groups involved in the comparison. For example, if the researcher wished to compare the Protestant and Catholic groups, the test of C_1 using the dummy coding system presented in Table 8.3.1A provides a test of the hypothesis.

Alternatively, these tests may be computed by hand using the output from the unweighted effects codes analysis. If neither of the groups being compared is the base group, Eq. (8.2.4) for comparing two groups i and j (neither of which is the base group) is used. This equation is reproduced here:

$$t = \frac{B_i - B_j}{\sqrt{SD^2_{Y-\hat{Y}}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}.$$

When the base group is included in the comparison, a different formula must be used since there is no regression coefficient directly available for the base group. This formula is as follows:

$$t = \frac{2B_i + \sum B_j}{\sqrt{SD^2_{Y-\hat{Y}}\left(\frac{1}{n_i} + \frac{1}{n_b}\right)}} \qquad (j \neq i) \qquad df = n - k - 1.$$

Note that the summation is over all of the B coefficients, except B_i, the regression coefficient for the group being compared with the base group. The numerator is a re-expression of $M_i - M_b$. Applying this formula to our running example for the comparison of Catholic and Protestant (omitted group),

$$t = \frac{2(-21.32) + (21.60 - 11.77)}{\sqrt{548.41\left(\frac{1}{9} + \frac{1}{13}\right)}} = -3.214.$$

This t value exceeds the critical value of $t = 2.037$ for $df = 32$ and $\alpha = .05$. Note that this is exactly the same value as the t for B_1 in the regression analysis using dummy codes presented in Table 8.2.5.

Once again, we encourage readers to exercise caution in conducting such comparisons, particularly when there is not a strong a priori prediction that two specific groups will differ. Multiple comparisons of group means increase the level of α for the study beyond the stated level, typically $\alpha = .05$, and special procedures described by Kirk (1995, Chapter 4) and Toothaker (1991) need to be taken.

Semipartial and Partial Correlations

As noted, the regression coefficients provide a contrast between a given group and the unweighted mean of all the groups. Because the mean of all the groups includes the group in question, this is functionally equivalent to contrasting this group with the remaining groups taken collectively. Thus, sr_i^2 is the proportion of Y variance accounted for by this contrast. Concretely, for our running example, sr_1^2 means that 19.4% of the variance in ATA scores is accounted for by the distinction between Catholic on one hand and equally weighted Protestant-Jewish-Other on the other hand. Thus, 19.4% of the ATA variance in the sample is accounted for by the "distinctiveness" of Catholics relative to the other groups.

The partial correlations relate the partialed effects-coded IV (e.g., $X_{1.23}$) with that part of Y left after the other variables have been removed. Thus, pr_1^2 gives

$$\frac{\text{ATA variance due to Catholic group distinctiveness}}{\text{ATA variance } \textit{not} \text{ accounted for by remaining groups' distinctiveness}}.$$

Recall that the denominator of pr_i^2 is always equal to or smaller than the denominator of sr_i^2 (which is 1.0). Hence, pr_i^2 will typically be larger than sr_i^2.

8.4 WEIGHTED EFFECTS CODING

8.4.1 Selection Considerations for Weighted Effects Coding

Weighted effects coding is most appropriate when the proportion of cases in each group in the sample can be considered to represent the corresponding proportion of cases in the population. This situation will most commonly occur when random or representative samples have been selected from a population. In weighted effects coding it is the comparison of each group with the aggregate population mean that is at issue. As a metaphor, unweighted effects codes may be thought of as the "Senate" option—since every state in the United States has two Senators regardless of its population size—whereas weighted effects codes would be the "House" option, since the 435 U.S. Representatives are divided among the states in proportion to their population size.

Like the situation with unweighted effects coding, one of the groups must be designated as the base group in the coding scheme. Once again, the group for which comparisons with the mean are of least interest will normally be chosen to be the base group because the MRC information regarding that group will be less accessible than for the other groups.

8.4.2 Constructing Weighted Effects

Table 8.4.1 presents four examples of weighted effects codes. These examples parallel exactly those in Table 8.3.1, permitting direct comparison of weighted and unweighted effect codes. Table 8.4.1A presents a set of weighted effect codes for religion using Catholic as the base group; Table 8.4.1B presents the set of weighted effect codes for religion using Protestant as the base group.[6] Table 8.4.1C presents a set of weighted effects codes for comparisons of three experimental treatment groups and Table 8.4.1D presents the weighted effect code for sex. In Tables 8.4.1A and B, we have written a general expression and the specific codes given the

[6]In practice, the decimal value corresponding to the fraction would be used for each code variable. Thus, in the Protestant base group in our example, the actual values entered would be $-.6923076$ ($=-9/13$), $-.4615384$ ($=-6/13$), and $-.6153846$ ($=-8/13$) for C_1, C_2, and C_3, respectively.

TABLE 8.4.1
Illustration of Weighted Effects Coding Systems

A. Religious groups: Catholic as base group.

Religion		General case: code variables C_1	C_2	C_3	Example: code variables C_1	C_2	C_3
Catholic	$(n_1 = 9)$	$-n_2/n_1$	$-n_3/n_1$	$-n_4/n_1$	$-13/9$	$-6/9$	$-8/9$
Protestant	$(n_2 = 13)$	1	0	0	1	0	0
Jewish	$(n_3 = 6)$	0	1	0	0	1	0
Other	$(n_4 = 8)$	0	0	1	0	0	1

B. Religious groups: Protestant as base group (used in chapter).

Religion		General case: code variables C_1	C_2	C_3	Example: code variables C_1	C_2	C_3
Catholic	$(n_1 = 9)$	1	0	0	1	0	0
Protestant	$(n_2 = 13)$	$-n_1/n_2$	$-n_3/n_2$	$-n_4/n_2$	$-9/13$	$-6/13$	$-8/13$
Jewish	$(n_3 = 6)$	0	1	0	0	1	0
Other	$(n_4 = 8)$	0	0	1	0	0	1

C. Unweighted effects coding of experimental treatment groups ($g = 3$).

Experimental group	Code variables C_1	C_2
Treatment 1	1	0
Treatment 2	0	1
Control	$-n_1/n_3$	$-n_2/n_3$

D. Dummy coding of sex ($g = 2$).

Sex	Code variable C_1
Female	1
Male	$-n_1/n_2$

stated sample sizes for the groups in the illustrative example. Whenever the sample size of the groups changes, the values of the unweighted effects codes for the base group change as well.

Once again, we arbitrarily designate one group to be the base group. Table 8.4.1B, which presents the coding scheme used in this section, designates Protestant as the base group. For the other groups, the coding exactly parallels that of unweighted effects codes. The critical difference between the unweighted effects coding system of Table 8.3.1 and the weighted effects coding system of Table 8.4.1 is in the set of codes assigned to the base group. The values of each code are weighted for the base group to reflect the different sample sizes of each of the groups. For each code variable, the base group receives the value of *minus* the ratio of the size of the group coded 1 and the size of the base group. Note that when sample sizes are equal across groups, $n_1 = n_2 = n_3 = n_4 = \cdots = n_g$, the codes for the base group simplify to $[-1 \quad -1 \quad -1 \quad -1 \quad \cdots \quad -1]$. Under these conditions, weighted effects codes are identical to unweighted effects codes.

In this coding system, C_1 contrasts Catholic, C_2 contrasts Jewish, and C_3 contrasts Other with the weighted mean of the four religious groups in the regression equation. Recall that the weighted mean is $\Sigma n_i M_i / \Sigma n_i = (\Sigma Y)/n$, where $n = \Sigma n_i$. This is the mean of the scores of all of the subjects on the dependent variable, ignoring group membership. Once again, each code variable contributes 1 *df* to the $g - 1$ *df* that comprise religion. If any of the code variables are omitted, the interpretation of the results can dramatically change. Unlike the other coding systems for nominal variables we have considered thus far, weighted effects codes are centered. In our example (and in general), the means for C_1, C_2, and C_3 are each equal to 0 in the sample. The correlations between pairs of code variables are also smaller than for unweighted effect codes, ranging from .347 to .395.

8.4.3 The R^2 and $\tilde{R}^2$ for Weighted Effects Codes

Table 8.4.2 presents the results from the MRC analysis. Once again, we note that $R^2 = .3549, F = 5.899$, and $\tilde{R}^2 = .2945$, values identical to those we obtained using dummy coding and unweighted effects coding. Each coding system taken as a set is equivalent because it represents the same nominal variable.

TABLE 8.4.2
Analysis of Illustrative Data: Attitude Toward Abortion

A. Weighted effects coding: regression coefficients.

C_i	β_i	B_i	SE_{B_i}	t_i
C_1	−.4975	−21.03	6.76	−3.11*
C_2	.3915	21.81	8.72	2.50*
C_3	−.2522	−11.57	7.30	−1.58

$R^2_{Y.123} = .3549$ $\quad F = 5.869^*$ $(df = 3, 32)$

$\tilde{R}^2_{Y.123} = .2945$

Note: $df = 32$. $^*p < .05$.

B. Predicted values in groups.

$$\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$$
$$= -21.03C_1 + 21.81C_2 - 11.57C_3 + 81.69.$$

Catholic:	$\hat{Y}_1 = -21.03(1) + 21.81(0) - 11.57(0) + 81.69$	$= 60.67 = \hat{Y}_1$.
Protestant:	$\hat{Y}_2 = -21.03(-9/13) + 21.81(-6/13) - 11.57(-8/13) + 81.69$	$= 93.31 = \hat{Y}_2$.
Jewish:	$\hat{Y}_3 = -21.03(0) + 21.81(1) - 11.57(0) + 81.69$	$= 103.50 = \hat{Y}_3$.
Other:	$\hat{Y}_4 = -21.03(0) + 21.81(0) - 11.57(1) + 81.69$	$= 70.13 = \hat{Y}_4$.

Note: $B_0 = 81.69$ is unweighted mean of four group means.

$$SD^2_{Y-\hat{Y}} = SD^2_Y(1 - R^2)\frac{n}{(n-k-1)} = 27.49^2(1 - .354944)\frac{36}{36-3-1} = 548.41.$$

8.4.4 Interpretation and Testing of B with Weighted Codes

To understand the meaning of each of the regression coefficients, we can take the usual strategy of substituting the value of each code variable into our standard regression equation. Although this procedure leads to the correct answers, the algebra becomes very tedious (see the appendix of West, Aiken, & Krull, 1996 for the algebraic derivation). Fortunately, the *results* of the algebra are not complex; these results are presented here:

Catholic:	$\hat{Y} = B_1 + B_0 = M_1$;
Protestant:	$\hat{Y} = -(n_1/n_2)B_1 - (n_3/n_2)B_2 - (n_4/n_2)B_3 + B_0 = M_2$;
Jewish:	$\hat{Y} = B_2 + B_0 = M_3$;
Other:	$\hat{Y} = B_3 + B_0 = M_4$.

In weighted effects coding, B_0 is the *weighted* mean of the groups. Each of the regression coefficients represents the deviation of the corresponding group mean from the weighted mean of the entire sample. Thus B_1 represents the difference between the mean of the group coded 1 on C_1 (Catholic) and the weighted mean, B_2 represents the difference between the mean of the group coded 1 on C_2 (Jewish) and the weighted mean, and, B_3 represents the difference between the mean of the group coded 1 on C_3 and the weighted mean. The difference between the mean of the Protestant base group and the weighted mean may be obtained by subtraction:

$$M_2 = -(n_1/n_2)B_1 - (n_3/n_2)B_2 - (n_4/n_2)B_3 + B_0.$$

These interpretations of B_1, B_2, and B_3 using weighted effects codes parallel those of unweighted effects except that the mean of each group is contrasted with the weighted mean rather than the unweighted mean of the set of groups.

To illustrate these relationships, we estimated our standard regression equation, $\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$, using the weighted effects codes shown in Table 8.4.1B with Protestant as the base group. The result is

$$\hat{Y} = -21.03C_1 + 21.81C_2 - 11.60C_3 + 81.69.$$

Table 8.4.2 shows the results of substituting the weighted effect codes into the standard regression equation. For each group, $\hat{Y}$ is again identical to the mean for the group.

Confidence Intervals and Significance Tests for B_i

Procedures for constructing confidence intervals and conducting significance tests for each of the unstandardized coefficients are identical to those presented in Sections 8.2.3 for dummy codes and 8.3.3 for unweighted effects codes. To construct confidence intervals, we take $B_i \pm tSE_{B_i}$, where $df = n - k - 1$. Substituting in the current values for B_i and SE_{B_i}, we find

B_0:	$81.69 \pm (2.037)(3.90)$	$= 73.75$	to	89.63;
B_1:	$-21.03 \pm (2.037)(6.76)$	$= -34.80$	to	-7.26;
B_2:	$21.81 \pm (2.037)(8.73)$	$= 4.03$	to	39.59;
B_3:	$-11.60 \pm (2.037)(7.30)$	$= -26.47$	to	3.27.

Significance tests divide each B_i by its corresponding $SE, t = B_i/SE_{B_i}$. These values are presented in Table 8.4.2. Note that these values differ from those presented in Table 8.2.5 for dummy codes and Table 8.3.4 for unweighted effects codes. These discrepancies reflect the differences in the meaning and the value of the mean comparisons reflected by the Bs in the three coding systems. Recall that in dummy coding, the Bs represent the comparison of each group mean with the reference group mean. In unweighted effects coding, the Bs represent the comparison of each group mean with the unweighted mean. And in weighted effects coding, the Bs represent the comparison of each group mean with the weighted mean. Analysts should decide in advance which coding scheme represents their research question of interest and report the corresponding results. In the present example, the discrepancy in the results for unweighted and weighted effects codes is not large because the ns did not differ greatly. Other data sets can produce larger (or smaller) differences between unweighted and weighted effects coding depending on the specific values of the n_is and the M_is in the groups.

Determining whether the mean of the base group differs from the weighted mean is most easily done by reanalyzing the data using a coding system that provides the answer directly.[7] In this case the use of another weighted effects coding system with a different omitted group will provide the answer. For example, using the coding system in Table 8.4.1A in which Catholic is the base group, the significance test of C_1 provides a test of the difference between the mean of the Protestant group and the unweighted mean. The significance of the difference between two group means can be tested through the use of dummy codes (see Section 8.2). A dummy coding scheme should be chosen in which the omitted group is one of the two groups of interest.

8.5 CONTRAST CODING

8.5.1 Considerations in the Selection of a Contrast Coding Scheme

Researchers often have specific research questions or formal hypotheses based on substantive theory that can be stated in terms of expected mean differences between groups or combinations of groups. For example, consider a team of drug abuse researchers who wish to evaluate the effectiveness of two new prevention programs. They have collected data on the amount of drug use (dependent variable) on three groups of children: (a) children who are exposed to the drug prevention program in their middle school classrooms (school-based-only program), (b) children who are exposed to the drug prevention in their middle school classrooms in addition to a home-based drug prevention program led by their parents (school-based plus home-based program), and (c) children not exposed to a prevention program (control group). The researchers have two explicit, a priori hypotheses:

1. Children in the two prevention groups ($a + b$) will have less drug use than children in the control group (c);
2. Children in the school-based plus home-based program (b) will have less drug use than children in the school-based-only program (a).

Contrast codes provide a method of testing such focused hypotheses. They are used when researchers have specific research questions or hypotheses, particularly those that involve comparison of means of combined groups. As we will see, the particular contrast codes

[7]Paralleling unweighted effects codes, algebraic expressions may be written for these significance tests for weighted effects codes. For comparisons involving the base group, the calculations quickly become very tedious as the number of groups increases. Using recode statements to develop a coding system that directly answers the questions of interest and then reanalyzing the data is a far simpler alternative. The exception is the test of the difference between two group means (not involving the base group), which is identical to the same test for dummy and unweighted effects codes (see Eq. 8.2.4).

selected depend on the researchers' hypotheses. Many methodologists (Abelson, 1995; Judd, McClelland, & Culhane, 1995; Rosenthal & Rosnow, 1985) strongly recommend the use of such contrasts in order to sharpen the interpretation of the results. The method allows the researcher to test the specific hypotheses of interest rather than some other hypothesis that is a consequence of simply using a default coding scheme such as dummy variables. The use of contrasts may increase the power of the statistical test to reject a false null hypotheses relative to less focused, omnibus tests. For readers familiar with ANOVA, contrast codes are the familiar a priori or planned comparisons discussed in traditional ANOVA texts (e.g., Kirk, 1995; Winer, Brown, & Michels, 1991). In the next section we initially focus on the mechanics of constructing contrast codes, returning later to the issue of choosing the set of contrast codes that best represents the researchers' hypotheses.

8.5.2 Constructing Contrast Codes

Overview and Illustration

To illustrate the use of code variables for contrasts, let us reconsider the three drug abuse prevention programs in the preceding example. The first hypothesis compares the unweighted mean of the two prevention groups with the mean of the control group. This hypothesis is represented in Table 8.5.1D by the C_1 code variable $[+\frac{1}{3} \quad +\frac{1}{3} \quad -\frac{2}{3}]$. The second hypothesis compares the mean of the school-based only program with the mean of the school-based plus home-based program. This hypothesis is represented in Table 8.5.1D by the C_2 code variable $[\frac{1}{2} \quad -\frac{1}{2} \quad 0]$. There are $g = 3$ groups in this example, so $g - 1 = 2$ code variables are necessary to represent the categorical IV of treatment group.

We use three rules to construct contrast codes. The first two rules are part of the formal statistical definition of a contrast. The third rule is not required, but it greatly simplifies the interpretation of the results.

Rule 1. The sum of the weights across all groups for each code variable must equal zero. In the prevention program example, $\frac{1}{3} + \frac{1}{3} - \frac{2}{3} = 0$ for C_1 and $\frac{1}{2} - \frac{1}{2} + 0 = 0$ for C_2.

Rule 2. The sum of the products of each pair of code variables, C_1C_2, must equal 0. In the example,

group 1 (school-based only): $(\frac{1}{3})(\frac{1}{2}) = \frac{1}{6}$;

group 2 (school + home–based): $(\frac{1}{3})(-\frac{1}{2}) = -\frac{1}{6}$;

group 3 (control): $(-\frac{2}{3})(0) = 0$.

The sum of the products of the code variables $= \frac{1}{6} - \frac{1}{6} + 0 = 0$. When the group sizes are equal ($n_1 = n_2 = n_3 = \cdots = n_g$), this rule guarantees that the contrast codes will be orthogonal so they will share no overlapping variance.

Rule 3. As will be discussed in detail in our more formal presentation, the difference between the value of the set of positive weights and the value of the set of negative weights should equal 1 for each code variable. In our example, $\frac{1}{3} - (-\frac{2}{3}) = 1$ for C_1 and $\frac{1}{2} - (-\frac{1}{2}) = 1$ for C_2. This rule ensures easy interpretation: Each unstandardized regression coefficient corresponds exactly to the difference between the unweighted means of the groups involved in the contrast.[8]

[8]If the present set of contrast codes are multiplied by a constant, the new set of codes also meets the formal definition of a contrast. Multiplying a set of contrast codes by a constant does not affect any standardized measure. The Pearson, partial, and semipartial correlations as well as the standardized regression coefficient are all unaffected by

More Formal Presentation

To facilitate a more formal presentation of contrast codes, we need to define each contrast weight in the table, C_{hi}. C_{hi} is the weight for the hth group (row) for the ith code variable (column). In our example (see Table 8.5.1D), $C_{31} = -\frac{2}{3}$ for control group, code variable 1; $C_{32} = 0$ for control group, code variable 2; and $C_{21} = +\frac{1}{3}$ for school-based plus home-based group, code variable 1. This notation allows us to refer to the specific weights corresponding to the value of the code variable for a specific group.

Let us first consider the construction of a code variable corresponding to a single contrast. A contrast for g sample means (or other statistics) is defined as any linear combination of them of the form

$$\text{Contrast} = C_{1i}M_1 + C_{2i}M_2 + \cdots + C_{gi}M_g$$

This equation represents the contrast produced by the ith code variable. The values corresponding to each group for the code variable must be chosen subject to the restriction stated earlier as Rule 1. Rule 1 is stated more formally below as Eq. (8.5.1), where h stands for group and i stands for code variable.

$$\sum_{h=1}^{h=g} C_{hi} = 0, \quad i.e., \; C_{1i} + C_{2i} + \cdots + C_{gi} = 0 \tag{8.5.1}$$

To see how the form of Eq. (8.5.1) expresses a contrast, consider any set of g groups comprising a nominal variable G. We partition the g groups into three subsets: (1) a subset u, (2) a subset v that we wish to contrast with subset u, and (3) a subset w containing groups, if any, that we wish to exclude from the contrast. The contrast compares the unweighted mean of the groups in subset u with the unweighted mean of the groups in subset v.

As a concrete example, consider the responses to a question about respondent's occupation in a survey. Respondents are classified into one of nine occupational groups based on their responses to the survey question. The first 4 ($=u$) response options represent "white collar" occupations (e.g., educator; medical professional), the next 3 ($=v$) response options represent "blue collar" occupations (e.g., skilled laborer; unskilled laborer), and the final two ($=w$) response options represent other occupational categories (e.g., unemployed, did not answer). The researcher is interested in comparing the responses of white collar and blue collar workers to other survey items measuring their beliefs about equal pay for women. To accomplish this comparison, we assign the contrast code for each group in subset u as $-v/(u+v)$ and the contrast for each group in subset v as $+u/(u+v)$. In the present example comparing white collar and blue collar workers, $-v/(u+v) = -3/(4+3)$ and $+u/(u+v) = +4/(4+3)$, so that the contrast may be expressed as

$$\text{contrast} = -\frac{3}{7}M_{Y_1} - \frac{3}{7}M_{Y_2} - \frac{3}{7}M_{Y_3} - \frac{3}{7}M_{Y_4} + \frac{4}{7}M_{Y_5} + \frac{4}{7}M_{Y_6} + \frac{4}{7}M_{Y_7} + 0M_{Y_8} + 0M_{Y_9}.$$

Note that the set of C_{hi} coefficients that comprise the contrast satisfy the restriction in Eq. (8.5.1):

$$\sum_{h=1}^{h=g} C_{hi} = -\frac{3}{7} - \frac{3}{7} - \frac{3}{7} - \frac{3}{7} + \frac{4}{7} + \frac{4}{7} + \frac{4}{7} + 0 + 0 = \frac{-12}{7} + \frac{+12}{7} = 0.$$

Thus, we can state generally how the values of the code variables for each group are assigned for a specific contrast represented by code variable C_i. The values of C_{hi} for each of the u

---

such transformation. Rule 3 is emphasized here because it permits straightforward interpretation of the unstandardized regression coefficients. With other contrast coding schemes, the unstandardized regression coefficient will represent the mean difference of interest multiplied by a constant.

groups is $-v/u+v$; the value of C_{hi} for each of the v groups is $u/u+v$; and the value of C_{hi} is 0 for each of the w groups, if any. Once again, the choice of these values for the code variable allow us to interpret the unstandardized regression coefficient B_i corresponding to the code variable as the difference between the unweighted means of the groups in u and the groups in v. If the unweighted mean of the groups in u and the unweighted mean of the groups in v are precisely equal, then the value of B_i for the contrast will be 0.

Throughout this chapter we have seen that the full representation of the information entailed in membership in one of g groups requires a set of $g-1$ code variables. The contrast described here is only one member of such a set. A total of $g-1$ code variables must be specified to represent the full set of contrasts. For example, 8 code variables will be required to represent the 9 occupational groups in our example.[9]

In our overview presentation of contrast codes, we also stated Rule 2. Rule 2 establishes the part of the definition of contrast codes that each possible pair of code variables must be linearly independent. Formally, this condition can be written as

$$\sum_{h=1}^{h=g} C_{hi}C_{hi'} = 0, \quad i.e., \; C_{1i}C_{1i'} + C_{2i}C_{2i'} + \cdots + C_{gi}C_{gi'} = 0,$$

where the C_{hi} represent the weights for code variable i of each of the groups and the $C_{hi'}$ represent the weights for code variable i' for each of the groups, where h identifies the group number from 1 to 9.

Three different sets of contrast codes for our example of religions and ATA are shown in Table 8.5.1A, B, and C. Applying the linear independence (Rule 2) condition to each of the three possible pairs of contrast codes (C_1C_2, C_1C_3, C_2C_3) in Part A, we find

$$C_1C_2\!: (½)(½) + (½)(-½) + (-½)(0) + (-½)(0) = 0;$$
$$C_1C_3\!: (½)(0) + (½)(0) + (-½)(½) + (-½)(-½) = 0;$$
$$C_2C_3\!: (½)(0) + (-½)(0) + (0)(½) + (0)(-½) = 0.$$

Rule 2 is met for all possible pairs of contrast codes. Note, however, that the set of contrasts will *not* be orthogonal (i.e., share no overlapping variance) unless all groups have equal sample sizes ($n_1 = n_2 = \cdots = n_g$). All of the sets of contrast codes displayed in each panel of Table 8.5.1 are independent.

Choosing Among Sets of Contrast Codes

A very large number of potential contrasts may be constructed. Indeed, the total number of different means of group means contrasts among g groups is $1 + [(3^g - 1)/2] - 2^g$. For $g = 4$, this number is 25; for $g = 6$, it is 301; and for $g = 10$, it is 28,501. However, relatively few of these contrasts will test hypotheses that are meaningful to the researcher. The goal in choosing contrast codes is to select a set of $g-1$ codes that map neatly onto the researcher's hypotheses of interest.

Consider the set of contrast codes displayed in Table 8.5.1A. Code variable 1 contrasts the unweighted mean of Protestant and Catholic with the unweighted mean for Jewish and Other. Code variable 2 contrasts Protestant with Catholic. Code variable 3 contrasts Jewish with Other. Suppose a researcher hypothesized that members of minority religions (Jewish, Other)

---

[9]In the case where the group sizes are precisely equal, the contrasts will be independent and code variables can be dropped from the regression equation. However, even when there are equal ns, including all of the $g-1$ contrasts provides an important check on model specification (see Sections 4.3.1 and 4.4.1). Abelson and Prentice (1997) strongly emphasize the importance of checking the fit of the contrast model.

TABLE 8.5.1

A. Contrast A: minority versus majority religions.

Religion	Code variables C_1	C_2	C_3
Catholic	1/2	1/2	0
Protestant	1/2	−1/2	0
Jewish	−1/2	0	1/2
Other	−1/2	0	−1/2

C. Contrast C: religious groupings.

Religion	Code variables C_1	C_2	C_3
Catholic	1/4	1/3	1/2
Protestant	1/4	1/3	−1/2
Jewish	1/4	−2/3	0
Other	−3/4	0	0

Note: C_1: Western religions vs. Other
C_2: Christian vs. Jewish
C_3: Catholic vs. Protestant

B. Contrast B: value on women's rights.

Religion	Code variables C_1	C_2	C_3
Catholic	−1/2	0	−1/2
Protestant	1/2	−1/2	0
Jewish	1/2	1/2	0
Other	−1/2	0	1/2

D. Contrast coding of experimental treatment groups ($g = 3$).

Experimental group	Code variables C_1	C_2
Treatment 1	1/3	1/2
Treatment 2	1/3	−1/2
Control	−2/3	0

are more opposed to restrictions on religious freedom than members of majority religions (Catholic, Protestant). If majority versus minority religious status is the only critical difference among the groups, the contrast associated with code variable 1 should show a large difference. However, the contrast associated with code variable 2, which compares the two majority religions, and the contrast associated with code variable 3, which compares the two minority religions, should not differ appreciably.

Table 8.5.1(B) presents a different set of contrast codes. Here, code variable 1 contrasts the unweighted mean of Protestant and Jewish with the unweighted mean of Catholic and Other. Code variable 2 contrasts Protestant with Jewish and code variable 3 contrasts Catholic with Other. Suppose the researcher has a theory that suggests that the Protestant and Jewish religions place a particularly high value on women's rights, whereas Catholic and Other religions place a lower value on women's rights. In this example, the contrast associated with code variable 1 would be expected to show a substantial difference, whereas the contrasts associated with code variables 2 and 3 should not differ appreciably. Table 8.5.1(C) provides yet another set of contrasts among the religious groups.

Each of the sets of contrast codes represents a different partitioning of the variance associated with the g groups. The strategy of contrast coding is to express the researcher's central hypotheses of interest in the form of $g-1$ independent contrast codes. The MRC analysis then directly yields functions of the contrast values, confidence intervals, and significance tests.

Contrast codes are centered; however, the corresponding contrast $= C_{1i}M_1 + C_{2i}M_2 + \cdots + C_{gi}M_g$ may or may not be centered. The means for the contrasts corresponding to C_1, C_2, and C_3 will equal 0 only if the number of subjects in each group is equal. However, when there is one categorical IV, the correlations among the contrast code variables will typically be low to moderate. In our running example of religion and attitudes toward abortion, the correlations between the contrast variables range from −.12 to +.11 for the codes presented in Table 8.5.1A and from −.26 to +.04 for the codes presented in Table 8.5.1B.

8.5.3 R^2 and $\tilde{R}^2$

Tables 8.5.2 and 8.5.3 present the results from the two separate MRC analysis for the sets of contrast codes presented in Table 8.5.1A and B, respectively. We again note that $R^2 = .3549, F = 5.899$, and $\tilde{R}^2 = .2945$ for both analyses. These values are identical to those obtained for dummy, unweighted effects, and weighted effects coding.

8.5.4 Partial Regression Coefficients

The contrast codes recommended here lead to unstandardized regression coefficients that are directly interpretable. The unstandardized regression coefficients will equal the difference between the unweighted mean of the means of the groups contained in u and the unweighted mean of the means of the groups contained in v. To illustrate, consider the contrast coding scheme in Table 8.5.1A: C_1 represents the difference between majority religions (Catholic, Protestant) and minority religions (Jewish, Other). We see the value of B_1 is −9.82. We also see that the mean ATA are Catholic = 60.67, Protestant = 93.31, Jewish = 103.50, and Other = 70.13. The difference in the unweighted means between the two sets is

$$\frac{M_1+M_2}{2}-\frac{M_3+M_4}{2}=\frac{60.67+93.31}{2}-\frac{103.50-70.13}{2}=-9.83.$$

Similarly, the value of $B_2 = -32.64$, which is equal to the difference between the Catholic (60.67) and Protestant (93.31), means and the value of $B_3 = 33.38$ which is equal to the

TABLE 8.5.2
Analysis of Illustrative Data: Attitude Toward Abortion

A. Contrast coding: majority versus minority religions.

C_i	pr_i	pr_i^2	sr_i	sr_i^2	β_i	B_i	SE_{B_i}	t_i
C_1	−.209	.0437	−.172	.0296	−.1742	−9.82	8.11	−1.21
C_2	−.494	.2331	−.456	.2083	−.4594	−32.64	10.15	−3.21*
C_3	.423	.1789	.375	.1404	.3770	33.38	12.65	2.64*

$R^2_{Y.123} = .3549$ $F = 5.869^*$ $(df = 3, 32)$

$\tilde{R}^2_{Y.123} = .2945$

Note: $^*p < .05$.

B. Predicted values in groups.

$$\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$$
$$= -9.83C_1 - 32.64C_2 + 33.38C_3 + 81.90$$

Catholic: $\hat{Y}_1 = -9.83(.5) - 32.64(.5) + 33.38(0) + 81.90 = 60.67 = \hat{Y}_1$

Protestant: $\hat{Y}_2 = -9.83\,(.5) + 21.60(-.5) + 33.38(0) + 81.90 = 93.31 = \hat{Y}_2$

Jewish: $\hat{Y}_3 = -9.83\,(-.5) + 21.60(0) + 33.38(.5) + 81.90 = 103.50 = \hat{Y}_3$

Other: $\hat{Y}_4 = -9.83(-.5) + 21.60(0) + 33.38(-.5) + 81.90 = 70.13 = \hat{Y}_4$

$$SD^2_{Y-\hat{Y}} = SD^2_Y(1-R^2)\frac{n}{(n-k-1)} = 27.49^2(1-.3549)\frac{36}{36-3-1} = 548.41.$$

Note: $B_0 = 81.90$ is unweighted mean of four group means. Decimal value of each contrast code value is used (e.g. $1/2 = .5$).

TABLE 8.5.3
Analysis of Illustrative Data: Attitude Toward Abortion

A. Contrast coding: pro- versus anti-women's rights religions.

C_i	pr_i	pr_i^2	sr_i	sr_i^2	β_i	B_i	SE_{B_i}	t_i
C_1	.584	.3411	.578	.3339	.599	33.01	8.11	4.07*
C_2	.154	.0237	.125	.0157	.130	10.19	11.56	.88
C_3	.145	.0120	.118	.0139	.118	9.46	11.38	.83

$R^2_{Y.123} = .3549$ $F = 5.869^*$ $(df = 3, 32)$

$\tilde{R}^2_{Y.123} = .2945$

Note: $^*p < .05$.

B. Predicted values in groups.

$$\hat{Y} = B_1C_1 + B_2C_2 + B_3C_3 + B_0$$
$$= 33.01C_1 + 10.19C_2 + 9.46C_3 + 81.90$$

Catholic: $\hat{Y}_1 = +33.01(-.5) + 10.19(0) + 9.46(.5) + 81.90 = 60.67 = \hat{Y}_1$

Protestant: $\hat{Y}_2 = +33.01(.5) + 10.19(-.5) + 9.46(-.5) + 81.90 = 93.31 = \hat{Y}_2$

Jewish: $\hat{Y}_3 = +33.01(.5) + 10.19(.5) + 9.46(0) + 81.90 = 103.50 = \hat{Y}_3$

Other: $\hat{Y}_4 = +33.01(-.5) + 10.10(0) + 9.46(0) + 81.90 = 70.13 = \hat{Y}_4$

$$SD^2_{Y-\hat{Y}} = SD^2_Y(1 - R^2)\frac{n}{(n-k-1)} = 27.49^2(1 - .3549)\frac{36}{36-3-1} = 548.41.$$

Note: $B_0 = 81.90$ is unweighted mean of four group means. Decimal value of each contrast code value is used (e.g. ½ = .5).

difference between the Jewish (103.50) and Other (70.13) means. Finally, the intercept, $B_0 =$ 81.90 equals the unweighted mean of the four group means, just as in unweighted effects coding.

Table 8.5.2 provides further insight into how contrast coding partitions the variance of the set of groups. When the contrast code values are substituted into the regression equation, the predicted value $\hat{Y}$ for each group is once again equal to the unweighted group mean. The contrast model yields the mean for a group by adding to the unweighted mean of the group means the effect provided by each group's role in the set of contrasts. For example, the Protestant mean M_2 comes about by adding the unweighted mean (81.90), one-half of the majority-minority contrast (C_1) [½(−9.82)] = −4.91, and minus one-half the value of the Protestant-Catholic contrast (C_2) [(−½) (−32.64)] = 16.32, but none of the irrelevant Jewish versus Other contrast (C_3) [(0)(33.38)] = 0. Thus, for Protestants, $M = 81.90 - 4.91 + 16.32 + 0 = 93.31$.

Contrast-coded unstandardized B_i values are *not* affected by varying sample sizes. Because they are a function only of unweighted means, the expected value of each contrast is invariant over changes in relative group size.[10] Unfortunately, standardized β_i coefficients lack this property, rendering them of little use in many applications involving categorical IVs. This same property of invariance of unstandardized B_i regression coefficients and lack of invariance of standardized β_i coefficients also holds for dummy and unweighted effects codes.

[10]Contrast coding is most frequently used to test a priori hypotheses in experiments. In observational studies in which large random samples have been selected from a population, it may be useful to construct weighted contrast codes that take sample size into account. In weighted contrast codes, the intercept is the weighted mean of the group means and each contrast represents the weighted mean of the groups in set u versus the weighted mean of the groups in set v. Serlin and Levin (1985) outline methods of constructing weighted effect codes.

Confidence Intervals and Significance Tests for B_i

Procedures for constructing confidence intervals and conducting significance tests for each of the unstandardized regression coefficients are identical to those presented in Section 8.2.5 for dummy variables. Again, we simply take $B_i \pm tSE_{B_i}$, with $df = n - k - 1$, where k is the number of code variables ($= g - 1$). Substituting in the values for B_i and SE_{B_i} for the regression analysis presented in Table 8.5.2 corresponding to the contrast code variables in Table 8.5.1A, we find

$$
\begin{aligned}
B_0&: \quad 81.90 \pm (2.037)(4.05) &&= 73.65 \text{ to } 90.15;\\
B_1&: \quad -9.83 \pm (2.037)(8.11) &&= -26.35 \text{ to } 6.69;\\
B_2&: \quad -32.64 \pm (2.037)(10.15) &&= -53.32 \text{ to } -11.97;\\
B_3&: \quad 33.38 \pm (2.037)(12.65) &&= 7.59 \text{ to } 59.17.
\end{aligned}
$$

Null hypothesis t tests for each B_i are calculated by dividing the coefficient by its SE. These values are presented in Table 8.5.2A.

Semipartial and Partial Correlations

The squared semipartial correlation, sr_i^2, is the proportion of the total Y variance accounted for by contrast i in the sample. Thus, from Table 8.5.2, .0296 of the total ATA variance is accounted for by C_1, the majority-minority religion contrast. Each of the groups contributing to the contrast is unweighted; religions within the majority category count equally and religions within the minority category count equally. However, the relative sizes of the two categories being contrasted (e.g., majority vs. minority religion) is influential, with sr being maximal when the two categories have equal sample sizes. Similarly, $sr_2^2 = .2082$ is the proportion of the ATA variance accounted for by C_2, the Protestant-Catholic distinction, and $sr_3^2 = .1404$ is the proportion of the ATA variance accounted for by C_3, the Jewish-Other distinction. Note that the sum of the sr_i^2s does not equal $R^2_{Y.123}$ because the n_is are not equal, and, as we saw earlier, the correlations between the contrasts are not equal to 0.

The squared partial correlation, pr_i^2, is the proportion of that part of the Y variance *not accounted for by the other contrasts* that is accounted for by contrast i. Thus, $pr_2^2 = .2331$ indicates that C_2 (the Protestant-Catholic distinction) accounts for 23.3% of the Y variance remaining after the variance due to contrasts C_1 and C_3 have been removed. As usual, $pr_2^2 = .2331$ is larger than $sr_2^2 = .2083$. Recall that in MRC, $pr_i^2 \geq sr_i^2$, the equality holding when the other IVs account for no variance.

The choice between sr and pr depends, as always, on what seems to be more appropriate interpretive framework, the total Y variance or the residual Y variance after the effects of the other variables have been removed. With contrast codes, the source of the overlapping variance between the code variables is unequal sample sizes among groups. When participants have been randomly or representatively sampled and **G** represents a set of naturally occurring categories, the difference in sample sizes will reflect true differences in the proportion of each group in the population. But, when **G** represents a set of experimental manipulations, the difference in sample size will usually be due to nonmeaningful, incidental sources like difficulty (or cost) in mounting each of the treatment conditions or simply the randomization process itself.[11] Typically, sr_i which uses the total variance may be more meaningful for natural categories and pr_i which considers only the unique, nonoverlapping variance will be more meaningful for experiments.

[11]One additional source of unequal ns in missing data due to participant nonresponse or dropout. Missing data may require special analysis procedures (see Chapter 12, Little & Rubin, 1990, and Schafer & Graham, in press).

8.5.5 Statistical Power and the Choice of Contrast Codes

With one categorical IV, many hypotheses that give rise to contrast analysis are comprised of two parts. First, researchers propose a central theoretical distinction that they believe will lead to strong differences between two sets of groups. Second, the researcher proposes that there are only negligible differences among the groups *within* each set involved in the central theoretical distinction.

To illustrate this, let us consider the results of the analyses of two researchers who have different hypotheses about the critical determinant of religion's influence on ATA. Researcher A hypothesizes that minority vs. majority status is the critical determinant. The test of B_1 using the coding system in Table 8.5.1A provides the test of the first part of this hypothesis (see Table 8.5.2). Using either the confidence interval or significance testing approach, we see there is little evidence for this hypothesis. The confidence interval overlaps 0; the test of the null hypothesis is not significant. The second part of the hypothesis is that B_2 and B_3 should show negligible effects.[12] Otherwise stated, we expect no difference between the means of the groups involved in each contrast. Examination of the two tests that jointly comprise the second part of the hypothesis shows that these two confidence intervals do *not* overlap 0, contrary to prediction. Failure of either part of the test of the hypothesis suggests that minority versus majority status is not the critical determinant of religion's influence on ATA. Even if the first part of the hypothesis were supported, statistically significant effects for the second (no difference) part of the hypothesis would suggest that the contrast model is misspecified.

Researcher Z proposes that the central determinant of religion's influence on ATA is the religion's view of the general rights of women. Based on her theorizing, this researcher uses the contrast coding scheme presented in Table 8.5.1B to test her hypothesis (see Table 8.5.3). The test of B_1, representing the central theoretical distinction, shows considerable support for this part of the hypothesis, 95% $CI = 26.49$ to 49.53, $t(32) = 4.07$, $p < .05$. The tests of B_2, $CI = -13.36$ to 33.74, $t(32) = 0.88$, *ns*, and of B_3, $CI = -13.72$ to 32.64, $t(32) = 0.83$, *ns*, show no evidence of differences within the critical distinction of women's general rights. Such an outcome provides support for researcher Z's hypothesis.

Recall that for both of the coding systems, the overall influence of religion on ATA was substantial, $R^2 = .355, F(3, 32) = 5.86, p < .05$. The juxtaposition of the results of the two coding systems in Tables 8.5.2 and 8.5.3 makes clear that the use of contrast codes rather than an omnibus test can either raise or lower statistical power. To simplify the comparison, we can report F tests corresponding to the 1 *df* contrasts since $F = t^2$ for this case. The 1 *df* contrast associated with the majority vs. minority religion contrast, $F(1, 32) = 1.46$, *ns*, explained 8.3% of the overall effect of religion on ATA. This value is calculated as follows:

$$\frac{sr_i^2}{R^2} = \frac{.0296}{.3549} = .083.$$

In comparison, the 1 *df* contrast associated with general attitude toward women's rights, $F(1, 32) = 16.56, p < .001$, was significantly significant. Since sr_i^2 for this contrast is .3339, $.3339/.3539 = 94.1\%$ of the variance in ATA accounted for by religion is accounted for by the 1 *df* general attitude toward women's rights contrast. When the sample sizes in each of the groups are equal, this quantity can also be expressed as the squared correlation (r^2) between the contrast code values and the observed means for each of the groups. Otherwise stated, this

[12] A joint test of the two parts of the within category equivalence may be performed using techniques discussed in Chapter 5. The full regression equation includes all three contrast code variables. The reduced regression equation only includes the central contrast of theoretical interest, here C_1. The gain in prediction (R^2) from the reduced to the full equation is tested. This test corresponds to a joint test of within-category equivalence.

quantity can be taken as a measure of the degree to which the contrast codes mimic the true pattern of means.

In conclusion, researchers having hypotheses about the central distinctions between groups can often use contrast codes to provide sharp tests of the hypotheses. When the observed pattern of means closely follows the predicted pattern, researchers also gain an extra benefit in terms of enhanced statistical power to reject the null hypothesis associated with the central contrasts of interest. However, when hypotheses about the patterning of the means are not sharp, or when the pattern of observed means does not closely match the pattern of predicted means, the use of contrast coding is less likely to detect true mean differences among groups.

8.6 NONSENSE CODING

We have repeatedly stated that $g - 1$ code variables are needed to carry information of membership in one of the g groups. These code variables must be nonredundant: The R of each code variable with the remaining $g - 2$ code variables must *not* equal 1.00 (see the discussion of exact collinearity in Section 10.5.1). These conditions characterize dummy, unweighted effects, weighted effects, and contrast codes. In general, any set of $g - 1$ code variables that meets these conditions will yield exactly the same R^2, F, and regression equations that solve for the group means.

We can explore the limits of these conditions by creating a set of $g - 1$ nonsense codes: nonredundant code variables created in an arbitrary manner. We selected four random numbers between -9 and $+9$ for C_1, then squared and cubed them to produce nonredundant values for C_2 and C_3, respectively. These values are shown in Table 8.6.1A, along with two other arbitrary sets of nonredundant code variables in Parts B and C. Applying the values in Part A to our

TABLE 8.6.1
Illustration of Three Nonsense Coding Systems: Religious Groups

A. Nonsense codes I.

	Code variables		
Religion	C_1	C_2	C_3
Catholic	5	25	125
Protestant	0	0	0
Jewish	−4	16	−64
Other	6	36	256

B. Nonsense codes II.

	Code variables		
Religion	C_1	C_2	C_3
Catholic	1	−7	0
Protestant	−1	−1	0
Jewish	4	.5	24
Other	1	6	−1

C. Nonsense codes III.

	Code variables		
Religion	C_1	C_2	C_3
Catholic	0	0	.71
Protestant	0	0	.04
Jewish	1	4	1
Other	2	108	2

D. Overall results for each nonsense coding scheme.

$R^2_{Y.123} = .3549$ $\quad F = 5.869^*$ ($df = 3, 32$)

$\tilde{R}^2_{Y.123} = .2945$

Note: $^*p < .05$.

running example of religion and ATA, we find that $R^2 = .3549$ and $F = 5.869$, the exact values we have obtained with the previously discussed coding systems. The unstandardized regression equation is

$$\hat{Y} = 10.69C_1 + 2.60C_2 - .6867C_3 + 60.67.$$

When the values of the nonsense codes in Table 8.6.1A are substituted into the equation, it correctly yields the means for each of the four groups. But, here the results diverge sharply from those of all of the coding systems we have considered previously. The values for all individual effects—unstandardized and standardized regression coefficients, simple correlations, partial and semipartial correlations—are gibberish.

The point is that despite the nonsensical character of the coding, any full set of $g - 1$ nonredundant codes carries complete information about group membership. Any result that depends upon the set as a whole will yield correct and meaningful results: $R^2, \tilde{R}^2, F$, and the group means from the regression equation. In contrast, all results from single variables will be nonsensical.

8.7 CODING SCHEMES IN THE CONTEXT OF OTHER INDEPENDENT VARIABLES

8.7.1 Combining Nominal and Continuous Independent Variables

Most analyses involving nominal IVs carried out by means of MRC are likely to include additional nominal or quantitative IVs as well. As one example, a political scientist might wish to study the effect of ethnic group (black, hispanic, white), sex, and family income level on political attitudes. As a second example, an educational researcher might wish to compare the mathematics achievement of public versus private school students, with family income held constant. In these analyses the other nominal or quantitative IVs may be included because their influence on Y is also of interest, as in the first example where ethnic group, sex, and family income are all believed to influence political attitudes. Or, the IV may serve as a control for the effect of one or more IVs of interest as in the second example. The IVs involved in these multiple regression equations may be nominal, quantitative, or combinations of the two.

Consider the first example. The researcher would choose a coding scheme for the nominal variables that best represented her research questions. Suppose the researcher were interested in comparing the political attitudes of the two minority groups with the white group. One way to specify the regression equation would be to use two dummy codes for ethnic group, using white as the reference group, and one dummy code for gender, using male as the reference group. The regression equation would be

(8.7.1) $$\hat{Y} = B_1C_1 + B_2C_2 + B_3Female + B_4Income + B_0 \qquad \text{(full model).}$$

In this coding scheme, C_1 is a code variable that equals 1 if the participant is hispanic and 0 otherwise. C_2 is a code variable that equals 1 if the participant is black and 0 otherwise. Female is a code variable that equals 1 if the participant is female and 0 otherwise (male). Income is the participant's yearly income, and $\hat{Y}$ is the participant's predicted attitude.

Recall from Chapter 3 that the intercept is the predicted value of Y when all IVs are equal to 0. Thus, B_0 represents the predicted attitude of a white, male participant with \$0 income. B_1 represents the mean difference in attitude between the hispanic and white groups and B_2 represents the mean difference in attitude between the black and white groups when sex and income are held constant. B_3 represents the mean difference in attitude between females and males when ethnic group and income are held constant. And B_4 represents the change in attitude for each 1-unit ($1) increase in income when ethnic group and sex are held constant. Each of these effects is a partialed effect. For example, the estimated values of B_1 and B_2 from the regression equation,

$$\hat{Y} = B_1C_1 + B_2C_2 + B_0, \tag{8.7.2}$$

would not in general equal those from Eq. (8.7.1). In Eq. (8.7.2), B_1 is the mean difference in attitude between hispanics and whites and B_2 is the mean difference in attitude between blacks and whites without controlling for sex or income. Only if income and sex are both unrelated to ethnic group will the regression coefficients B_1 and B_2 be equal in the two equations. When the same proportion of females are in each ethnic group, sex and ethnic group will be unrelated.

The interpretation we have presented holds only for the dummy coding scheme described here. As we have seen in this chapter, the interpretation of the individual partial regression coefficients depends on the coding scheme that is chosen. Dummy codes compare each group with a reference group. Unweighted effects codes compare each group with the unweighted mean of the groups, weighted effects codes compare each group with the weighted mean of the groups, and contrast codes compare the unweighted means of the two sets of groups that are involved in the contrast. The inclusion of additional quantitative or nominal IVs changes these from unconditional to conditional comparisons. Otherwise stated, in the comparison the effect of the additional IVs has been partialed out (held constant).

In Chapter 3 we learned about testing partialed effects. The procedures discussed there and earlier in this chapter can be used to construct confidence intervals and to conduct tests of significance for each B_i. In addition, the significance of nominal IVs with three or more groups can be tested using the gain in prediction formula presented in Section 5.5. For example, to test the effect of ethnic group, we would specify the full model, here represented by Eq. (8.7.1). We would then specify a reduced model that included all terms except for the nominal IV of interest (ethnic group), here represented by Eq. (8.7.3):

$$\hat{Y} = B_3Female + B_4Income + B_0 \quad \text{(reduced model)}. \tag{8.7.3}$$

Applying the significance test for the gain in prediction, Eq. (5.5.1), we would have

$$F = \frac{(R^2_{\text{full}} - R^2_{\text{reduced}})/(g-1)}{(1 - R^2_{\text{full}})/(n-k-1)} \quad \text{with} \quad df = (g-1, n-k-1).$$

where R^2_{full} represents the full model (here, Eq. 8.7.1), R^2_{reduced} represents the reduced model (here, Eq. 8.7.3), g is the number of groups comprising the nominal variable (here, $g = 3$), and k is the number of terms in the full model, here 4. As we have seen throughout this chapter, the result of the significance test of the nominal variable will be identical regardless of whether a dummy, unweighted or weighted effects, or contrast coding scheme is used.

8.7.2 Calculating Adjusted Means for Nominal Independent Variables

Beyond reporting significance tests or confidence intervals, it is sometimes desired to report the "adjusted means" that correspond to these comparisons. Adjusted means are the predicted means for each group involved in the regression equation. To illustrate the calculation of adjusted means in a regression equation with two nominal IVs, suppose we estimate a simplified version of Eq. (8.7.1) in which family income is not considered and obtain the following hypothetical results:

$$\hat{Y} = 10C_1 + 5C_2 - 3Female + 50. \tag{8.7.4}$$

To calculate the adjusted means, we would simply substitute in the values of the code variables corresponding to each group. In our example, there are 3 ethnic groups by 2 genders, yielding 6 total adjusted group means. Using the dummy coding scheme described earlier in this section, we would have:

Group	C_1	C_2	Female	Adjusted mean
White Male	0	0	0	50
Hispanic Male	1	0	0	60
Black Male	0	1	0	55
White Female	0	0	1	47
Hispanic Female	1	0	1	57
Black Female	0	1	1	52

The calculation of the adjusted means for two of the groups is:

$$\text{White Male:} \quad \text{adjusted mean} = 10(0) + 5(0) - 3(0) + 50 = 50;$$

$$\text{Hispanic Female:} \quad \text{adjusted mean} = 10(1) + 5(0) - 3(1) + 50 = 57.$$

When there is more than one IV in the regression equation, these adjusted group means will not in general equal the original group means. The adjusted means are estimated based on a particular regression equation, here Eq. (8.7.4). If the regression equation is misspecified, then there may be substantial differences between the adjusted means and the actual means of the groups. In the present example, if there is a true interaction between ethnic group and sex in determining political attitudes, the original and the adjusted means will differ substantially because the interaction is not represented in Eq. (8.7.4). We will return to the topic of interactions between nominal variables in Chapter 9.

8.7.3 Adjusted Means for Combinations of Nominal and Quantitative Independent Variables

When quantitative IVs are included in the regression equation, the approach of substituting in the possible values of the IV to calculate adjusted means is no longer feasible. A simple way to calculate adjusted means is to use a regression equation in which all quantitative variables have been centered by subtracting their respective means. Given our focus on mean differences in this section, we will refer to the other IVs as covariates. If the nominal variable has been dummy coded, the intercept will now reflect the adjusted mean of the reference group. Each of the B_is will reflect the mean difference between the group coded 1 and the reference group, again adjusted for the group differences in the quantitative IVs.

Let us consider a fictitious investigation of background factors and altruism. The researchers have hypothesized that there are influences of population density on altruism, and have drawn samples of residents of a city and the surrounding area outside the city (noncity). As a first analysis, the researchers compared respondents living in the city (City $= 1$) with those not living in the city (City $= 0$), finding that city-dwellers were 18.37 points lower on the DV of altruism (see Table 8.7.1). However, the researchers are concerned that this difference may only reflect differences in neuroticism between city and noncity respondents. They therefore carry out a second regression analysis in which a dummy variable city and Neurot_C are used as IVs, where centered Neuroticism, $\text{Neurot}_C = \text{Neuroticism} - \text{mean(Neuroticism)}$. Neurot_C correlates $-.247$ with altruism and .169 with city residence. The resulting regression equation is

$$\hat{Y} = -17.62\, City - .22\, Neurot_C + 52.88 \tag{8.7.5}$$

CH08EX02

TABLE 8.7.1
Original and Neuroticism-Partialed Scores for Altruism and Residence

	n	$M_{\text{altruism}}(sd)$	$M_{\text{neuroticism}}(sd)$
Total sample	150	46.42 (14.48)	56.29 (9.72)
City residents	55	34.79 (10.92)	58.44 (10.44)
Noncity residents	95	53.16 (11.79)	55.04 (9.10)
Mean difference	($t_{148\,df}$)	18.37 (6.45) $p < .01$	3.40 (2.09) $p < .05$

Partial Data: First 10 Non–City Dwellers and Last 10 City Dwellers

	Altruism	City	Neuroticism	Altruism · Neurot_C	City · Neurot_C
1	69.13	0	61.01	70.86	−0.039
2	56.18	0	73.55	62.53	−0.145
3	65.57	0	41.15	60.00	0.128
4	65.85	0	55.93	65.72	0.003
5	63.08	0	46.41	59.44	0.084
6	50.61	0	58.21	51.31	−0.016
7	63.47	0	54.57	62.84	0.015
8	69.11	0	47.46	65.86	0.075
9	69.87	0	58.94	70.84	−0.022
10	49.05	0	49.05	46.39	0.061
⋮	⋮	⋮	⋮	⋮	⋮
141	32.18	1	61.35	34.04	0.958
142	40.47	1	34.21	32.34	1.186
143	15.21	1	66.48	18.96	0.914
144	27.23	1	52.31	25.76	1.034
145	48.44	1	34.32	40.36	1.185
146	26.52	1	71.87	32.25	0.870
147	39.36	1	55.73	39.15	1.005
148	54.30	1	35.44	46.63	1.176
149	33.77	1	63.78	36.52	0.937
150	40.15	1	65.39	43.50	0.924

In this equation the intercept $B_0 = 52.88$ is the adjusted mean level of altruism for those not living in the city and $-17.62 = B_{\text{City}}$ is the adjusted mean difference between those living in the city minus those not living in the city. The adjusted mean level of altruism for city residents is $52.88 - 17.62 = 35.26$. From Table 8.7.1, we see that the original means on altruism were 34.79 for city residents and 53.16 for noncity residents so that the difference is 18.37. The inclusion of Neurot_C in the regression equation has slightly reduced the difference between city and noncity residents in altruism. The difference in the two results stems from the partialing process that is so central to MRC analysis.

It is useful here to review briefly the meaning of partialing. Partialing means that we are removing any variation that is associated with other IVs in the regression equation (see Sections 3.3 and 3.4). One way to do this is to follow the procedure used to construct added variable plots, presented in Section 4.4.2. Y is regressed on the other covariates (excluding the variable of interest, IV_i). Then IV_i is regressed on the other covariates. In this present two-variable example, altruism and then city would be regressed on Neurot_C. For this sample, the regression equation predicting altruism is

$$\hat{Y}_{\text{Altruism}} = -.368\, Neurot_C + 67.13. \tag{8.7.6}$$

The regression equation predicting City is:

$$\hat{Y}_{\text{City}} = .008\, Neurot_C - .11. \tag{8.7.7}$$

We then save the residuals from each regression equation and add the original means back in for altruism and city, respectively, to return to the original scaling of each of the variables (recalling that residuals always have a mean of zero). The new variables Altruism $\cdot$ Neurot_C and City $\cdot$ Neurot_C represent the values of Altruism and City, respectively, with the linear effect of Neurot_C partialed out.

In Table 8.7.1 we present data that illustrates the results of this procedure. We include the first ten of the noncity cases (1–10) and the last ten of the city cases (141–150) from this sample. Columns 1, 2, and 3 present the original scores on Altruism, City, and Neuroticism (not centered). Column 4 (Altruisum $\cdot$ Neurot_C) presents the Altruism score adjusted for level of Neurot_C and Column 5 (City $\cdot$ Neurot_C) presents the City score, adjusted for Neurot_C. For example, case 1 was a noncity dweller with an Altruism score of 69.13 and a Neuroticism score of 61.01. Using Eq. 8.7.5, the predicted Altruism score is $(-.368)(61.01) + 67.13 = 44.68$. The residual is thus $69.13 - 44.68 = 24.45$, indicating that this respondent was more altruistic than his or her Neurot_C score would have led one to expect. Adding the mean back gives $46.42 + 24.45 = 70.87$, which is the partialed score (within rounding error).

Carrying out the same operation for the city dichotomy using Eq. (8.7.6), we find this value to $= (.0084)(61.01) - .108 = .406$ for case 1. Our "observed" value for this noncity dweller was, of course, 0, so the residual is $-.406$. Adding in the mean for city of .367 (the proportion of the sample coded 1), the partialed score is $.367 - .406 = -.039$.

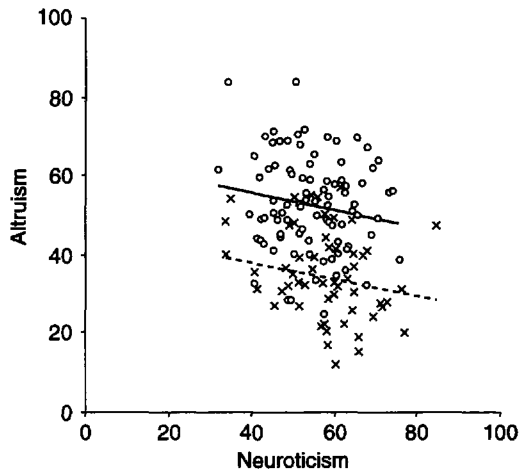
The critical point is what happens when partialed Altruism is regressed on partialed City. Exactly the same estimates for the intercept B_0 (the adjusted mean for the noncity dwellers) and the regression weight B_{City} (representing the adjusted mean difference between noncity and city) are obtained as in the equation in which the unpartialed variables were employed in combination with Neurot_C. Thus, the effect of partialing of Neurot_C, or any other covariate(s), can be seen to be equivalent whether it is accomplished by inclusion in a single prediction equation with IV_i or by prior removal of covariate influence using residuals from the separate equations for Y and IV_i. In addition, the standard errors in the significance tests will also be equal. And the adjusted means have exactly this meaning—means on partialed Y corresponding to scores of 0 and 1 on this partialed dichotomy.

Figure 8.7.1(A) and (B) presents two scatterplots of these same data. Figure 8.7.1(A) presents the original neuroticism covariate on the x axis; Figure 8.7.1 (B) presents the centered Neurot_C covariate on the x axis. Both plots are identical except for the values represented by the scaling of the x axis. In both panels, non–city residents are indicated by open circles and city residents are indicated by $\times$. Two parallel lines have been fit to the data, the solid line for the non–city residents and the dashed line for the city residents. We see that the two lines are always the same distance apart (17.62, the distance between the adjusted city and noncity means) regardless of the value of the IV on the x axis. The specific value of the adjusted means is the value of $\hat{Y}$ for each group estimated at the mean of value of neuroticism (=56.29) or equivalently at the 0 value of Neurot_C. Substituting into Eq. 8.7.5, the adjusted means are

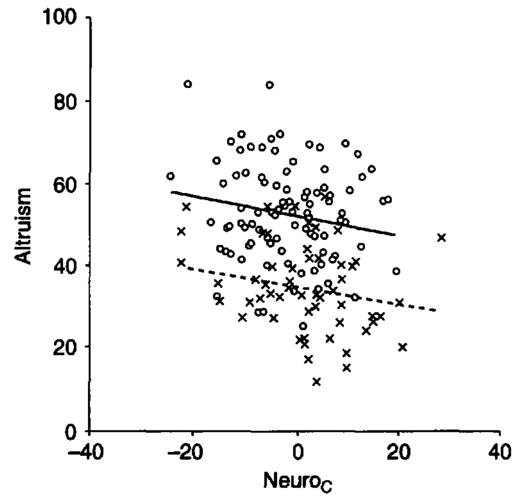
$$\hat{Y}_{\text{City}} = -(17.62)(1) - (.22)(0.0) + 52.88 = \text{adjusted } M_{\text{City}}$$

$$\hat{Y}_{\text{Noncity}} = -(17.62)(0) - (.22)(0.0) + 52.88 = \text{adjusted } M_{\text{Noncity}}$$

(A) Uncentered Neuroticism.



(B) Centered Neuroticism ($Neurot_C$).



Note: City residents are designated by ×; noncity residents are designated by o. The dashed line is the regression line for the city group; the solid line is the regression line for the noncity group. The regression equation including neuroticism is $\hat{Y} = -17.62\ City - 0.22\ Neurot_C + 65.24$. The regression equation including Neurot_C is $\hat{Y} = -17.62\ City - 0.22\ Neurot_C + 52.88$.

FIGURE 8.7.1 Scatterplot of altruism versus neuroticism.

Figure 8.7.1(A) presents a plot of the data in which neuroticism has not been centered. When quantitative IVs have not been centered, the adjusted means in the equations that include them need to take the means of the other IVs into account by adding the products of their B coefficients times their means to the estimate. For our running example of altruism, the equation with the uncentered neuroticism is

$$\hat{Y} = -17.62\ \mathrm{City} - .22\ \mathrm{Neuroticism} + 65.24$$

for which we see that only the intercept B_0 is changed. As is illustrated in Fig. 8.7.1(A), the intercept now represents the predicted value of a noncity resident who has a score of 0 on neuroticism, a score that falls far outside the range of observed values of neuroticism in the sample. Estimating the adjusted mean now requires that the mean value of neuroticism (= 56.29) be entered into the equation:

$$\hat{Y}_{\text{City}} = (-17.62)(1) + (-.22)(56.29) + 65.24 = -17.62 - 12.39 + 65.24 = 35.26$$

$$\hat{Y}_{\text{Noncity}} = (-17.62)(0) + (-.22)(56.29) + 65.24 = 0 - 12.39 + 65.24 = 52.88.$$

8.7.4 Adjusted Means for More Than Two Groups and Alternative Coding Methods

The two principles just articulated generalize to any number of groups and any of the methods of coding group membership (the nominal or categorical variable) described in this chapter:

1. When covariates have been centered, their inclusion in the equation will produce adjusted means by the same methods used to produce unadjusted group means in their absence.
2. When covariates have not been centered one must add (or subtract) the products of their B coefficients in the equation times their full sample means—$\Sigma B_{\text{cov}_i} M_{\text{cov}_i}$—to the intercept in order for the methods described earlier for retrieving Y means to work.

To illustrate these principles, let us develop our running example in more detail. Suppose we had actually sampled altruism in three subpopulations: city, small town, and rural areas. (The last two subpopulations were combined into a single noncity group for our previous analysis). Our theory leads us to hypothesize that those living in small towns are more altruistic than those living in either rural areas or cities. In addition, previous research has suggested that neuroticism and socioeconomic status (SES) may also be related to altruism. Consequently, we are also worried that our findings may be contaminated by these two factors.

Because our theory suggests that the small town residents will be different from both city and rural respondents, we have used dummy codes in which town is the reference group. C_1 represents city residence and C_2 represents rural residence. Table 8.7.2A presents the mean Altruism, Neuroticism, and SES by residential area. Table 8.7.2B presents the correlation matrix including the two dummy codes. Of importance, some of the relationships between the dummy variables, (Rural, City), SES, and Neuroticism are statistically significant, suggesting the possibility that any differences between residence areas may be contaminated (in part) by the effects of Neuroticism and SES.

For pedagogical purposes, we estimate two regression equations. Equation (8.7.8) does not include the covariates:

$$\text{(8.7.8)} \qquad \text{Predicted } \mathit{Altruism} = -24.94\ \mathit{City} - 11.15\ \mathit{Rural} + 59.73.$$

Equation (8.7.9) now includes the two centered covariates, Neurot_C and SES_C, to partial out any effects of these two variables.

$$\text{(8.7.9)} \qquad \text{Predicted } \mathit{Altruism} = -24.97\ \mathit{City} - 10.21\ \mathit{Rural} - 0.190 \mathit{Neurot}_\text{C} + 0.196\ \mathit{SES}_\text{C} + 59.39.$$

Comparing the results of Eq. (8.7.8) with Eq. (8.7.9), we note that the difference in mean Altruism between the City dwellers and those from small towns shows hardly any net effect of the partialing of the covariate (B_{city} – 24.94 versus –24.97). However, the magnitude of the difference in mean Altruism between rural and small towns decreases by about 1 unit

TABLE 8.7.2
Altruism and Three Types of Residential Area

A. Mean scores by area.

	Altruism	Neuroticism	SES
Rural ($n = 56$)	48.58	56.11	44.06
Small town ($n = 39$)	59.73	53.51	46.31
City ($n = 55$)	34.79	58.44	51.21
Total ($n = 150$)	46.42 ($sd = 14.48$)	56.29 ($sd = 9.72$)	47.27 ($sd = 10.72$)

B. Correlation matrix including dummy-variable coded area.

	Altruism	City	Rural	Neuroticism	SES
Altruism	1.0				
City	−.613	1.0			
Rural	.115	−.587	1.0		
Neuroticism	−.247	.169	−.014	1.0	
SES	−0.25	.281	−.231	.118	1.0

C. Regression equation without covariates.

$$\textit{Predicted Altruism} = -29.94\textit{City} - 11.15\textit{Rural} + 59.73$$

D. Regression equation with centered covariates.

$$\textit{Predicted Altruism} = -24.97\textit{City} - 10.21\textit{Rural} - 0.190\textit{Neurot}_C + 0.196\textit{SES}_C + 59.39$$

E. Regression equation with uncentered covariates.

$$\textit{Predicted Altrusim} = -24.97\textit{City} - 10.21\textit{Rural} - 0.190\textit{Neurotism} + 0.196\textit{SES} + 60.83$$

Note: City residents are coded $C_1 = 1, C_2 = 0$. Small town residents are coded $C_1 = 0, C_2 = 0$ (reference group). Rural residents are coded $C_1 = 0, C_2 = 1$. C_1 is referred to as *City* and C_2 is referred to as *Rural* in the text.

(B_{rural}: -11.15 versus -10.21). This change is attributable to the influence of the Neuroticism and SES covariates. In general, the effect of including covariates on the group difference is indeterminate. The group difference may be the same, larger, smaller, or even reversed in sign, depending on the structure of the data.

To calculate the adjusted mean Altruism for the three groups in our example, we substitute the appropriate values into Eq. (8.7.9):

$$\begin{aligned}
\text{Adjusted } M_{Y(\text{Town})} &= (B_{\text{City}})(0) + (B_{\text{Rural}})(0) + (B_{\text{Neurot}_C})(M_{\text{Neurot}_C}) + (B_{\text{SES}_C})(M_{\text{SES}_C}) + B_0 \\
&= 0 + 0 + 0 + 0 + B_0 = B_0 = 59.39; \\
\text{Adjusted } M_{Y(\text{City})} &= (B_{\text{City}})(1) + (B_{\text{Rural}})(0) + (B_{\text{Neurot}_C})(M_{\text{Neurot}_C}) + (B_{\text{SES}_C})(M_{\text{SES}_C}) + B_0 \\
&= -24.97 + 0 + 0 + 0 + 59.39 = 34.42; \\
\text{Adjusted } M_{Y(\text{Rural})} &= (B_{\text{City}})(0) + (B_{\text{Rural}})(1) + (B_{\text{Neurot}_C})(M_{\text{Neurot}_C}) + (B_{\text{SES}_C})(M_{\text{SES}_C}) + B_0 \\
&= 0 - 10.21 + 0 + 0 + 59.39 = 49.18.
\end{aligned}$$

If we had not centered the covariates, we would need to add in the product of their regression weights times their respective means, so

$$\begin{aligned}
\text{Adjusted } M_{Y(\text{Town})} &= (B_{\text{City}})(0) + (B_{\text{Rural}})(0) + (B_{\text{Neuroticism}})(M_{\text{Neuroticism}}) + (B_{\text{SES}})(M_{\text{SES}}) + B_0 \\
&= 0 + 0 + (-0.190)(56.29) + (0.196)(47.27) + 60.83 = 59.39
\end{aligned}$$

$$\begin{aligned}\text{Adjusted } M_{Y(\text{City})} &= (B_{\text{City}})(1) + (B_{\text{Rural}})(0) + (B_{\text{Neuroticism}})(M_{\text{Neuroticism}}) + (B_{\text{SES}})(M_{\text{SES}}) + B_0 \\ &= -24.97(1) + 0 + (-0.190)(56.29) + (0.196)(47.27) + 60.83 = 34.42\end{aligned}$$

and

$$\begin{aligned}\text{Adjusted } M_{Y(\text{Rural})} &= (B_{\text{City}})(0) + (B_{\text{Rural}})(1) + (B_{\text{Neuroticism}})(M_{\text{Neuroticism}}) + (B_{\text{SES}})(M_{\text{SES}}) + B_0 \\ &= 0 - 10.21(1) + (-0.190)(56.29) + (0.196)(47.27) + 60.83 = 49.18\end{aligned}$$

as before.

CH08EX03

The analyst may use dummy codes, unweighted effect codes, weighted effect codes, or contrast codes to calculate adjusted means. When the covariates have been centered, the code values for the particular coding system are simply substituted into the corresponding regression equation. When the covariates have not been centered, the sum of the products of the covariates regression weights times their full sample means is added to the intercept term.[13]

Once again, we caution the reader that the adjusted means are estimated based on a specific regression equation, here Eq. (8.7.9). If the regression equation has been misspecified, then there may be substantial differences between the adjusted means and the true means in the population. Of particular importance, Eq. (8.7.9) assumes that possible interactions between the code variables representing the group and the covariates are 0. Methods of investigating interactions between nominal and quantitative variables are presented in Chapter 9.

8.7.5 Multiple Regression/Correlation with Nominal Independent Variables and the Analysis of Covariance

Analysis of Covariance (ANCOVA) is an analysis strategy typically applied to assess the impact of one (or more) group factors (e.g., treatment groups; gender) while statistically controlling for other IVs (called the covariates). ANCOVA is often treated as an extension of ANOVA that includes additional variables to be controlled. ANCOVA is treated in statistical packages (SAS, SPSS, SYSTAT) within the ANOVA framework. Readers familiar with ANCOVA may wonder about its relationship to MRC. Paralleling our observations in Section 8.2.7 about one-way ANOVA, they are identical. The F test for the Group effect in ANCOVA will be identical to the gain in prediction tests associated with the Group in MRC. The gain in R^2 associated with the Group ($R^2_{\text{full}} - R^2_{\text{reduced}}$) will equal η^2 from the ANCOVA. MRC analysis has the advantage relative to standard ANCOVA of allowing for the choice of a coding scheme that optimally represents the researchers' questions of interest.

The juxtaposition of MRC and ANCOVA also highlights several important issues in the interpretation of group differences in these analyses. As we have repeatedly emphasized, the use of the MRC approach assumes that the regression model has been properly specified. In Chapter 4, we considered five assumptions that we can apply in the present context.

1. The form of the relationship between each of the IVs and the DV is properly specified. In this chapter we have focused only on linear relationships. If there is a nonlinear relationship between one or more of the covariates and the DV, the model will be misspecified. Or, if there is a group $\times$ covariate interaction such that the regression of the DV on the covariate differs in each of the groups, the model will be misspecified. Regression models that include interactions between nominal IVs and between nominal and continuous IVs are considered in Chapter 9.

2. All relevant IVs are included in the model. MRC only adjusts for those covariates that are included in the regression equation.[14] Any other unmeasured covariates that (a) are

[13]The reader may wish to practice obtaining the desired contrasts with these data.

[14]In experiments, covariates are measured prior to treatment and mediators are measured following treatment.

associated with group status and (b) affect the DV lead to bias in the estimates of the group effect (see Section 4.5.2). One special exception occurs in the case of mediational analysis (see Section 12.2.1; Baron & Kenny, 1986) in which the group variable is presumed to cause changes in the mediator, which, in turn, causes changes in the DV (indirect effect).

3. All IVs are measured with perfect reliability. In the present context, group status (e.g., treatment versus control; city versus rural) is likely to be measured with near perfect reliability. With a single covariate, the effect of unreliability is to lead to too little adjustment of the mean difference for the covariate. With multiple unreliable covariates, the direction of bias is uncertain, although it will typically be toward underadjustment of mean differences.

4. The variance of the residuals within each group is equal (homoscedasticity). When large differences in the residual variances are obtained, the actual alpha level of significance tests may be too high or too low relative to the stated alpha level (typically, $\alpha = .05$, see Section 4.4.6).

5. The residuals from the regression equation should follow a normal distribution. Although this assumption is not critical in moderate or large n studies, non-normally distributed residuals may be a symptom of model misspecification.

In the context of MRC, these five are standard assumptions that should always be checked. In the context of ANCOVA, only some of these assumptions have received emphasis. In large part, this difference in emphasis stems from the application of ANCOVA to randomized experimental designs. In the context of randomized experiments, assumption 2 is automatically met because the treatment group will on average not be correlated with any of the covariates. Violation of assumption 3 (reliable covariates) can lead only to underestimation of the magnitude of the treatment effect (low power statistical test) but not to spurious results indicating a treatment effect exists when in fact it does not. Finally, violation of assumption 4 (equality of residual variances in each treatment group) does not typically occur—in practice relatively few treatments have large effects on the variance of the DV. Consequently, the emphasis in the traditional ANCOVA literature has been on the issues of possible nonlinear covariate-DV relationships and possible treatment × covariate interactions (assumption 1).

However, when MRC or ANCOVA is used to compare groups in nonexperimental contexts, it is important that each of the five assumptions be examined. Assumptions 1–4 are critical to inferring a treatment effect, and violation of assumption 5 can provide clues about model misspecification. Whether the groups are naturally existing groups (males and females; city, small town, rural residents) or are treatment groups in a quasi-experimental design (e.g., community 1 receives treatment; community 2 receives control), it cannot be definitely presumed that the groups are initially equivalent. Reichardt (1979); Shadish, Cook, and Campbell (2002); and West, Biesanz, and Pitts (2000) present full discussions of the conceptual and statistical issues associated with randomized and nonrandomized designs.

8.8 SUMMARY

A nominal independent variable G that partitions observations into g groups can be represented as $g - 1$ independent variables by various coding systems. All coding systems yield identical R^2, adjusted R^2, F for R^2, and, via the regression equations, group means on Y. All of these results are identical with those of analysis of variance. In contrast, the coding systems differ sharply in the meaning of the results for the individual code variables, C_i. Researchers should choose the coding system that provides the best answers to the specific research questions they are posing. Remaining questions that are not addressed by the coding system can typically be answered by using a second coding system that provides a direct answer to those questions. The various alternative coding methods facilitate the interpretation of the results for a

TABLE 8.8.1
Comparison of Coding Systems

Coding	r^2_{Yi}	B_0	B_i	sr^2_i	pr^2_i
Dummy	PV[a] due to i vs. non-i dichotomy	M_b (mean of reference group)	$M_i - M_b$ (difference between means of reference and ith groups)	PV due to i vs. non-i dichotomy	PV due to i vs. non-i dichotomy, excluding other effects
Unweighted effects	Ambiguous[b]	M_U (unweighted mean)	$M_i - M_u$ (difference between mean of ith group and unweighted mean)	PV due to i's effect	PV due to i's effect, excluding other effects
Weighted effects	Ambiguous	M_W (weighted mean)	$M_i - M_W$ (difference between mean of ith group and weighted mean)	—	—
Contrast	Ambiguous	M_U (unweighted mean)	varies	PV due to the i contrast	PV due to the i contrast, excluding other contrasts
Nonsense[c]	—	—	—	—	—

Note: [a]Proportion of variance. [b]Ambiguous interpretation unless sample sizes are equal in all groups. [c]All individual results are meaningless. For ease of presentation, we have assumed that the ith code variables also refers to the ith group.

single C_i for different frames of reference or desired comparisons. Table 8.8.1 provides a concise summary of the meaning of regression coefficients and simple, semipartial, and partial correlation coefficients for the different coding methods.

1. ***Dummy-variable coding.*** One group is selected as the reference group. The intercept is the mean of the reference group, and each of the unstandardized regression coefficients is the difference between the mean of one of the groups and the mean of the reference group. This coding system is particularly appropriate for research in which one group is a control group and the others are to be compared with it (Section 8.2).

2. ***Unweighted effects coding.*** The reference point here is the unweighted mean of all of the group means; this value is the intercept. The unstandardized regression coefficients represent the difference between each group's mean and the unweighted mean. This coding system is useful with experimental designs (Section 8.3).

3. ***Weighted effects coding.*** The intercept is the weighted mean of the individual group means. The unstandardized regression coefficients represent the difference between each group's mean and the weighted mean. This coding system is particularly useful when the groups represent natural categories and a large random sample has been taken from the population (Section 8.4).

4. ***Contrast coding.*** The $g - 1$ variables are constructed so as to provide a set of independent comparisons between means or unweighted means of means, selected in accordance with the researchers' specific research hypotheses. The intercept is the unweighted mean of the individual group means. The unstandardized regression coefficient provides the difference between the unweighted mean of the means of the groups in set u and the unweighted mean of the groups in set v, where u and v and the two sets of groups involved in the contrast. Contrast codes are particularly important when researchers have a priori hypotheses about the specific pattern of differences about the group means (Section 8.5).

5. ***Nonsense coding.*** Here the coding values are arbitrary and individual results are meaningless, although those for the set as a whole are identical to those of the other coding systems. Nonsense codes were included because they illustrate how any set of $g - 1$ nonredundant code variables fully captures the information in a nominal IV (Section 8.6).

Finally, we noted that coding systems for nominal independent variables provide the foundation for understanding the meaning of more complex regression models including both categorical and continuous scale variables. Group means on the dependent variables adjusted for covariates are easily determined, especially when covariates have been centered. Any of the coding methods may be employed in these models combining groups and covariates. However, as always, the researcher is cautioned about the consequences of inadequate reliability in the covariates (Section 8.7).