12 DUMMY PREDICTOR VARIABLES IN MULTIPLE REGRESSION

12.1 ♦ Research Situations Where Dummy Predictor Variables Can Be Used

Previous examples of regression analysis have used scores on quantitative X variables to predict scores on a quantitative Y variable. However, it is possible to include group membership or categorical predictor variables as predictors in regression analysis. This can be done by creating dummy (dichotomous) predictor variables to represent information about group membership. A dummy or dichotomous predictor variable provides yes/no information for questions about group membership. For example, a simple dummy variable to represent gender corresponds to the following question: Is the participant female (0) or male (1)? Gender is an example of a two-group categorical variable that can be represented by a single dummy variable.

When we have more than two groups, we can use a set of dummy variables to provide information about group membership. For example, suppose that a study includes members of $k = 3$ political party groups. The categorical variable political party has the following scores: 1 = Democrat, 2 = Republican, and 3 = Independent. We might want to find out whether mean scores on a quantitative measure of political conservatism (Y) differ across these three groups. One way to answer this question is to perform a one-way analysis of variance (ANOVA) that compares mean conservatism (Y) across the three political party groups. In this chapter, we will see that we can also use regression analysis to evaluate how political party membership is related to scores on political conservatism.

However, we should not set up a regression to predict scores on conservatism from the *multiple-group* categorical variable political party, with party membership coded 1 = Democrat, 2 = Republican, and 3 = Independent. Multiple-group categorical variables usually do not work well as predictors in regression, because scores on a quantitative outcome variable, such as "conservatism," will not necessarily increase linearly with the score on the categorical variable that provides information about political party. The score values that represent political party membership may not be rank ordered in a way that is **monotonically** associated with changes in conservatism; as we move from Group 1 = Democrat to Group 2 = Republican, scores on conservatism may increase, but as we

move from Group 2 = Republican to Group 3 = Independent, conservatism may decrease. Even if the scores that represent political party membership are rank ordered in a way that is monotonically associated with level of conservatism, the amount of change in conservatism between Groups 1 and 2 may not be equal to the amount of change in conservatism between Groups 2 and 3. In other words, scores on a multiple-group categorical predictor variable (such as political party coded 1 = Democrat, 2 = Republican, and 3 = Independent) are not necessarily *linearly* related to scores on quantitative variables.

If we want to use the categorical variable political party to predict scores on a quantitative variable such as conservatism, we need to represent the information about political party membership in a different way. Instead of using one categorical predictor variable with codes 1 = Democrat, 2 = Republican, and 3 = Independent, we can create two dummy or dichotomous predictor variables to represent information about political party membership, and we can then use these two dummy variables as predictors of conservatism scores in a regression. Political party membership can be assessed by creating dummy variables (denoted by D_1 and D_2) that correspond to two yes/no questions. In this example, the first dummy variable D_1 corresponds to the following question: Is the participant a member of the Democratic Party? Coded 1 = yes, 0 = no. The second dummy variable D_2 corresponds to the following question: Is the participant a member of the Republican Party? Coded 1 = yes, 0 = no. We assume that group memberships for individuals are mutually exclusive and exhaustive—that is, each case belongs to only one of the three groups, and every case belongs to one of the three groups identified by the categorical variable. When these conditions are met, a third dummy variable is not needed to identify the members of the third group because, for Independents, the answers to the first two questions that correspond to the dummy variables D_1 and D_2 would be no. In general, when we have k groups or categories, a set of $k - 1$ dummy variables is sufficient to provide complete information about group membership. Once we have represented political party group membership by creating scores on two dummy variables, we can set up a regression to predict scores on conservatism (Y) from the scores on the two dummy predictor variables D_1 and D_2:

$$Y' = b_0 + b_1 D_1 + b_2 D_2. \tag{12.1}$$

In this chapter, we will see that the information about the association between group membership (represented by D_1 and D_2) and scores on the quantitative Y variable that can be obtained from the regression analysis in Equation 12.1 is equivalent to the information that can be obtained from a one-way ANOVA that compares means on Y across groups or categories. It is acceptable to use *dichotomous* predictor variables in regression and correlation analysis. This works because (as discussed in Chapter 8) a dichotomous categorical variable has only two possible score values, and the only possible relationship between scores on a dichotomous predictor variable and a quantitative outcome variable is a linear one. That is, as you move from a score of 0 = female on gender to a score of 1 = male on gender, mean height or mean annual salary may increase, decrease, or stay the same; any change that can be observed across just two groups can be represented as linear. Similarly, if we represent political party membership using two dummy variables, each dummy variable represents a contrast between the means of two groups; for example, the D_1

dummy variable can represent the difference in mean conservatism between Democrats and Independents, and the D_2 dummy variable can represent the mean difference in conservatism between Republicans and Independents.

This chapter uses empirical examples to demonstrate that regression analyses that use dummy predictor variables (similar to Equation 12.1) provide information that is equivalent to the results of more familiar analyses for comparison of group means (such as ANOVA). There are several reasons why it is useful to consider dummy predictor variables as predictors in regression analysis. First, the use of dummy variables as predictors in regression provides a simple demonstration of the fundamental equivalence between ANOVA and multiple regression; ANOVA and regression are both special cases of a more general analysis called the general linear model (GLM). Second, researchers often want to include group membership variables (such as gender) along with other predictors in a multiple regression. Therefore, it is useful to examine examples of regression that include dummy variables along with quantitative predictors.

The computational procedures for regression remain the same when we include one or more dummy predictor variables. The most striking difference between dummy variables and quantitative variables is that the scores on dummy variables usually have small integer values (such as 1, 0, and −1). The use of small integers as codes simplifies the interpretation of the regression coefficients associated with dummy variables. When dummy variables are used as predictors in a multiple regression, the b raw score slope coefficients provide information about differences between group means. The specific group means that are compared differ depending on the method of coding that is used for dummy variables, as explained in the following sections. Except for this difference in the interpretation of regression coefficients, regression analysis remains essentially the same when dummy predictor variables are included.

12.2 ♦ Empirical Example

The hypothetical data for this example are provided by a study of predictors of annual salary in dollars for a group of $N = 50$ college faculty members; the complete data appear in Table 12.1. Predictor variables include the following: gender, coded 0 = female and 1 = male; years of job experience; college, coded 1 = Liberal Arts, 2 = Sciences, 3 = Business; and an overall merit evaluation. Additional columns in the SPSS data worksheet in Figure 12.1, such as D_1, D_2, E_1, and E_2, represent alternative ways of coding group membership, which are discussed later in this chapter. All subsequent analyses in this chapter are based on the data in Table 12.1.

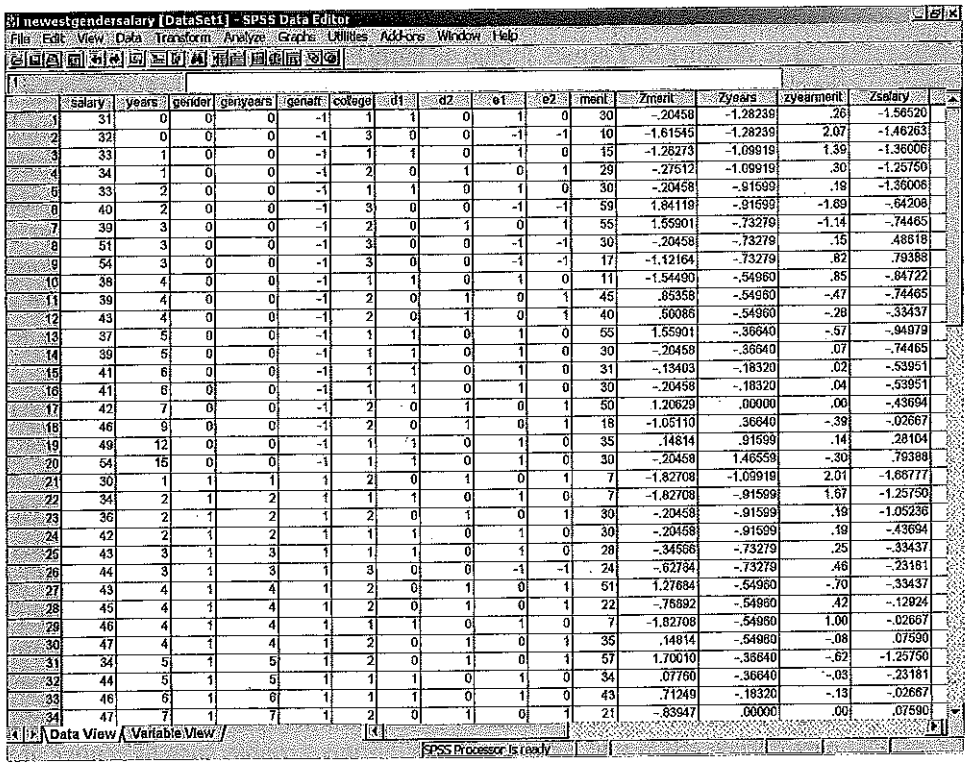
The first research question that can be asked using these data is whether there is a significant difference in mean salary between males and females (ignoring all other predictor variables). This question could be addressed by conducting an independent samples t test to compare male and female means on salary. In this chapter, a one-way ANOVA is performed to compare mean salary for male versus female faculty; then, salary is predicted from gender by doing a regression analysis to predict salary scores from a dummy variable that represents gender. The examples presented in this chapter demonstrate that ANOVA and regression analysis provide equivalent information about gender differences in mean salary. Examples or demonstrations such as the ones presented in this chapter do

Table 12.1 ◆ Hypothetical Data for Salary and Predictors of Salary for $N = 50$ College Faculty

Salary	*Years*	*Gender*	*College*
31	0	0	1
32	0	0	3
33	1	0	1
34	1	0	2
33	2	0	1
40	2	0	3
39	3	0	2
51	3	0	3
54	3	0	3
38	4	0	1
39	4	0	2
43	4	0	2
37	5	0	1
39	5	0	1
41	6	0	1
41	6	0	1
42	7	0	2
46	9	0	2
49	12	0	1
54	15	0	1
30	1	1	2
34	2	1	1
42	2	1	1
36	2	1	2
43	3	1	1
44	3	1	3
46	4	1	1
43	4	1	2
45	4	1	2
47	4	1	2
44	5	1	1
34	5	1	2
46	6	1	1
51	7	1	1
47	7	1	2
50	8	1	1
51	8	1	1
51	9	1	2
51	9	1	3
54	10	1	3
63	10	1	3
56	12	1	3
58	13	1	3
58	14	1	1
59	14	1	1
58	14	1	2
66	17	1	1
67	19	1	2
59	20	1	2
64	22	1	3

NOTES: Salary, annual salary in thousands of dollars; years, years of job experience; gender, dummy-coded gender (0 = female and 1 = male); college, membership coded 1 = Liberal Arts, 2 = Sciences, 3 = Business.

Figure 12.1 ♦ SPSS Data Worksheet for Hypothetical Faculty Salary Study



newestgendersalary [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	salary	years	gender	genyears	genaff	college	d1	d2	e1	e2	merit	Zmerit	Zyears	zyearmerit	Zsalary
1	31	0	0	0	-1	1	1	0	1	0	30	-.20458	-1.28239	.26	-1.56520
2	32	0	0	0	-1	3	0	0	-1	-1	10	-1.61545	-1.28239	2.07	-1.46263
3	33	1	0	0	-1	1	1	0	1	0	15	-1.26273	-1.09919	1.39	-1.36006
4	34	1	0	0	-1	2	0	1	0	1	29	-.27512	-1.09919	.30	-1.25750
5	33	2	0	0	-1	1	1	0	1	0	30	-.20458	-.91599	.19	-1.36006
6	40	2	0	0	-1	3	0	0	-1	-1	59	1.84119	-.91599	-1.69	-.64208
7	39	3	0	0	-1	2	0	1	0	1	55	1.55901	-.73279	-1.14	-.74465
8	51	3	0	0	-1	3	0	0	-1	-1	30	-.20458	-.73279	.15	.48618
9	54	3	0	0	-1	3	0	0	-1	-1	17	-1.12164	-.73279	.82	.79388
10	38	4	0	0	-1	1	1	0	1	0	11	-1.54490	-.54960	.85	-.84722
11	39	4	0	0	-1	2	0	1	0	1	45	.85358	-.54960	-.47	-.74465
12	43	4	0	0	-1	2	0	1	0	1	40	.50086	-.54960	-.28	-.33437
13	37	5	0	0	-1	1	1	0	1	0	55	1.55901	-.36640	-.57	-.94979
14	39	5	0	0	-1	1	1	0	1	0	30	-.20458	-.36640	.07	-.74465
15	41	6	0	0	-1	1	1	0	1	0	31	-.13403	-.18320	.02	-.53951
16	41	6	0	0	-1	1	1	0	1	0	30	-.20458	-.18320	.04	-.53951
17	42	7	0	0	-1	2	0	1	0	1	50	1.20629	.00000	.00	-.43694
18	46	9	0	0	-1	2	0	1	0	1	18	-1.05110	.36640	-.39	-.02667
19	49	12	0	0	-1	1	1	0	1	0	35	.14814	.91599	.14	.28104
20	54	15	0	0	-1	1	1	0	1	0	30	-.20458	1.46559	-.30	.79388
21	30	1	1	1	1	2	0	1	0	1	7	-1.82708	-1.09919	2.01	-1.66777
22	34	2	1	2	1	1	1	0	1	0	7	-1.82708	-.91599	1.67	-1.25750
23	36	2	1	2	1	2	0	1	0	1	30	-.20458	-.91599	.19	-1.05236
24	42	2	1	2	1	1	1	0	1	0	30	-.20458	-.91599	.19	-.43694
25	43	3	1	3	1	1	1	0	1	0	28	-.34566	-.73279	.25	-.33437
26	44	3	1	3	1	3	0	0	-1	-1	24	-.62784	-.73279	.46	-.23181
27	43	4	1	4	1	2	0	1	0	1	51	1.27684	-.54960	-.70	-.33437
28	45	4	1	4	1	2	0	1	0	1	22	-.75892	-.54960	.42	-.12924
29	46	4	1	4	1	1	1	0	1	0	7	-1.82708	-.54960	1.00	-.02667
30	47	4	1	4	1	2	0	1	0	1	35	.14814	-.54960	-.08	.07590
31	34	5	1	5	1	2	0	1	0	1	57	1.70010	-.36640	-.62	-1.25750
32	44	5	1	5	1	1	1	0	1	0	34	.07760	-.36640	-.03	-.23181
33	46	6	1	6	1	1	1	0	1	0	43	.71249	-.18320	-.13	-.02667
34	47	7	1	7	1	2	0	1	0	1	21	-.83947	.00000	.00	.07590

Data View | Variable View

SPSS Processor is ready

not constitute formal mathematical proof. Mathematical statistics textbooks provide formal mathematical proof of the equivalence of ANOVA and regression analysis.

The second question that will be addressed is whether there are significant differences in salary across the three colleges. This question will be addressed by doing a one-way ANOVA to compare mean salary across the three college groups and by using dummy variables that represent college group membership as predictors in a regression. This example demonstrates that membership in k groups can be represented by a set of $(k-1)$ dummy variables.

12.3 ♦ Screening for Violations of Assumptions

When we use one or more dummy variables as predictors in regression, the assumptions are essentially the same as for any other regression analysis (and the assumptions for a one-way ANOVA). As in other applications of ANOVA and regression, scores on the outcome variable Y should be quantitative and approximately normally distributed. If the Y outcome variable is categorical, logistic regression analysis should be used instead of linear regression; a brief introduction to binary logistic regression is presented in Chapter 21. Potential violations of the assumption of an approximately normal distribution shape for the Y outcome variable can be assessed by examining a histogram of scores on Y; the shape of this distribution should be reasonably close to normal. As described in Chapter 4, if there are extreme outliers or if the distribution shape is drastically different from normal, it

may be appropriate to drop a few extreme scores, modify the value of a few extreme scores, or, by using a data transformation such as the logarithm of *Y*, make the distribution of *Y* more nearly normal.

The variance of *Y* scores should be fairly homogeneous across groups—that is, across levels of the dummy variables. The *F* tests used in ANOVA are fairly robust to violations of this assumption, unless the numbers of scores in the groups are small and/or unequal. When comparisons of means on *Y* across multiple groups are made using the SPSS *t* test procedure, one-way ANOVA, or GLM, the Levene test can be requested to assess whether the homogeneity of variance is seriously violated. SPSS multiple regression does not provide a formal test of the assumption of homogeneity of variance across groups. It is helpful to examine a graph of the distribution of scores within each group (such as a boxplot) to assess visually whether the scores of the *Y* outcome variable appear to have fairly homogeneous variances across levels of each *X* dummy predictor variable.

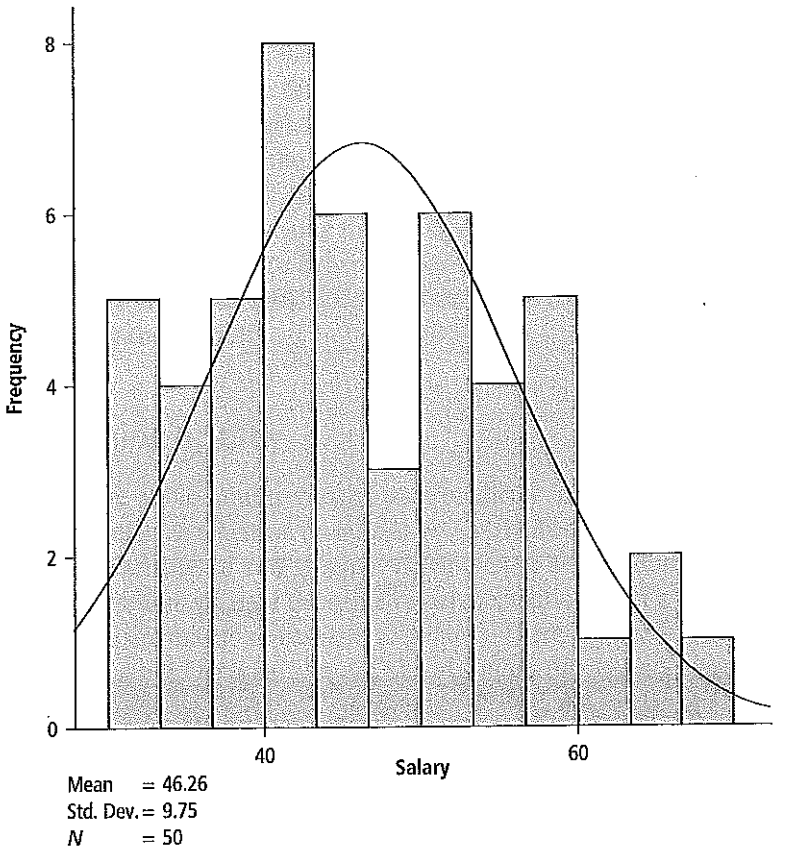
The issue of group size should also be considered in preliminary data screening (i.e., How many people are there in each of the groups represented by codes on the dummy variables?). For optimum statistical power and greater robustness to violations of assumptions, such as the homogeneity of variance assumption, it is preferred that there are equal numbers of scores in each group.[1] The minimum number of scores within each group should be large enough to provide a reasonably accurate estimate of group means. For any groups that include fewer than 10 or 20 scores, estimates of the group means may have confidence intervals that are quite wide. The guidelines about the minimum number of scores per group from Chapter 5 (on the independent samples *t* test) and Chapter 6 (between-subjects [between-S] one-way ANOVA) can be used to judge whether the numbers in each group that correspond to a dummy variable, such as gender, are sufficiently large to yield reasonable statistical power and reasonable robustness against violations of assumptions.

For the hypothetical faculty salary data in Table 12.1, the numbers of cases within the groups are (barely) adequate. There were 20 female and 30 male faculty in the sample, of whom 22 were Liberal Arts faculty, 17 were Sciences faculty, and 11 were Business faculty. Larger group sizes are desirable in real-world applications of dummy variable analysis; relatively small numbers of cases were used in this example to make it easy for students to verify computations, such as group means, by hand.

All the SPSS procedures that are used in this chapter, including boxplot, scatter plot, Pearson correlation, one-way ANOVA, and linear regression, have been introduced and discussed in more detail in earlier chapters. Only the output from these procedures appears in this chapter. For a review of the menu selections and the SPSS dialog windows for any of these procedures, refer to the chapter in which each analysis was first introduced.

To assess possible violations of assumptions, the following preliminary data screening was performed. A histogram was set up to assess whether scores on the quantitative outcome variable salary were reasonably normally distributed; this histogram appears in Figure 12.2. Although the distribution of salary values was multimodal, the salary scores did not show a distribution shape that was drastically different from a normal distribution. In real-life research situations, salary distributions are often positively skewed, such that there is a long tail at the upper end of the distribution (because there is usually no fixed upper limit for salary) and a truncated tail at the lower end of the distribution

Figure 12.2 ♦ Histogram of Salary Scores



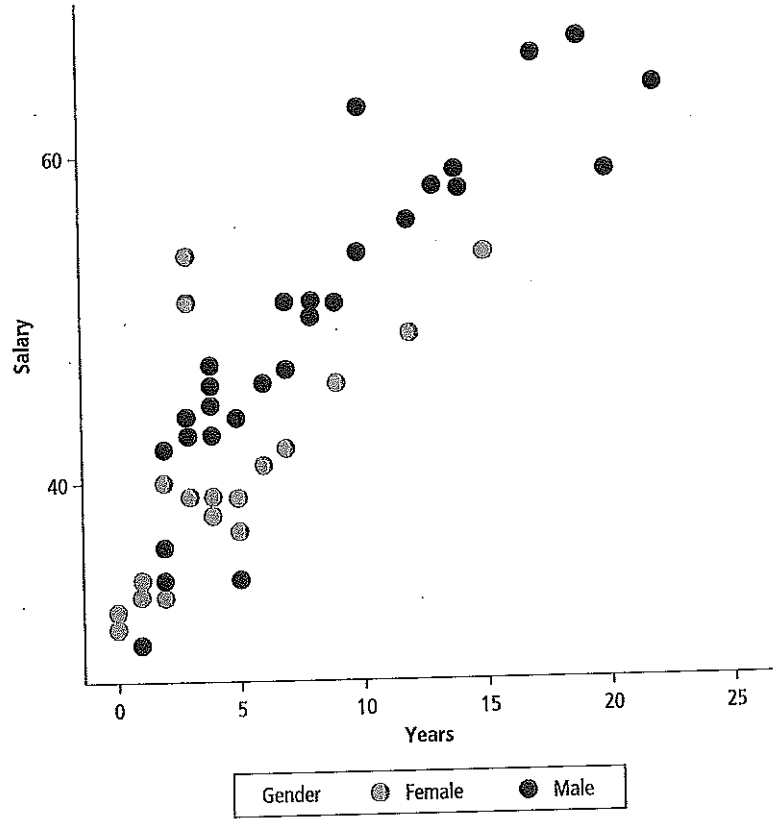
(because salary values cannot be lower than 0). Skewed distributions of scores can sometimes be made more nearly normal in shape by taking the base 10 or natural log of the salary scores. Logarithmic transformation was judged unnecessary for the artificial salary data that are used in the example in this chapter. Also, there were no extreme outliers in salary, as can be seen in Figure 12.2.

A boxplot of salary scores was set up to examine the distribution of outcomes for the female and male groups (Figure 12.3). A visual examination of this boxplot did not reveal any extreme outliers within either group; median salary appeared to be lower for females than for males. The boxplot suggested that the variance of salary scores might be larger for males than for females. This is a possible violation of the homogeneity of variance assumption; in the one-way ANOVA presented in a later section, the Levene test was requested to evaluate whether this difference between salary variances for the female and male groups was statistically significant (and the Levene test was not significant).

Figure 12.4 shows a scatter plot of salary (on the *Y* axis) as a function of years of job experience (on the *X* axis) with different case markers for female and male faculty. The relation between salary and years appears to be linear, and the slope that predicts salary from years appears to be similar for the female and male groups, although there appears

Figure 12.3 ♦ Boxplot of Salary Scores for Female and Male Groups

Figure 12.4 ♦ Scatter Plot of Salary by Years With Case Markers for Gender



to be a tendency for females to have slightly lower salary scores than males at each level of years of job experience; also, the range for years of experience was smaller for females (0–15) than for males (0–22). Furthermore, the variance of salary appears to be reasonably homogeneous across years in the scatter plot that appears in Figure 12.4, and there were no extreme bivariate outliers. Subsequent regression analyses will provide us with more specific information about gender and years as predictors of salary. We will use a regression analysis that includes the following predictors: years, a dummy variable to represent gender, and a product of gender and years. The results of this regression will help answer the following questions: Is there a significant increase in predicted salary associated with years of experience (when gender is statistically controlled)? Is there a significant difference in predicted salary between males and females (when years of experience is statistically controlled)?

12.4 ♦ Issues in Planning a Study

Essentially, when we use a dummy variable as a predictor in a regression, we have the same research situation as when we do a *t* test or ANOVA (both analyses predict scores on the *Y* outcome variable for two groups). When we use several dummy variables as predictors in a regression, we have the same research situation as in a one-way ANOVA (both analyses compare means across several groups). Therefore, the issues reviewed in planning studies that use *t* tests (in Chapter 5) and studies that use one-way ANOVA (in Chapter 6) are also relevant when we use a regression analysis as the method of data analysis. To put it briefly, if the groups that are compared received different "dosage" levels of some treatment variable, the dosage levels need to be far enough apart to produce detectable differences in outcomes. If the groups are formed on the basis of participant characteristics (such as age), the groups need to be far enough apart on these characteristics to yield detectable differences in outcome. Other variables that might create within-group variability in scores may need to be experimentally or statistically controlled to reduce the magnitude of error variance, as described in Chapters 5 and 6.

It is important to check that every group has a reasonable minimum number of cases. If any group has fewer than 10 cases, the researcher may decide to combine that group with one or more other groups (if it makes sense to do so) or exclude that group from the analysis. The statistical power tables that appeared in Chapters 5 and 6 can be used to assess the minimum sample size needed per group to achieve reasonable levels of statistical power.

12.5 ♦ Parameter Estimates and Significance Tests for Regressions With Dummy Variables

The use of one or more dummy predictor variables in regression analysis does not change any of the computations for multiple regression described in Chapter 11. The estimates of *b* coefficients, the *t* tests for the significance of individual *b* coefficients, and the overall *F* test for the significance of the entire regression equation are all computed using the methods described in Chapter 11. Similarly, sr^2 (the squared semipartial correlation), an estimate of the proportion of variance uniquely predictable from each dummy predictor variable, can be calculated and interpreted for dummy predictors in a manner similar to

that described in Chapter 11. Confidence intervals for each b coefficient are obtained using the methods described in Chapter 11.

However, the *interpretation* of b coefficients when they are associated with dummy-coded variables is slightly different from the interpretation when they are associated with continuous predictor variables. Depending on the method of coding that is used, the b_0 (intercept) coefficient may correspond to the mean of one of the groups or to the grand mean of Y. The b_i coefficients for each dummy-coded predictor variable may correspond to contrasts between group means or to differences between group means and the grand mean.

12.6 ♦ Group Mean Comparisons Using One-Way Between-S ANOVA

12.6.1 ♦ Gender Differences in Mean Salary

A one-way between-S ANOVA was performed to assess whether mean salary differed significantly between female and male faculty. No other variables were taken into account in this analysis. A test of the homogeneity of variance assumption was requested (the Levene test).

The Levene test was not statistically significant: $F(1, 48) = 2.81, p = .1$. Thus, there was no statistically significant difference between the variances of salary for the female and male groups; the homogeneity of variance assumption was not violated. The mean salary for males (M_{male}) was 49.9 thousand dollars per year; the mean salary for females (M_{female}) was 40.8 thousand dollars per year. The difference between mean annual salary for males and females was statistically significant at the conventional $\alpha = .05$ level: $F(1, 48) = 13.02$, $p = .001$. The difference between the means ($M_{male} - M_{female}$) was +9.1 thousand dollars (i.e., on average, male faculty earned about 9.1 thousand dollars more per year than female faculty). The eta-squared effect size for this gender difference was .21; in other words, about 21% of the variance in salaries could be predicted from gender. This corresponds to a large effect size (refer to Table 5.2 for suggested verbal labels for values of η^2 and R^2).

This initial finding of a gender difference in mean salary is not necessarily evidence of gender bias in salary levels. Within this sample, there was a tendency for female faculty to have fewer years of experience and for male faculty to have more years of experience, as well as salary increases as a function of years of experience. It is possible that the gender difference we see in this ANOVA is partly or completely accounted for by differences between males and females in years of experience. Subsequent analyses will address this by examining whether gender still predicts different levels of salary when "years of experience" is statistically controlled by including it as a second predictor of salary.

Note that when the numbers of cases in the groups are unequal, there are two different ways in which a grand mean can be calculated. An unweighted grand mean of salary can be obtained by simply averaging male and female mean salary, ignoring sample size. The unweighted grand mean is $(40.8 + 49.9)/2 = 45.35$. However, in many statistical analyses, the estimate of the grand mean is weighted by sample size. The weighted grand mean in this example is found as follows:

$$[(n_{male} \times M_{male}) + (n_{female} \times M_{female})]/(n_{male} + n_{female})$$
$$= [(20 \times 40.8) + (30 \times 49.9)]/(20+30)$$
$$= 46.26.$$

Figure 12.5 ♦ One-Way Between-S ANOVA: Mean Salary for Females and Males

Descriptives

salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Female	20	40.80	6.986	1.562	37.53	44.07	31	54
Male	30	49.90	9.714	1.774	46.27	53.53	30	67
Total	50	46.26	9.750	1.379	43.49	49.03	30	67

Test of Homogeneity of Variances

salary

Levene Statistic	df1	df2	Sig.
2.809	1	48	.100

ANOVA

salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	993.720	1	993.720	13.019	.001
Within Groups	3663.900	48	76.331		
Total	4657.620	49			

The grand mean for salary reported in the one-way ANOVA in Figure 12.5 corresponds to the weighted grand mean. (This weighted grand mean is equivalent to the sum of the 50 individual salary scores divided by the number of scores in the sample.) In the regression analyses reported later in this chapter, the version of the grand mean that appears in the results corresponds to the unweighted grand mean.

12.6.2 ♦ College Differences in Mean Salary

In a research situation that involves a categorical predictor variable with more than two levels or groups and a quantitative outcome variable, the most familiar approach to data analysis is a one-way ANOVA. To evaluate whether mean salary level differs for faculty across the three different colleges in the hypothetical dataset in Table 12.1, we can conduct a between-S one-way ANOVA using the SPSS ONEWAY procedure. The variable college is coded 1 = Liberal Arts, 2 = Sciences, and 3 = Business. The outcome variable, as in previous analyses, is annual salary in thousands of dollars. Orthogonal contrasts between colleges were also requested by entering custom contrast coefficients. The results of this one-way ANOVA appear in Figure 12.6.

The means on salary were as follows: 44.8 for faculty in Liberal Arts, 44.7 in Sciences, and 51.6 in Business. The overall F for this one-way ANOVA was not statistically significant: $F(2, 47) = 2.18, p = .125$. The effect size, eta squared, was obtained by taking the ratio $SS_{between}/SS_{total}$; for this ANOVA, $\eta^2 = .085$. This is a medium effect. In this situation, if we want to use this sample to make inferences about some larger population of faculty, we would not have evidence that the proportion of variance in salary that is predictable from

Figure 12.6 ♦ One-Way Between-S ANOVA: Mean Salary Across Colleges

salary

Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Liberal Arts	22	44.82	9.261	1.975	40.71	48.92	31	66
Sciences	17	44.71	9.777	2.371	39.68	49.73	30	67
Business	11	51.55	9.658	2.912	45.06	58.03	32	64
Total	50	46.26	9.750	1.379	43.49	49.03	30	67

Test of Homogeneity of Variances

salary

Levene Statistic	df1	df2	Sig.
.005	2	47	.995

ANOVA

salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	394.091	2	197.045	2.172	.125
Within Groups	4263.529	47	90.713		
Total	4657.620	49			

Contrast Coefficients

Contrast	college		
	Liberal Arts	Sciences	Business
1	1	-1	0
2	1	1	-2

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
salary	Assume equal variances	1	.11	3.076	.037	47	.971
		2	-13.57	6.515	-2.082	47	.043
	Does not assume equal variances	1	.11	3.086	.036	33.580	.971
		2	-13.57	6.591	-2.058	16.027	.056

college is significantly different from 0. However, if we just want to describe the strength of the association between college and salary within this sample, we could say that, for this sample, about 8.5% of the variance in salary was predictable from college. The orthogonal contrasts that were requested made the following comparison. For Contrast 1, the custom contrast coefficients were +1, −1, and 0; this corresponds to a comparison of the mean salaries between College 1 (Liberal Arts) and College 2 (Sciences); this contrast was not statistically significant: $t(47) = .037$, $p = .97$. For Contrast 2, the custom contrast coefficients were +1, +1, and −2; this corresponds to a comparison of the mean salary for Liberal Arts and Sciences faculty combined, compared with the Business faculty. This contrast was statistically significant: $t(47) = -2.082$, $p = .043$. Business faculty had a significantly higher mean salary than the two other colleges (Liberal Arts and Sciences) combined.

The next sections show that regression analyses with dummy predictor variables can be used to obtain the same information about the differences between group means. Dummy variables that represent group membership (such as female and male, or Liberal Arts, Sciences, and Business colleges) can be used to predict salary in regression analyses. We will see that the information about differences among group means that can be obtained by doing regression with dummy predictor variables is equivalent to the information that we can obtain from one-way ANOVA. In future chapters, both dummy variables and quantitative variables are used as predictors in regression.

12.7 ♦ Three Methods of Coding for Dummy Variables

The three coding methods that are most often used for dummy variables are given below (the details regarding these methods are presented in subsequent sections along with empirical examples):

1. Dummy coding of dummy variables
2. Effect coding of dummy variables
3. Orthogonal coding of dummy variables

In general, when we have k groups, we need only $(k - 1)$ dummy variables to represent information about group membership. Most dummy variable codes can be understood as answers to yes/no questions about group membership. For example, to represent college group membership, we might include a dummy variable D_1 that corresponds to the question, "Is this faculty member in Liberal Arts?" and code the responses as 1 = yes, 0 = no. If there are k groups, a set of $(k - 1)$ yes/no questions provides complete information about group membership. For example, if a faculty member reports responses of "no" to the questions, "Are you in Liberal Arts?" and "Are you in Science?" and there are only three groups, then that person must belong to the third group (in this example, Group 3 is the Business college).

The difference between dummy coding and effect coding is in the way in which codes are assigned for members of the last group—that is, the group that does not correspond to an explicit yes/no question about group membership. In dummy coding of dummy variables, members of the last group receive a score of 0 on all the dummy variables. (In effect coding of dummy variables, members of the last group are assigned scores of −1 on all the dummy variables.) This difference in codes results in slightly different interpretations of the b coefficients in the regression equation, as described in the subsequent sections of this chapter.

12.7.1 ♦ Regression With Dummy-Coded Dummy Predictor Variables

12.7.1.1 ♦ *Two-Group Example With a Dummy-Coded Dummy Variable*

Suppose we want to predict salary (Y) from gender; gender is a **dummy-coded dummy variable** with codes of 0 for female and 1 for male participants in the study. In a previous section, this difference was evaluated by doing a one-way between-S ANOVA to compare mean salary across female and male groups. We can obtain equivalent information

about the magnitude of gender differences in salary from a regression analysis that uses a dummy-coded variable to predict salary. For this simple two-group case (prediction of salary from dummy-coded gender), we can write a regression equation using gender to predict salary (Y) as follows:

$$\text{Salary}' \text{ or } Y' = b_0 + b_1 \times \text{Gender}. \tag{12.2}$$

From Equation 12.2, we can work out two separate prediction equations: one that makes predictions of Y for females and one that makes predictions of Y for males. To do this, we substitute the values of 0 (for females) and 1 (for males) into Equation 12.2 and simplify the expression to obtain these two different equations:

$$Y' = b_0 \text{ (for females, gender} = 0), \tag{12.3}$$

$$Y' = b_0 + b_1 \text{ (for males, gender} = 1). \tag{12.4}$$

These two equations tell us that the constant value b_0 is the best prediction of salary for females, and the constant value $(b_0 + b_1)$ is the best prediction of salary for males. This implies that b_0 = mean salary for females, $b_0 + b_1$ = mean salary for males, and b_1 = the difference between mean salary for the male versus female groups. The slope coefficient b_1 corresponds to the difference in mean salary for males and females. If the b_1 slope is significantly different from 0, it implies that there is a statistically significant difference in mean salary for males and females.

The results of the regression in Figure 12.7 provide the numerical estimates for the raw score regression coefficients (b_0 and b_1) for this set of data:

$$\text{Salary}' = 40.8 + 9.1 \times \text{Gender}.$$

For females, with gender = 0, the predicted mean salary given by this equation is $40.8 + 9.1 \times 0 = 40.8$. Note that this is the same as the mean salary for females in the one-way ANOVA output in Figure 12.5. For males, with gender = 1, the predicted salary given by this equation is $40.8 + 9.1 = 49.9$. Note that this value is equal to the mean salary for males in the one-way ANOVA in Figure 12.5.

The b_1 coefficient in this regression was statistically significant: $t(48) = 3.61$, $p = .001$. The F test reported in Figure 12.6 is equivalent to the square of the t test value for the null hypothesis that $b_1 = 0$ in Figure 12.7 ($t = +3.608$, $t^2 = 13.02$). Note also that the eta-squared effect size associated with the ANOVA ($\eta^2 = .21$) and the R^2 effect size associated with the regression were equal; in both analyses, about 21% of the variance in salaries was predictable from gender.

When we use a dummy variable with codes of 0 and 1 to represent membership in two groups, the value of the b_0 intercept term in the regression equation is equivalent to the mean of the group for which the dummy variable has a value of 0. The b_1 "slope" coefficient represents the difference (or contrast) between the means of the two groups. The slope, in this case, represents the change in the mean level of Y when you move from a code of 0 (female) to a code of 1 (male) on the dummy predictor variable.

Figure 12.7 ♦ Regression to Predict Salary From Dummy-Coded Gender

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gender [a]		Enter

a. All requested variables entered.

b. Dependent Variable: salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.462 [a]	.213	.197	8.737

a. Predictors: (Constant), gender

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	993.720	1	993.720	13.019	.001 [a]
	Residual	3663.900	48	76.331		
	Total	4657.620	49			

a. Predictors: (Constant), gender

b. Dependent Variable: salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	40.800	1.954		20.884	.000			
	gender	9.100	2.522	.462	3.608	.001	.462	.462	.462

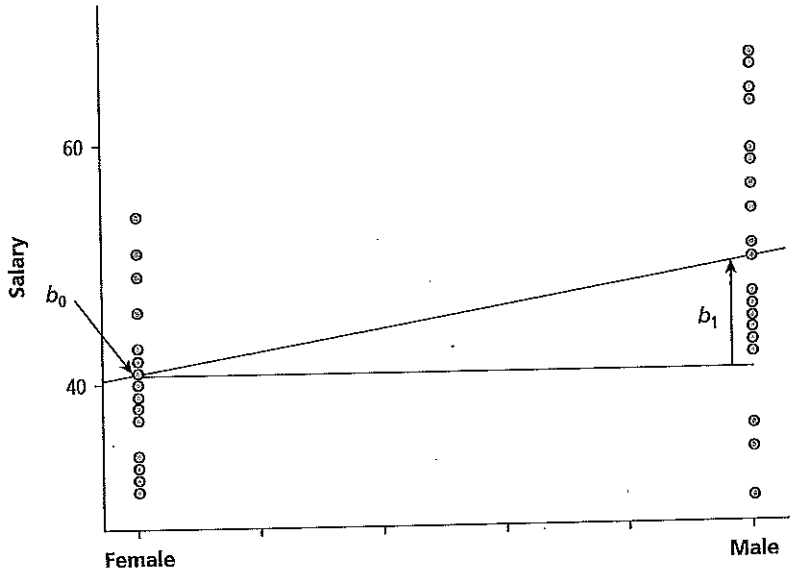
a. Dependent Variable: salary

Figure 12.8 shows a scatter plot of salary scores (on the Y axis) as a function of gender code (on the X axis). In this graph, the intercept b_0 corresponds to the mean on the dependent variable (salary) for the group that had a dummy variable score of 0 (females). The slope, b_1, corresponds to the difference between the means for the two groups—that is, the change in the predicted mean when you move from a code of 0 to a code of 1 on the dummy variable. That is, $b_1 = M_{male} - M_{female}$, the change in salary when you move from a score of 0 (female) to a score of 1 (male). The test of the statistical significance of b_1 is equivalent to the t test of the difference between the mean Y values for the two groups represented by the dummy variable in the regression.

12.7.1.2 ♦ *Multiple-Group Example With Dummy-Coded Dummy Variables*

When there are multiple groups (number of groups = k), group membership can be represented by scores on a set of $(k - 1)$ dummy variables. Each dummy variable essentially represents a yes/no question about group membership. In the preceding example, there are $k = 3$ college groups in the faculty data. In this example, we will use two dummy variables, denoted by D_1 and D_2, to represent information about college membership in a

Figure 12.8 ◆ Graph for Regression to Predict Salary From Dummy-Coded Gender (0 = Female, 1 = Male)



regression analysis. D_1 corresponds to the following question: Is the faculty member from the Liberal Arts college? 1 = yes, 0 = no. D_2 corresponds to the following question: Is the faculty member from the Sciences college? 1 = yes, 0 = no. For dummy coding, members of the last group receive a score of 0 on all the dummy variables. In this example, faculty from the Business college received scores of 0 on both the D_1 and D_2 dummy-coded dummy variable. For the set of three college groups, the dummy-coded dummy variables that provide information about college group membership were coded as follows:

	D_1	D_2
Liberal Arts	1	0
Sciences	0	1
Business	0	0

Now that we have created dummy variables that represent information about membership in the three college groups as scores on a set of dummy variables, mean salary can be predicted from college groups by a regression analysis that uses the dummy-coded dummy variables shown above as predictors:

$$Y' = b_0 + b_1 D_1 + b_2 D_2. \tag{12.5}$$

The results of the regression (using dummy-coded dummy variables to represent career group membership) are shown in Figure 12.9. Note that the overall F reported for the

regression analysis in Figure 12.9 is identical to the overall F reported for the one-way ANOVA in Figure 12.6, $F(2, 47) = 2.17$, $p = .125$, and that the η^2 for the one-way ANOVA is identical to the R^2 for the regression ($\eta^2 = R^2 = .085$). Note also that the b_0 coefficient in the regression results in Figure 12.9 (b_0 = constant = 51.55) equals the mean salary for the group that was assigned score values of 0 on all the dummy variables (the mean salary for the Business faculty was 51.55). Note also that the b_i coefficients for each of the two dummy variables represent the difference between the mean of the corresponding group and the mean of the comparison group whose codes were all 0; for example, $b_1 = -6.73$, which corresponds to the difference between the mean salary of Group 1, Liberal Arts ($M_1 = 44.82$) and mean salary of the comparison group, Business ($M_3 = 51.55$); the value of the b_2 coefficient ($b_2 = -6.84$) corresponds to the difference between mean salary for the Science ($M = 44.71$) and Business ($M = 51.55$) groups. This regression analysis with dummy variables to represent college membership provided information equivalent to a one-way ANOVA to predict salary from college.

Figure 12.9 ♦ Regression to Predict Salary From Dummy-Coded College Membership

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	d2, d1[a]		Enter

a. All requested variables entered.

b. Dependent Variable: salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.291[a]	.085	.046	9.524

a. Predictors: (Constant), d2, d1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	394.091	2	197.045	2.172	.125[a]
	Residual	4263.529	47	90.713		
	Total	4657.620	49			

a. Predictors: (Constant), d2, d1

b. Dependent Variable: salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	51.545	2.872		17.949	.000			
	d1	-6.727	3.517	-.346	-1.913	.062	-.132	-.269	-.267
	d2	-6.840	3.685	-.336	-1.856	.070	-.116	-.261	-.259

a. Dependent Variable: salary

12.7.2 ♦ Regression With Effect-Coded Dummy Predictor Variables

12.7.2.1 ♦ *Two-Group Example With an Effect-Coded Dummy Variable*

We will now code the scores for gender slightly differently, using a method called "effect coding of dummy variables." In effect coding of dummy variables, a score value of 1 is used to represent a "yes" answer to a question about group membership; membership in the group that does not correspond to a "yes" answer on any of the group membership questions is represented by a score of −1. In the following example, the **effect-coded dummy variable** "geneff" (Is the participant male, yes or no?) is coded +1 for males and −1 for females. The variable geneff is called an effect-coded dummy variable because we used −1 (rather than 0) as the value that represents membership in the last group. Our overall model for the prediction of salary (Y) from gender, represented by the effect-coded dummy variable geneff, can be written as follows:

$$Y' = b_0 + b_1 \times \text{Geneff.} \tag{12.6}$$

Substituting the values of +1 for males and −1 for females, the predictive equations for males and females become

$$Y' = b_0 + b_1 \text{ (for males, with geneff coded +1),} \tag{12.7}$$

$$Y' = b_0 - b_1 \text{ (for females, with geneff coded −1).} \tag{12.8}$$

From earlier discussions on t tests and ANOVA, we know that the best predicted value of Y for males is equivalent to the mean on Y for males, M_{male}; similarly, the best predicted value of Y for females is equal to the mean on Y for females, M_{female}. The two equations above, therefore, tell us that $M_{\text{male}} = b_0 + b_1$ and $M_{\text{female}} = b_0 - b_1$. What does this imply for the values of b_0 and b_1? The mean for males is b_1 units above b_0; the mean for females is b_1 units below b_0. With a little thought, you will see that the intercept b_0 must equal the grand mean on salary for both genders combined.

Note that when we calculate a grand mean by combining group means, there are two different possible ways to calculate the grand mean. If the groups have the same numbers of scores, these two methods yield the same result, but when the groups have unequal numbers of cases, these two methods for computation of the grand mean yield different results. Whenever you do analyses with unequal numbers in the groups, you need to decide whether the unweighted or the weighted grand mean is a more appropriate value to report. In some situations, it may not be clear what default decision a computer program uses (i.e., whether the program reports the weighted or the unweighted grand mean), but it is possible to calculate both the weighted and unweighted grand means by hand from the group means; when you do this, you will be able to determine which version of the grand mean was reported on the SPSS printout.

The unweighted grand mean for salary for males and females is obtained by ignoring the number of cases in the groups and averaging the group means together for males and females. For the male and female salary data that appeared in Table 12.1, the **unweighted mean** is $(M_{\text{male}} + M_{\text{female}})/2 = (40.80 + 49.90)/2 = 45.35$. Note that the b_0 constant or

intercept term in the regression in Figure 12.10 that uses effect-coded gender to predict salary corresponds to this unweighted grand mean of 45.35. When you run a regression to predict scores on a quantitative outcome variable from effect-coded dummy predictor variables, and the default methods of computation are used in SPSS, the b_0 coefficient in the regression equation corresponds to the unweighted grand mean, and effects (or differences between group means and grand means) are reported relative to this unweighted grand mean as a reference point. This differs slightly from the one-way ANOVA output in Figure 12.5, which reported the weighted grand mean for salary (46.26).

When effect-coded dummy predictor variables are used, the slope coefficient b_1 corresponds to the "effect" of gender; that is, $+b_1$ is the distance between the male mean and the grand mean, and $-b_1$ is the distance between the female mean and the grand mean. In Chapter 6 on one-way ANOVA, the terminology used for these distances (group mean minus grand mean) was *effect*. The effect of membership in Group i in a one-way ANOVA is represented by α_i, where $\alpha_i = M_i - M_Y$, the mean of Group i minus the grand mean of Y across all groups.

This method of coding (+1 vs. −1) is called "effect coding," because the intercept b_0 in Equation 12.5 equals the unweighted grand mean for salary, M_Y, and the slope coefficient b_i for each effect-coded dummy variable E represents the effect for the group that has a code of 1 on that variable—that is, the difference between that group's mean on Y and the unweighted grand mean. Thus, when effect-coded dummy variables are used to represent group membership, the b_0 intercept term equals the grand mean for Y, the outcome variable, and each b_i coefficient represents a contrast between the mean of one group versus the unweighted grand mean (or the "effect" for that group). The significance of b_1 for the effect-coded variable geneff is a test of the significance of the difference between the mean of the corresponding group (in this example, males) and the unweighted grand mean. Given that geneff is coded −1 for females and +1 for males and given that the value of b_1 is significant and positive, the mean salary of males is significantly higher than the grand mean (and the mean salary for females is significantly lower than the grand mean).

Note that we do not have to use a code of +1 for the group with the higher mean and a code of −1 for the group with the lower mean on Y. The sign of the b coefficient can be either positive or negative; it is the combination of signs (on the code for the dummy variable and the b coefficient) that tells us which group had a mean that was lower than the grand mean.

The overall F result for the regression analysis that predicts salary from effect-coded gender (geneff) in Figure 12.10 is identical to the F value in the earlier analyses of gender and salary reported in Figures 12.5 and 12.7: $F(1, 48) = 13.02, p = .001$. The effect size given by η^2 and R^2 is also identical across these three analyses ($R^2 = .21$). The only difference between the regression that uses a dummy-coded dummy variable (Figure 12.7) and the regression that uses an effect-coded dummy variable to represent gender (Figure 12.10) is in the way in which the b_0 and b_1 coefficients are related to the grand mean and group means.

12.7.1.2 ♦ *Multiple-Group Example With Effect-Coded Dummy Variables*

If we used effect coding instead of dummy coding to represent membership in the three college groups used as an example earlier, group membership could be coded as follows:

	E_1	E_2
Liberal Arts	1	0
Sciences	0	1
Business	−1	−1

That is, E_1 and E_2 still represent yes/no questions about group membership. E_1 corresponds to the following question: Is the faculty member in Liberal Arts? Coded 1 = yes, 0 = no. E_2 corresponds to the following question: Is the faculty member in Sciences? Coded 1 = yes, 0 = no. The only change when effect coding (instead of dummy coding) is used is that members of the Business college (the one group that does not correspond directly to a yes/no question) now receive codes of −1 on both the variables E_1 and E_2.

We can run a regression to predict scores on salary from the two effect-coded dummy variables E_1 and E_2:

$$\text{Salary}' \text{ or } Y' = b_0 + b_1E_1 + b_2E_2. \tag{12.9}$$

Figure 12.10 ♦ Regression to Predict Salary From Effect-Coded Gender

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	geneff [a]		Enter

a. All requested variables entered.

b. Dependent Variable: salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.462 [a]	.213	.197	8.737

a. Predictors: (Constant), geneff

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	993.720	1	993.720	13.019	.001 [a]
	Residual	3663.900	48	76.331		
	Total	4657.620	49			

a. Predictors: (Constant), geneff

b. Dependent Variable: salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	45.350	1.261		35.962	.000			
	geneff	4.550	1.261	.462	3.608	.001	.462	.462	.462

a. Dependent Variable: salary

The results of this regression analysis are shown in Figure 12.11.

When effect coding is used, the intercept or b_0 coefficient is interpreted as an estimate of the (unweighted) grand mean for the Y outcome variable, and each b_i coefficient represents the effect for one of the groups—that is, the contrast between a particular group mean and the grand mean. (Recall that when dummy coding was used, the intercept b_0 was interpreted as the mean of the "last" group—namely, the group that did not correspond to a "yes" answer on any of the dummy variables—and each b_i coefficient corresponded to the difference between one of the group means and the mean of the "last" group, the group that is used as the reference group for all comparisons.) In Figure 12.11, the overall F value and the overall R^2 are the same as in the two previous analyses that compared mean salary across college (in Figures 12.6 and 12.9): $F(2, 47) = 2.12, p = .125$; $R^2 = .085$. The b coefficients for Equation 12.9, from Figure 12.11, are as follows:

$$\text{Salary}' = 47.03 - 2.205 \times E_1 - 2.317 \times E_2.$$

The interpretation is as follows: The (unweighted) grand mean of salary is 47.03 thousand dollars per year. Members of the Liberal Arts faculty have a predicted annual salary that is 2.205 thousand dollars less than this grand mean; members of the Sciences faculty

Figure 12.11 ♦ Regression to Predict Salary From Effect-Coded College Membership

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	e2, e1 [a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.291 [a]	.085	.046	9.524

a. Predictors: (Constant), e2, e1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	394.091	2	197.045	2.172	.125 [a]
	Residual	4263.529	47	90.713		
	Total	4657.620	49			

a. Predictors: (Constant), e2, e1

b. Dependent Variable: salary

Coefficients[b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	47.023	1.403		33.525	.000			
	e1	-2.205	1.828	-.179	-1.206	.234	-.238	-.173	-.168
	e2	-2.317	1.935	-.177	-1.197	.237	-.237	-.172	-.167

a. Dependent Variable: salary

have a predicted salary that is 2.317 thousand dollars less than this grand mean. Neither of these differences between a group mean and the grand mean is statistically significant at the $\alpha = .05$ level.

Because members of the Business faculty have scores of −1 on both E_1 and E_2, the predicted mean salary for Business faculty is

$$\text{Salary}' = 47.03 + 2.205 + 2.317 = 51.5 \text{ thousand dollars per year.}$$

We do not have a significance test to evaluate whether the mean salary for Business faculty is significantly higher than the grand mean. If we wanted to include a significance test for this contrast, we could do so by rearranging the dummy variable codes associated with group membership, such that membership in the Business college group corresponded to an answer of "yes" on either E_1 or E_2.

12.7.3 ♦ Orthogonal Coding of Dummy Predictor Variables

We can set up contrasts among group means in such a way that the former are orthogonal (the term *orthogonal* is equivalent to uncorrelated or independent). One method of creating orthogonal contrasts is to set up one contrast that compares Group 1 versus Group 2 and a second contrast that compares Groups 1 and 2 combined versus Group 3, as in the example below:

	Group		
	1 (Liberal Arts)	*2 (Sciences)*	*3 (Business)*
O_1	+1	−1	0
O_2	+1	+1	−2

The codes across each row should sum to 0. For each **orthogonally coded dummy variable**, the groups for which the code has a positive sign are contrasted with the groups for which the code has a negative sign; groups with a code of 0 are ignored. Thus, O_1 compares the mean of Group 1 (Liberal Arts) with the mean of Group 2 (Sciences).

To figure out which formal null hypothesis is tested by each contrast, we form a weighted linear composite that uses these codes. That is, we multiply the population mean μ_k for Group k by the contrast coefficient for Group k and sum these products across the k groups; we set that weighted linear combination of population means equal to 0 as our null hypothesis.

In this instance, the null hypotheses that correspond to the contrast specified by the two O_i orthogonally coded dummy variables are as follows:

$$H_0 \text{ for } O_1\text{: } (+1)\mu_1 + (-1)\mu_2 + (0)\mu_3 = 0.$$

That is,

$$H_0 \text{ for } O_1\text{: } \mu_1 - \mu_2 = 0 \text{ (or } \mu_1 = \mu_2).$$

The O_2 effect-coded dummy variable compares the average of the first two group means (i.e., the mean for Liberal Arts and Sciences combined) with the mean for the third group (Business):

$$H_0 \text{ for } O_2\text{: } (+1)\mu_1 + (+1)\mu_2 + (-2)\mu_3,$$
$$H_0 \text{ for } O_2\text{: } (\mu_1 + \mu_2) - 2\mu_3 = 0,$$

or

$$\frac{\mu_1 + \mu_2}{2} - \mu_3 = 0 \text{ or } \frac{\mu_1 + \mu_2}{2} = \mu_3.$$

We can assess whether the contrasts are orthogonal by taking the cross products and summing the corresponding coefficients. Recall that products between sets of scores provide information about covariation or correlation; see Chapter 7 for details. Because each of the two variables O_1 and O_2 has a sum (and, therefore, a mean) of 0, each code represents a deviation from a mean. When we compute the sum of cross products between corresponding values of these two variables, we are, in effect, calculating the numerator of the correlation between O_1 and O_2. In this example, we can assess whether O_1 and O_2 are orthogonal or uncorrelated by calculating the following sum of cross products. For O_1 and O_2, the sum of cross products of the corresponding coefficients is

$$(+1)(+1) + (-1)(+1) + (0)(-2) = 0.$$

Because this sum of the products of corresponding coefficients is 0, we know that the contrasts specified by O_1 and O_2 are orthogonal.

Of course, as an alternate way to see whether the O_1 and O_2 predictor variables are orthogonal or uncorrelated, SPSS can also be used to calculate a correlation between O_1 and O_2; if the contrasts are orthogonal, Pearson's r between O_1 and O_2 will equal 0 (provided that the numbers in the groups are equal).

For each contrast specified by a set of codes (e.g., the O_1 set of codes), any group with a 0 coefficient is ignored, and groups with opposite signs are contrasted. The direction of the signs for these codes does not matter; the contrast represented by the codes (+1, −1, and 0) represents the same comparison as the contrast represented by (−1, +1, 0). The b coefficients obtained for these two sets of codes would be opposite in sign, but the significance of the difference between the means of Group 1 and Group 2 would be the same whether Group 1 was assigned a code of +1 or −1.

Figure 12.12 shows the results of a regression in which the dummy predictor variables O_1 and O_2 are coded to represent the same orthogonal contrasts. Note that the t tests for significance of each contrast are the same in both Figure 12.6, where the contrasts were requested as an optional output from the ONEWAY procedure, and Figure 12.12, where the contrasts were obtained by using orthogonally coded dummy variables as predictors of salary. Only one of the two contrasts, the contrast that compares the mean salary for Liberal Arts and Sciences faculty with the mean salary for Business faculty, was statistically significant.

Figure 12.12 ♦ Regression to Predict Salary From Dummy Variables That Represent Orthogonal Contrasts

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	o2, o1[a]		Enter

a. All requested variables entered.

b. Dependent Variable: salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.291[a]	.085	.046	9.524

a. Predictors: (Constant), o2, o1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	394.091	2	197.045	2.172	.125[a]
	Residual	4263.529	47	90.713		
	Total	4657.620	49			

a. Predictors: (Constant), o2, o1

b. Dependent Variable: salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	47.023	1.403		33.525	.000			
	o1	.056	1.538	.005	.037	.971	-.013	.005	.005
	o2	-2.261	1.086	-.291	-2.082	.043	-.291	-.291	-.291

a. Dependent Variable: salary

Orthogonal coding of dummy variables can also be used to perform a trend analysis—for example, when the groups being compared represent equally spaced dosage levels along a continuum. The following example shows orthogonal coding to represent linear versus quadratic trends for a study in which the groups receive three different dosage levels of caffeine.

	0 mg	*150 mg*	*300 mg*
O_1	−1	0	+1
O_2	+1	−2	+1

Note that the sum of cross products is again 0 [(−1)(+1) + (0)(−2) + (+1)(+1)], so these contrasts are orthogonal. A simple way to understand what type of trend is represented

by each line of codes is to visualize the list of codes for each dummy variable as a template or graph. If you place values of −1, 0, and +1 from left to right on a graph, it is clear that these codes represent a linear trend. The set of coefficients +1, −2, and +1 or, equivalently, −1, +2, and −1 represent a quadratic trend. So, if b_1 (the coefficient for O_1) is significant with this set of codes, the linear trend is significant and b_1 is the amount of change in the dependent variable Y from 0 to 150 mg or from 150 to 300 mg. If b_2 is significant, there is a quadratic (curvilinear) trend.

12.8 ♦ Regression Models That Include Both Dummy and Quantitative Predictor Variables

We can do a regression analysis that includes one (or more) dummy variables and one (or more) continuous predictors, as in the following example:

$$Y' = b_0 + b_1 D + b_2 X. \tag{12.10}$$

How is the b_1 coefficient for the dummy variable, D, interpreted in the context of this multiple regression with another predictor variable? The b_1 coefficient still represents the estimated difference between the means of the two groups; if gender was coded 0, 1 as in the first example, b_1 still represents the difference between mean salary Y for males and females. However, in the context of this regression analysis, this difference between means on the Y outcome variable for the two groups is assessed while statistically controlling for any differences in the quantitative X variable (such as years of job experience).

Numerical results for this regression analysis appear in Figure 12.13. From these results we can conclude that both gender and years are significantly predictive of salary. The coefficient to predict salary from gender (controlling for years of experience) was $b_2 = 3.36$, with $t(47) = 2.29$, $p = .026$. The corresponding squared semipartial (or part) correlation for gender as a predictor of salary was $sr^2 = (.159)^2 = .03$. The coefficient to predict salary from years of experience, controlling for gender, was $b = 1.44$, $t(47) = 10.83$, $p < .001$; the corresponding sr^2 effect size for years was $.749^2 = .56$. This analysis suggests that controlling for years of experience partly accounts for the observed gender differences in salary, but it does not completely account for gender differences; even after years of experience is taken into account, males still have an average salary that is about 3.36 thousand dollars higher than females at each level of years of experience. However, this gender difference is relatively small in terms of the proportion of explained variance. About 56% of the variance in salaries is uniquely predictable from years of experience (this is a very strong effect). About 3% of the variance in salary is predictable from gender (this is a medium-sized effect). Within this sample, for each level of years of experience, females are paid about 3.35 thousand dollars less than males who have the same number of years of experience; this is evidence of possible gender bias. Of course, it is possible that this remaining gender difference in salary might be accounted for by other variables. Perhaps more women are in the college of Liberal Arts, which has lower salaries, and more men are in the Business and Science colleges, which tend to receive higher salaries. Controlling for other variables such as college might help us to account for part of the gender difference in mean salary levels.

The raw score b coefficient associated with gender in the regression in Figure 12.13 had a value of $b = 3.355$; this corresponds to the difference between the intercepts of the regression lines for males and females in Figure 12.14. For this set of data, it appears that gender differences in salary may be due to a difference in starting salaries (i.e., the salaries paid to faculty with 0 years of experience) and not due to differences between the annual raises in salary given to male and female faculty. The model Results section at the end of the chapter provides a more detailed discussion of the results that appear in Figures 12.13 and 12.14.

The best interpretation for the salary data in Table 12.1, based on the analyses that have been performed so far, appears to be the following: Salary significantly increases as a function of years of experience; there is also a gender difference in salary (such that males are paid significantly higher salaries than females) even when the effect of years of experience is statistically controlled. However, keep in mind that statistically controlling for additional predictor variables (such as college and merit) in later analyses could substantially change the apparent magnitude of gender differences in salary.

Figure 12.13 ♦ Regression to Predict Salary From Gender and Years

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gender, years[a]		Enter

a. All requested variables entered.

b. Dependent Variable: salary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.880[a]	.775	.765	4.722

a. Predictors: (Constant), gender, years

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3609.469	2	1804.734	80.926	.000[a]
	Residual	1048.151	47	22.301		
	Total	4657.620	49			

a. Predictors: (Constant), gender, years

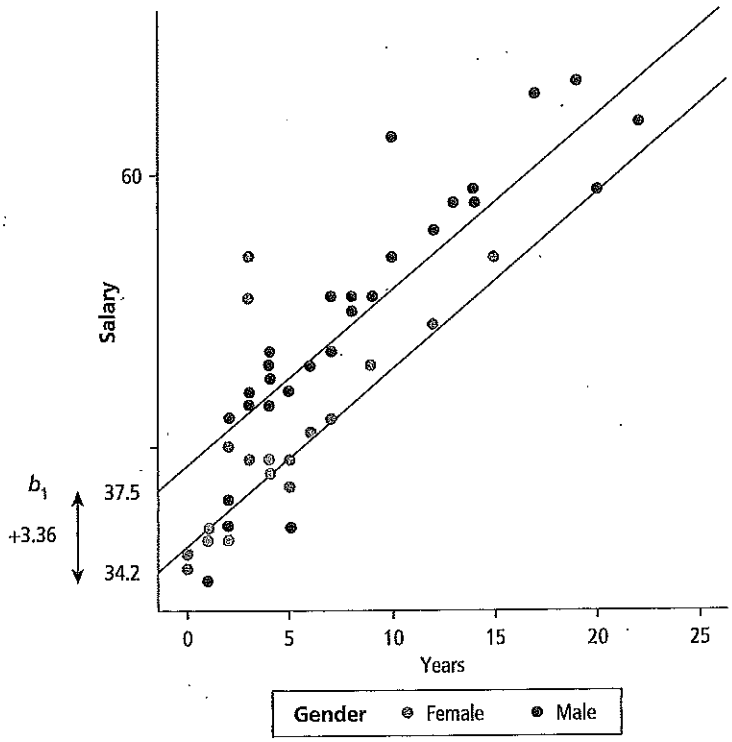
b. Dependent Variable: salary

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	34.193	1.220		28.038	.000			
	years	1.436	.133	.804	10.830	.000	.866	.845	.749
	gender	3.355	1.463	.170	2.293	.026	.462	.317	.159

a. Dependent Variable: salary

Figure 12.14 ♦ Graph of the Regression Lines to Predict Salary From Years of Experience Separately for Males and Females



NOTE: This is based on the regression analysis in Figure 12.13: Salary′ = 34.2 + 3.36 × Gender + 1.44 × Years; gender coded 0 = female and 1 = male.

12.9 ♦ Effect Size and Statistical Power

As discussed in Chapter 11, we can represent the effect size for the regression as a whole (i.e., the proportion of variance in Y is predictable from a set of variables that may include dummy-coded and/or continuous variables) by reporting multiple R and multiple R^2 as our overall effect-size measures. We can represent the strength of the unique predictive contribution of any particular variable by reporting the estimate of sr^2_{unique} for each predictor variable, as in Chapter 11. The proportion of variance given by sr^2 can be used to describe the proportion of variance uniquely predictable from the contrast specified by a dummy variable, in the same manner in which it describes the proportion of variance uniquely predictable from a continuous predictor variable.

When only dummy variables are included in regression analysis, the regression is essentially equivalent to a one-way ANOVA (or a t test, if only two groups are being compared). Therefore, the tables presented in Chapter 5 (independent samples t test) and Chapter 6 (one-way ANOVA) can be used to look up reasonable minimum sample sizes per group for anticipated effect sizes that are small, medium, or large. Whether the method used to make predictions and compare means across groups is ANOVA or regression, none

of the groups should have a very small n. If $n < 20$ per group, nonparametric analyses may be more appropriate.

12.10 ♦ Nature of the Relationship and/or Follow-Up Tests

When dummy-coded group membership predictors are included in a regression analysis, the information that individual coefficients provide is equivalent to the information obtained from planned contrasts between group means in an ANOVA. The choice of the method of coding (dummy, effect, orthogonal) and the decision as to which group to code as the "last" group determine which set of contrasts the regression will include.

Whether we compare multiple groups by performing a one-way ANOVA or by using a regression equation with dummy-coded group membership variables as predictors, the written results should include the following: means and standard deviations for scores in each group, confidence intervals for each group mean, an overall F test to report whether there were significant differences in group means, planned contrasts or post hoc tests to identify which specific pairs of group means differed significantly, and a discussion of the direction of differences between group means. Effect-size information about the proportion of explained variance (in the form of an η^2 or sr^2) should also be included.

12.11 ♦ Results

The hypothetical data showing salary scores for faculty (in Table 12.1) were analyzed in several different ways to demonstrate the equivalence between ANOVA and regression with dummy variables and to illustrate the interpretation of b coefficients for dummy variables in regression. The text below reports the regression analysis for prediction of salary from years of experience and gender (as shown in Figure 12.13).

Results

To assess whether gender and years of experience significantly predict faculty salary, a regression analysis was performed to predict faculty annual salary in thousands of dollars from gender (dummy-coded 0 = female, 1 = male) and years of experience. The distribution of salary was roughly normal, the variances of salary scores were not significantly different for males and females, and scatter plots did not indicate nonlinear relations or bivariate outliers. No data transformations were applied to scores on salary and years, and all 50 cases were included in the regression analysis.

The results of this regression analysis (SPSS output in Figure 12.13) indicated that the overall regression equation was significantly predictive of salary; $R = .88$, $R^2 = .78$, adjusted $R^2 = .77$, $F(2, 47) = 80.93$, $p < .001$. Salary could be predicted almost perfectly from gender and years of job experience. Each of the two individual predictor variables was statistically significant. The raw score coefficients for the predictive equation were as follows:

$$\text{Salary}' = 34.19 + 3.36 \times \text{Gender} + 1.44 \times \text{Years}.$$

When controlling for the effect of years of experience on salary, the magnitude of the gender difference in salary was 3.36 thousand dollars. That is, at each level of years of experience, male annual salary was about 3.36 thousand dollars higher than female salary. This difference was statistically significant: $t(47) = 2.29$, $p = .026$.

For each 1-year increase in experience, the salary increase was approximately 1.44 thousand dollars for both females and males. This slope for the prediction of salary from years of experience was statistically significant: $t(47) = 10.83$, $p < .001$. The graph in Figure 12.14 illustrates the regression lines to predict salary for males and females separately. The intercept (i.e., predicted salary for 0 years of experience) was significantly higher for males than for females.

The squared semipartial correlation for years as a predictor of salary was $sr^2 = .56$; thus, years of experience uniquely predicted about 56% of the variance in salary (when gender was statistically controlled). The squared semipartial correlation for gender as a predictor of salary was $sr^2 = .03$; thus, gender uniquely predicted about 3% of the variance in salary (when years of experience was statistically controlled). The results of this analysis suggest that there was a systematic difference between salaries for male and female faculty and that this difference was approximately the same at all levels of years of experience. Statistically controlling for years of job experience, by including it as a predictor of salary in a regression that also used gender to predict salary, yielded results that suggest that the overall gender difference in mean salary was partly, but not completely, accounted for by gender differences in years of job experience.

12.12 ♦ Summary

This chapter presented examples that demonstrated the equivalence of ANOVA and regression analyses that use dummy variables to represent membership in multiple groups. This discussion has presented demonstrations and examples rather than formal proofs; mathematical statistics textbooks provide formal proofs of equivalence between ANOVA and regression. ANOVA and regression are different special cases of the GLM.

If duplication of ANOVA using regression were the only application of dummy variables, it would not be worth spending so much time on them. However, dummy variables have important practical applications. Researchers often want to include group membership variables (such as gender) among the predictors that they use in multiple regression, and it is important to understand how the coefficients for dummy variables are interpreted.

An advantage of choosing ANOVA as the method for comparing group means is that the SPSS procedures provide a wider range of options for follow-up analysis—for example, post hoc protected tests. Also, when ANOVA is used, interaction terms are generated automatically for all pairs of (categorical) predictor variables or factors, so it is less likely that a researcher will fail to notice an interaction when the analysis is performed as a factorial ANOVA (as discussed in Chapter 13) than when a comparison of group means is performed using dummy variables as predictors in a regression. ANOVA does not assume a linear relationship between scores on categorical predictor variables and scores on quantitative outcome variables. A quantitative predictor can be added to an ANOVA model (this type of analysis, called analysis of covariance or ANCOVA, is discussed in Chapter 15).

On the other hand, an advantage of choosing regression as the method for comparing group means is that it is easy to use quantitative predictor variables along with group membership predictor variables to predict scores on a quantitative outcome variable. Regression analysis yields equations that can be used to generate different predicted scores for cases with different score values on both categorical and dummy predictor variables. A possible disadvantage of the regression approach is that interaction terms are not automatically included in a regression; the data analyst must specifically create a new variable (the product of the two variables involved in the interaction) and add that new variable as a predictor. Thus, unless they specifically include interaction terms in their models (as discussed in Chapter 15), data analysts who use regression analysis may fail to notice interactions between predictors. A data analyst who is careless may also set up a regression model that is "nonsense"; for example, it would not make sense to predict political conservatism (Y) from scores on a categorical X_1 predictor variable that has codes 1 for Democrat, 2 for Republican, and 3 for Independent. Regression assumes a linear relationship between predictor and outcome variables; political party represented by just one categorical variable with three possible score values probably would not be linearly related to an outcome variable such as conservatism. To compare group means using regression in situations where there are more than two groups, the data analyst needs to create dummy variables to represent information about group membership. In some situations, it may be less convenient to create new dummy variables (and run a regression) than to run an ANOVA.

Ultimately, however, ANOVA and regression with dummy predictor variables yield essentially the same information about predicted scores for different groups. In many research situations, ANOVA may be a more convenient method to assess differences among group means. However, regression with dummy variables provides a viable alternative, and in some research situations (where predictor variables include both categorical and quantitative variables), a regression analysis may be a more convenient way of setting up the analysis.

Note

1. Unequal numbers in groups make the interpretation of b coefficients for dummy variables more complex. For additional information about issues that should be considered when using dummy or effect codes to represent groups of unequal sizes, see Hardy (1993).

Comprehension Questions

1. Suppose that a researcher wants to do a study to assess how scores on the dependent variable heart rate (HR) differ across groups that have been exposed to various types of stress. Stress group membership was coded as follows:

   Group 1, no stress/baseline

   Group 2, mental arithmetic

   Group 3, pain induction

   Group 4, stressful social role play

   The basic research questions are whether these four types of stress elicited significantly different HRs overall and which specific pairs of groups differed significantly.

   a. Set up dummy-coded dummy variables that could be used to predict HR in a regression.

      Note that it might make more sense to use the "no-stress" group as the one that all other group means are compared with, rather than the group that happens to be listed last in the list above (stressful role play). Before working out the contrast coefficients, it may be helpful to list the groups in a different order:

      Group 1, mental arithmetic

      Group 2, pain induction

      Group 3, stressful social role play

      Group 4, no stress/baseline

      Set up dummy-coded dummy variables to predict scores on HR from group membership for this set of four groups.

      Write out in words which contrast between group means each dummy variable that you have created represents.

   b. Set up effect-coded dummy variables that could be used to predict HR in a regression.

      Describe how the numerical results for these effect-coded dummy variables (in 1b) differ from the numerical results obtained using dummy-coded dummy variables (in 1a). Which parts of the numerical results will be the same for these two analyses?

   c. Set up the coding for orthogonally coded dummy variables that would represent these orthogonal contrasts:

      Group 1 versus 2

      Groups 1 and 2 versus 3

      Groups 1, 2, and 3 versus 4

2. Suppose that a researcher does a study to see how level of anxiety (A_1 = low, A_2 = medium, A_3 = high) is used to predict exam performance (Y). Here are hypothetical data for this research situation. Each column represents scores on Y (exam scores).

A_1, Low Anxiety	A_2, Medium Anxiety	A_3, High Anxiety
72	86	65
81	93	79
54	81	74
66	80	80
71	92	74

a. Would it be appropriate to do a Pearson correlation (and/or linear regression) between anxiety, coded 1, 2, 3 for (low, medium, high), and exam score? Justify your answer.
b. Set up orthogonally coded dummy variables (O_1, O_2) to represent linear and quadratic trends, and run a regression analysis to predict exam scores from O_1 and O_2. What conclusions can you draw about the nature of the relationship between anxiety and exam performance?
c. Set up dummy-coded dummy variables to contrast each of the other groups with Group 2, medium anxiety; run a regression to predict exam performance (Y) from these dummy-coded dummy variables.
d. Run a one-way ANOVA on these scores; request contrasts between Group 2, medium anxiety, and each of the other groups. Do a point-by-point comparison of the numerical results for your ANOVA printout with the numerical results for the regression in (2c), pointing out where the results are equivalent.

3. Why is it acceptable to use a dichotomous predictor variable in a regression when it is not usually acceptable to use a categorical variable that has more than two values as a predictor in regression?
4. Why are values such as +1, 0, and –1 generally used to code dummy variables?
5. How does the interpretation of regression coefficients differ for dummy coding of dummy variables versus effect coding of dummy variables? (Hint: in one type of coding, b_0 corresponds to the grand mean; in the other, b_0 corresponds to the mean of one of the groups.)
6. If you have k groups, why do you only need $k - 1$ dummy variables to represent group membership? Why is it impossible to include k dummy variables as predictors in a regression when you have k groups?
7. How does orthogonal coding of dummy variables differ from dummy and effect coding?
8. Write out equations to show how regression can be used to duplicate a t test or a one-way ANOVA.