8 One-Way ANOVA

Models with a Single Categorical Predictor

To this point we have relied on regression models where the predictor variables have been treated as continuous variables. Our purpose in this chapter and the following two is to examine our basic approach to data analysis when predictors are categorical variables. In the language of traditional statistics books, earlier chapters concerned multiple regression. The present chapter concerns one-way analysis of variance (ANOVA) models or, equivalently, models with a single categorical predictor. In the next chapter we consider models having multiple categorical predictors, or higher order ANOVA models. Chapter 10 is devoted to models in which some predictors are categorical and some are continuous. Such models have been traditionally labeled analysis of covariance models. By integrating these into a common approach, we will not only explore these traditional topics but also consider others that considerably extend the sorts of questions that we are able to ask of our data, in the context of categorical predictor variables. Throughout we will continue to use our basic approach to statistical inference, testing null hypotheses by the comparison of augmented and compact models.

THE CASE OF A CATEGORICAL PREDICTOR WITH TWO LEVELS

Figure 8.1 contains hypothetical data from a study in which the impact of a SAT training course was evaluated. Twenty high-school seniors were randomly assigned to either take the 2-week training course, designed to improve SAT performance, or to a control no-course condition. As we can see, 10 students wound up in each of the two groups. At the end of the 2-week period, all 20 students took the SAT test and their performance was recorded. What we would like to do is examine whether the course made a difference in subsequent performance. Our question thus is whether we can reliably predict subsequent SAT performance as a function of whether a student was in the Course group or the No Course control group.

If we are to tell the computer to specify a model in which SAT performance is predicted by which of the two groups a student was in, we need some way of coding or numerically representing the group variable. This variable is categorical rather than numerical or continuous, meaning that students in the two groups differ on whether or not they have taken the course, but no automatic numerical representation of that difference is implied. Such variables require some coding scheme to represent them numerically, so that they can be used as predictor variables in models. It turns out that any numerical representation of this group variable would do, as long as we used that

numerical representation consistently. By consistent use, we mean that if a given value on the variable that represents group (Course versus No Course) numerically is assigned to one group, then every student in that group has that same value on the variable, and no student in the other group has that value.

FIGURE 8.1 Two-group SAT data

Student	*Group*	*SAT*
1	Course	580
2	Course	560
3	Course	660
4	Course	620
5	Course	600
6	Course	580
7	Course	590
8	Course	640
9	Course	620
10	Course	600
11	No Course	580
12	No Course	530
13	No Course	590
14	No Course	550
15	No Course	610
16	No Course	590
17	No Course	600
18	No Course	530
19	No Course	590
20	No Course	600

To illustrate, suppose we created a variable X_i to represent group numerically, arbitrarily assigning the value of –1 to students in the No Course group and +1 to students in the Course group. Since every student is in one group or the other, all students have values of either –1 or +1 on variable X_i. Notice that our purpose in creating this variable is simply to differentiate numerically between the two groups. Since group is a categorical variable, no rank order or interval information need be preserved in our coding scheme. We could just as easily have given the value of –1 to the Course group and +1 to the No Course group. Similarly, we could have given the value of 203 to students in the Course group and the value of –20.5 to students in the No Course group. The point is that the values that represent the categorical variable are arbitrarily defined, but they must be consistently used.

Throughout all of the rest of the book we will use a convention for coding nominal predictors known as *contrast codes*. Contrast codes are simply one of the possible arbitrary coding schemes for numerically representing categorical predictors. Two conditions define contrast codes and differentiate them from other coding schemes. For now, we will only define one of these two conditions. The other is only relevant when the categorical variable has more than two categories and will be given later in this chapter. Let us define a value on a contrast-coded categorical variable X_i as λ_k ("lambda$_k$"), where the subscript k refers to the level of the categorical variable being coded. In this case, k refers to the two levels of the group variable: Course versus No Course. Across levels of k, or across all categories of the variable, a contrast code is one where:

$$\sum_k \lambda_k = 0$$

Notice that we are summing here across levels or categories rather than across individual observations. In other words, the condition is that the values of the contrast variable sum to zero across the two categories, not across the individual observations in those two categories.

In our example, the values of +1 for students in the Course group and –1 for students in the No Course group constitute values of a contrast-coded variable, since the sum of these two values across the two categories equals zero. Another valid contrast-coded variable would use values of +.5 for the Course group and –.5 for the No Course group. Notice, however, that the values of 203 for the Course group and –20.5 for the No Course group do not meet the condition for a contrast-coded variable. Note also that in this example the values of a contrast-coded variable, say +1 and –1, sum to zero not only

across the two categories but also across the 20 students in those two categories. This will be the case when a contrast-coded variable is used and when there are an equal number of observations in the two groups or categories. Had we had more students in one of the two groups than in the other, then the sum of the values across the two groups would be zero, but the sum of the values across all the students would not be.

In the following sections, we will use two different contrast-coded predictors to predict SAT performance with the data presented in Figure 8.1. We will first use the values of +1 and –1 and then we will use the values of +.5 and –.5. These possible values of a contrast-coded predictor are simply two from an infinite number of such values that might be used. At a later point, we will also briefly discuss the estimation of models using coding conventions other than contrast codes.

Model Estimation and Inference with a Contrast-Coded Predictor

We start by estimating a model in which we predict SAT performance with a contrast-coded predictor having values of +1 for students in the Course group and –1 for students in the No Course group. SAT is thus our Y_i variable and our predictor, X_i, is the contrast-coded predictor. Our model is the simple regression model with a single predictor variable of Chapter 5. We will want to compare this model, making predictions of SAT conditional on X_i, with a compact one in which we predict the same value for every student regardless of whether they were in the Course group or the No Course group:

MODEL A: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

MODEL C: $Y_i = \beta_0 + \varepsilon_i$

Assuming that $\beta_1 \neq 0$, the augmented model makes conditional predictions of SAT performance, conditional on whether students were in the Course group or the No Course group. On the other hand, the compact model makes the same prediction for all students, regardless of their group.

The least-squares estimates for these models are:

MODEL A: $\hat{Y}_i = 591 + 14X_i$

MODEL C: $\hat{Y}_i = 591$

Both of these models, as well as the data on which they are based, are graphed in Figure 8.2. The horizontal prediction function is Model C, while the prediction line that makes differential predictions is Model A. The sums of squared errors of these two models are 16,060 for Model A and 19,980 for Model C.

The calculations of these two sums of squares are shown in Figure 8.3 where we derive for each observation the predicted value from each model, the residual, and the squared residual. The sums of these squared residuals, given in the last row of Figure 8.3, are the sums of squared errors for the two models.

The comparison of these two models, asking whether the predictions conditional on the contrast-coded predictor do a better job than the unconditional predictions, yields:

$$\text{PRE} = \frac{19{,}980 - 16{,}060}{19{,}980} = \frac{3920}{19{,}980} = .196$$

FIGURE 8.2 Models A and C for the SAT data

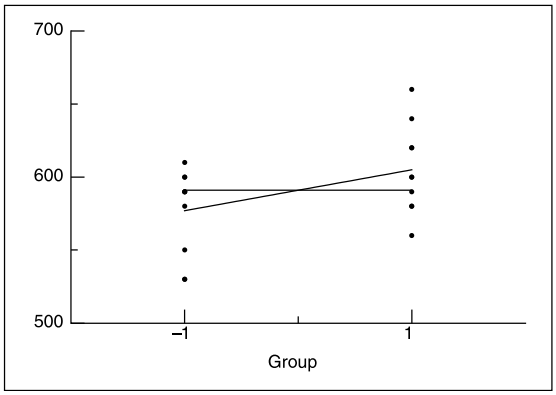

FIGURE 8.3 Calculation of SSE for Models C and A

				Compact			*Augmented*		
Student	*Group*	X_i	*SAT* (Y_i)	$\hat{Y}_i$	e_i	e_i^2	$\hat{Y}_i$	e_i	e_i^2
1	Course	1	580	591	–11	121	605	–25	625
2	Course	1	560	591	–31	961	605	–45	2025
3	Course	1	660	591	69	4761	605	55	3025
4	Course	1	620	591	29	841	605	15	225
5	Course	1	600	591	9	81	605	–5	25
6	Course	1	580	591	–11	121	605	–25	625
7	Course	1	590	591	–1	1	605	–15	225
8	Course	1	640	591	49	2401	605	35	1225
9	Course	1	620	591	29	841	605	15	225
10	Course	1	600	591	9	81	605	–5	25
11	No Course	–1	580	591	–11	121	577	3	9
12	No Course	–1	530	591	–61	3721	577	–47	2209
13	No Course	–1	590	591	–1	1	577	13	169
14	No Course	–1	550	591	–41	1681	577	–27	729
15	No Course	–1	610	591	19	361	577	33	1089
16	No Course	–1	590	591	–1	1	577	13	169
17	No Course	–1	600	591	9	81	577	23	529
18	No Course	–1	530	591	–61	3721	577	–47	2209
19	No Course	–1	590	591	–1	1	577	13	169
20	No Course	–1	600	591	9	81	577	23	529
					SSE(C) = 19,980			SSE(A) = 16,060	

Model A contains two parameters and Model C one, hence PA – PC equals 1 and n – PA equals 18. Accordingly, we can compute the F statistic for this comparison either from the computed value of PRE or from the values of the sums of squares:

$$F_{1,18} = \frac{\text{PRE}/(\text{PA} - \text{PC})}{(1 - \text{PRE})/(n - \text{PA})} = \frac{.196/1}{(1 - .196)/18} = 4.39$$

$$F_{1,18} = \frac{\text{SSR}/(\text{PA} - \text{PC})}{\text{SSE(A)}/(n - \text{PA})} = \frac{3920/1}{16{,}060/18} = 4.39$$

These statistics fall just short of their critical values, with 1 and 18 degrees of freedom, with α set at .05. Hence, we are unable to conclude that the conditional predictions of SAT are significantly better than the unconditional one made by Model C.

So far, once we have coded our categorical predictor, there is nothing different about this simple regression model from those simple models using continuous predictors that were discussed in Chapter 5. The estimation of the model parameters and the calculations of PRE and F proceed just as before. Since all of the assumptions underlying the use of the critical values of PRE and F involve assumptions about the distribution of Y_i, or really of ε_i, and since the categorical nature of X_i has no effect on the distribution of ε_i, none of the assumptions underlying this analysis have been impacted by the use of the categorical predictor.

What has changed somewhat, however, is the interpretation of the estimated parameters of the model and, accordingly, the interpretation of the statistical inference results. It is not the case that the old interpretations (those developed in Chapter 5) are incorrect, for the value of the intercept in the augmented model, 591, is still the predicted value of Y_i when X_i equals zero. The coefficient for X_i, 14, is still a slope—the amount by which the predicted value of Y_i changes for each unit increase in X_i. And PRE and F still tell us about the reduction in errors of prediction. Rather, when we have categorical predictors, new interpretations of these statistics become possible.

To understand these new interpretations, it is helpful to consider the predicted values that the augmented model makes. These are contained in Figure 8.3. If we are dealing with a student in the Course group, the predicted value from the augmented model is:

$$\hat{Y}_{+1} = 591 + 14(+1) = 605$$

And if we are dealing with a student in the No Course group, the predicted value from the augmented model is:

$$\hat{Y}_{-1} = 591 + 14(-1) = 577$$

These predicted values turn out to be the mean SAT scores of the 10 students in each of the two groups. That is, 605 is the mean SAT score of those students who received the course, and 577 is the mean SAT score of those students who did not. And given our use of +1 and –1 as the values of the contrast-coded predictor, the slope associated with that predictor equals half the difference between the means of the two groups:

$$\begin{aligned}\bar{Y}_C - \bar{Y}_{NC} &= \hat{Y}_{+1} - \hat{Y}_{-1} = 605 - 577 \\ &= (591 + 14(+1)) - (591 + 14(-1)) \\ &= 14(+1) - 14(-1) \\ &= 2(14)\end{aligned}$$

In general, the least-squares parameter estimate or slope associated with a contrast-coded predictor is given by:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

which, in the example at hand, is evaluated as:

$$\frac{(-1)577 + (+1)605}{(-1)^2 + (+1)^2} = \frac{28}{2} = 14$$

This is a very general and useful formula for interpreting the slope associated with any contrast-coded predictor. The numerator represents a comparison between group means, in this case the difference between the mean for the Course group and that for the No Course group, and the denominator is a scaling factor dependent on the specific values used for the contrast-coded predictor. The important point is that the regression coefficient associated with any contrast-coded predictor tells us about the difference between group means, the direction of that difference being determined by which group is coded with a positive value and which group is coded with a negative value.

The estimated intercept of 591 equals, as always, the predicted value of Y_i when X_i equals zero. Since X_i equals zero halfway between the two values of +1 and –1 that code the two groups, the estimated value of the intercept is necessarily equal to the average of the two group means. This result is made clear by the graph of the model in Figure 8.2. It can also be shown algebraically as follows:

$$\hat{Y}_{+1} = \bar{Y}_C = 591 + 14(+1)$$

$$\hat{Y}_{-1} = \bar{Y}_{NC} = 591 + 14(-1)$$

Adding these two equalities gives:

$$\bar{Y}_C + \bar{Y}_{NC} = (591 + 14(+1)) + (591 + 14(-1))$$
$$\bar{Y}_C + \bar{Y}_{NC} = (2)591$$
$$\frac{\bar{Y}_C + \bar{Y}_{NC}}{2} = 591$$

Notice that this interpretation of the intercept in the augmented model, including the contrast-coded predictor, is not the same as the interpretation of the intercept in the compact model, the one making unconditional predictions. The intercept in the compact model is estimated as the mean of all the observations, what we might call the grand mean, $\bar{Y}$, defined as:

$$b_{0C} = \bar{Y} = \frac{\sum_i Y_i}{n}$$

On the other hand, the intercept in the augmented model is the mean of the group means, defined as:

$$b_{0A} = \frac{\sum_k \bar{Y}_k}{m}$$

where m is the total number of groups, in this case 2.

In the dataset that we have been using, the values of these two intercepts, one from the compact model and one from the augmented, are identical, i.e., 591, because there

are an equal number of students in the two groups. In general, however, they estimate different things. The intercept in the compact model, the one making unconditional predictions, is the mean of all the observations. The intercept in the augmented model, the one making predictions conditional on group, is the mean of the group means.

One last result when using contrast-coded predictors is important to know. Just as the regression coefficient for any contrast-coded predictor can be represented as a comparison among group means:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

so too can the SSR associated with any such predictor be similarly expressed. As always, the SSR associated with a predictor is the difference between the SSE(A) and SSE(C) for compact and augmented models with and without that predictor. And, as always, that SSR equals:

$$\text{SSR} = \sum(\hat{Y}_{iA} - \hat{Y}_{iC})^2$$

In the case of a categorical predictor, as we have seen, the predicted values for the augmented model in this expression are the group or category means, $\bar{Y}_k$, and the predicted value from the compact model is the grand mean of all the observations, $\bar{Y}$. As a result, it is possible to show that the SSR associated with any contrast-coded predictor can be expressed as a function of the category means (and the number of observations in each group, n_k) as follows:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)}$$

In the case at hand, we have seen that SSE(C) equals 19,980 and SSE(A) equals 16,060, resulting in an SSR associated with the contrast-coded predictor of 3920. That is equivalently computed as:

$$\frac{((-1)577 + (+1)605)^2}{((-1)^2/10) + ((+1)^2/10)} = 3920$$

Estimation with Alternative Values for the Contrast-Coded Predictor

If we had defined the values of the contrast-coded predictor as +.5 for the Course group and –.5 for the No Course group, rather than +1 and –1, the estimated model would be:

$$\text{MODEL A: } \hat{Y}_i = 591 + 28X_i'$$

where X_i' is the new contrast-coded predictor. Importantly, this model makes exactly the same predictions for students in the two groups, i.e., their respective group means:

$$\hat{Y}_{+.5} = 591 + 28(+.5) = 605 = \bar{Y}_C$$

$$\hat{Y}_{-.5} = 591 + 28(-.5) = 577 = \bar{Y}_{NC}$$

Accordingly, in a deep sense, it is the same augmented model with the same sum of squared errors. The slope for the contrast-coded predictor now equals 28 instead of 14, since now there is a one-unit difference on X_i– *that separates the two groups* (between –.5 and +.5) rather than the two-unit difference that separated them on X_i (between +1 and –1). And that slope now equals the difference between the two group means:

$$b_1 = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} = \frac{(-.5)577 + (+.5)605}{(-.5)^2 + (+.5)^2} = \frac{.5(605 - 577)}{.5} = 605 - 577 = 28$$

Of course, since this Model A is in a deep sense the same as the one where the contrast-coded predictor had values of +1 and –1, the model comparison of it with the compact model, making predictions that were not conditional on group, yields the exact same SSR, PRE and *F* statistics. Recomputing the SSR for this comparison, using these new codes and the formula given for the SSR in terms of the group means, gives us:

$$\frac{((-.5)577 + (+.5)605)^2}{((-.5)^2/10) + ((+.5)^2/10)} = 3920$$

Equivalence with ANOVA and Two-Group *t*-Test

The slope in the augmented model, making conditional predictions, informs us about the difference between the two group means (regardless of the values of λ_k used to construct the contrast-coded predictor). Therefore, the comparison of this augmented model with the compact one making unconditional predictions asks both whether the parameter associated with the contrast-coded predictor departs from zero and whether the two group means differ from each other. In other words, the model comparison we have examined addresses the following equivalent null hypotheses:

H_0: $\beta_1 = 0$

H_0: $\mu_C = \mu_{NC}$

where μ_C and μ_{NC} are the true but unknown means of the two groups.

In more traditional statistical textbooks, the test of a null hypothesis about the difference between two group means is usually conducted by computing a two-group ANOVA or a two-group independent samples *t*-test. It is therefore important to show that our model comparison and its associated PRE and *F* statistics are identical to those generated by these more traditional approaches.

In Figure 8.4 we give the ANOVA source table for the model comparison we have just conducted, using the formulas developed in Chapter 5 and earlier. We also provide the source table using the numeric values generated by our data example.

The formula for the SSR for the model comparison that we have given before is:

$$\text{SSR} = \sum_i (\hat{Y}_A - \hat{Y}_C)^2$$

FIGURE 8.4 ANOVA source tables

Source	*SS*	*df*	*MS*	*F*	*p*	*PRE*
Reduction	SSR	PA – PC	MSR	$\frac{\text{MSR}}{\text{MSE(A)}}$		$\frac{\text{SSR}}{\text{SSE(C)}}$
Error	SSE(A)	n – PA	MSE(A)			
Total	SSE(C)	n – 1	MSE(C)			

Source	*SS*	*df*	*MS*	*F*	*p*	*PRE*
Reduction	3920	1	3920.00	4.39	< .10	.196
Error	16060	18	892.22			
Total	19980	19	1051.58			

summing across all individual observations. In the present case, $\hat{Y}_C$ is the mean of all of the observations, $\bar{Y}$, and the predicted values from the augmented model, $\hat{Y}_A$, are the two group means $\bar{Y}_C$ and $\bar{Y}_{NC}$. Let us generically represent these group means as $\bar{Y}_k$, thereby indicating the group mean for the *k*th group. Then we can write the above expression for the sum of squares reduced as:

$$\text{SSR} = \sum_i (\hat{Y}_A - \hat{Y}_C)^2 = \sum_i (\bar{Y}_k - \bar{Y})^2 = \sum_k n_k (\bar{Y}_k - \bar{Y})^2$$

where n_k is the number of observations in the *k*th group.

We can also re-express the formula for SSE(A) and SSE(C) in terms of means, since those are the predicted values from each model:

$$\text{SSE(A)} = \sum_i (Y_i - \hat{Y}_{iA})^2 = \sum_i (Y_i - \bar{Y}_k)^2$$

$$\text{SSE(C)} = \sum_i (Y_i - \hat{Y}_{iC})^2 = \sum_i (Y_i - \bar{Y})^2$$

Accordingly, the formulas in the source table that we have used all along for summarizing our computations of PRE and *F* (given in the top half of Figure 8.4) can, in this case, be equivalently written with the formulas used to compute an analysis of variance to compare group means in traditional statistics textbooks. This revised version of the source table is given in Figure 8.5.

The names given to the rows in this version of the source table have been changed to reflect those traditionally used in analysis of variance. So, the sum of squares reduced and its mean square are traditionally called the sum of squares and mean square *between groups*, and the sum of squares and mean square from Model A are traditionally called the sum of squares and mean square *within groups*. But fundamentally and algebraically this is the same source table, yielding the exact same *F* and PRE, as the one we are more used to, given in Figure 8.4.

The square root of *F* is the *t* statistic (since n – PA =1), with $n - 2$ degrees of freedom, that is traditionally called the two-group independent samples *t*-test.

All of this is simply to demonstrate that our integrated approach to statistical inference, resting on model comparisons estimated with any least-squares multiple

FIGURE 8.5 ANOVA source table

Source	*SS*	*df*	*MS*	*F*	*p*	*PRE*
Between	$\sum_k n_k (\bar{Y}_k - \bar{\bar{Y}})^2$	1	MSB	$\frac{\text{MSB}}{\text{MSW}}$		$\frac{\text{SSB}}{\text{SST}}$
Within	$\sum_i (Y_i - \bar{Y}_k)^2$	$n - 2$	MSW			
Total	$\sum_i (Y_i - \bar{\bar{Y}})^2$	$n - 1$	MST			

regression program, yields the same results as the cookbook recipes given in more traditional statistical textbooks. Our model comparison in this case, testing whether the regression coefficient for a single contrast-coded predictor departs from zero, is exactly equivalent to a two-sample *t*-test for examining whether the means of two groups differ from each other.

Confidence Interval for the Slope of a Contrast-Coded Predictor

The formula that we gave in Chapter 6 for the confidence interval for a regression coefficient continues to be applicable in the situation where predictor variables are contrast-coded. The confidence interval for the slope associated with any predictor variable was given there as:

$$b \pm \sqrt{\frac{F_{\text{crit}}\,\text{MSE}}{(n-1)s_X^2\,(tol)}}$$

Since the slope of a contrast-coded predictor informs us about the magnitude of the difference between group means, so its confidence interval also informs us about the confidence interval associated with that mean difference. To see this, let us take the case where the contrast-coded predictor used values of –.5 for the No Course group and +.5 for the Course group. The resulting slope in this case equaled 28 and the confidence interval for that slope is calculated as:

$$28 \pm \sqrt{\frac{4.41(892.22)}{19(0.263)1}}$$
$$28 \pm 28.06$$

where 4.41 is the critical *F* value with 1 and 18 degrees of freedom, 892.22 is the mean square error from our Model A, $n - 1$ equals 19, the variance of the 20 individuals on the contrast-coded predictor is 0.263, and its tolerance is of course 1 since it is the only predictor in the model. This confidence interval can also be written as:

$$-0.06 \leq \beta_1 \leq 56.06$$

Since the parameter that is being estimated here, with this contrast-coded predictor, is also an estimate of the true mean difference between the two groups, this confidence interval can be equivalently expressed as

$$-0.06 \leq \mu_C - \mu_{NC} \leq 56.06$$

When the contrast-coded predictor used the values of –1 and +1 rather than –.5 and +.5, the estimated slope was half the difference in the group means (i.e., 14) and its confidence interval is computed as:

$$14 \pm \sqrt{\frac{4.41(892.22)}{19(1.052)1}}$$

$$14 \pm 14.03$$

The term that is different in this interval (other than the value of the estimated slope itself) is the variance of the predictor (coded +1 and –1), which is 1.052 rather than 0.263.

In this case, the confidence interval is given equivalently as:

$$-0.03 \leq \beta_1 \leq 28.03$$

$$-0.03 \leq \frac{\mu_C - \mu_{NC}}{2} \leq 28.03$$

Thus, it continues to tell us about the confidence interval for the mean difference, except with these codes of course it is the confidence interval for half the mean difference. If we multiply this expression by 2, we get the confidence interval for the mean difference.

CATEGORICAL PREDICTORS WITH MORE THAN TWO LEVELS

Suppose a developmental psychologist is interested in the effects of feedback about performance on subsequent motivation to do a task. She hypothesizes that subsequent motivation will decline if children are told that they earlier failed at the task. To test this hypothesis, she randomly assigns children to three conditions; in one condition they are told that they failed on the task; in a second condition they are given no feedback; and in a third condition they are told they succeeded. The experimenter then monitors the number of tasks they subsequently complete, after the differential feedback has been given. Twenty-four children are run in total, eight in each of the three conditions. The hypothetical raw data are given in Figure 8.6.

FIGURE 8.6 Hypothetical experimental data for three conditions (values represent number of tasks each subject completes)

	Failure	*No Feedback*	*Success*
	2	4	4
	2	3	6
	2	4	5
	3	5	4
	4	5	6
	4	2	4
	3	4	3
	4	3	3
$\bar{Y}_k$	3.000	3.750	4.375

Contrast Codes for Multilevel Categorical Predictors

In order to examine the effects of feedback on the number of tasks subsequently completed, we need to derive a coding scheme to represent the three levels of the categorical feedback variable. We might think that a single variable that codes all three conditions would be appropriate, giving observations from the Success condition a higher value on the variable than observations for the No Feedback condition who in turn receive a higher value than observations from the Failure condition. We could then see if such a coded variable would be predictive of Y_i. The problem with using a single variable to code the three levels of this categorical variable is that with such a coding scheme we are assuming that the categories can be ordered in an a priori manner and that the relationship between the values of the single predictor variable and Y_i is a linear one. While we may have a reason for expecting that Y_i should be lower in the Failure condition than in the other two, we do not have any reason for assigning particular values to the groups, expecting linear predictions as a function of those particular values. In other words, a single-predictor variable that codes the three conditions with particular values does not make much sense, given that we are dealing with a categorical variable whose levels do not differ in a neat linear way.

To ask whether we can predict Y_i as a function of some categorical variable having in general m levels or groups is equivalent to asking whether there are differences among the m group means ($\bar{Y}_k$) across those levels (with k varying from 1 to m). To answer this question, we need to employ $m - 1$ contrast-coded predictor variables in our model. We will then be able to ask about mean differences among the groups, allowing for all possible orderings of those means. To define these $m - 1$ contrast-coded predictors, it is now time to introduce the second defining condition for contrast codes. The first condition, you will recall, was that for a contrast-coded variable the sum of the λ values across the groups or levels of the categorical variable must equal zero: $\sum_k \lambda_k = 0$. When we use more than a single contrast-coded predictor to code a categorical variable having more than two levels, the second condition that must be met is that across levels of the categorical variable all pairs of contrast codes must be orthogonal to each other. Given that the first condition is met, this second condition of orthogonality will be met whenever the sum (across k or the levels of the categorical variable) of the products of the λ values from pairs of contrast codes equals zero.

In our example, we have three levels of the categorical variable. We will therefore use two contrast codes to code it. Each value of λ now has two subscripts, the first one designating which contrast code we are talking about and the second one designating the level of the categorical variable (k). The condition of orthogonality is met when:

$$\sum_k \lambda_{1k} \lambda_{2k} = 0$$

To make this second defining condition of contrast codes more understandable, let us illustrate codes that do and do not meet it for the example at hand. In Figure 8.7, two sets of codes, with two codes in each set, are given for coding the three levels of the categorical predictor variable: Failure, No Feedback, and Success.

Each of the four codes meets the first defining condition for a contrast code, in that the sum of the λ values for any given code, computed across the three levels of the categorical variable, equals zero. The second defining condition, however, is only met

FIGURE 8.7 Sets of codes for a three-level categorical predictor

	Failure	*No Feedback*	*Success*
Set A			
λ_{1k}	–2	1	1
λ_{2k}	0	–1	1
Set B			
λ_{1k}	–1	0	1
λ_{2k}	0	–1	1

by the codes in Set A. If we multiply the value of λ_{1k} by the value of λ_{2k} at each of the levels of the categorical predictor variable and then we add up the resulting three products, we get a sum of 0 from Set A (i.e., 0 + (–1) + 1 = 0) and a sum of 1 from Set B (i.e., 0 + 0 + 1 = 1). Accordingly, only the codes in Set A can legitimately be called contrast codes.

This second defining condition means that a given code cannot be defined as a contrast code in isolation. We could not, for instance, look at the code for λ_{1k} in Set A and identify it as a contrast code, unless we looked at the other code or codes with which it is used in combination to code the categorical predictor variable. For instance, if we changed the values of λ_{2k} in Set A to be –1, –1, and 2 for Failure, No Feedback, and Success, respectively, then the codes in Set A would no longer be contrast codes, even though we had not changed the values of λ for the first code. This set of codes would no longer be contrast codes since the sum of the products of the λ values across the category levels would no longer equal zero.

If our categorical predictor variable had four levels, we would need three contrast codes to code it completely. The second defining condition for contrast codes would be met in such a case if the sums of the products of the λ values for all possible pairs of codes equaled zero. Suppose, for instance, that we had a categorical variable with four levels, as in Figure 8.8. There we define three contrast codes with values of λ_{1k}, λ_{2k}, and λ_{3k}. We then have three pairs of codes, and for each of these pairs the sum of the products of the λ values must equal zero. For codes 1 and 2, the sum of the products of the λ_{jk} values equals (–3)0 + 1(–2) + 1(1) + 1(1) = 0. For codes 1 and 3, the sum of the products of the λ_{jk} values equals (–3)0 + 1(0) + 1(–1) + 1(1) = 0. And for codes 2 and 3, the sum of the products of the λ_{jk} values equals 0(0) + 0(–2) + 1(–1) + 1(1) = 0.

With a categorical predictor having three levels, then, we need two contrast codes and a single sum of products of λ values must equal zero. With a categorical predictor having four levels, we need three contrast codes. Those three codes result in three possible pairs of codes, and hence three sums of products of λ values must equal zero. In general, with a categorical variable having m levels, we need $m - 1$ contrast codes to code it. From these $m - 1$ contrast codes, there are $(m - 1)(m - 2)/2$ pairs of codes. This many sums of products of λ values must equal zero to meet the second defining condition of contrast codes.

For any given categorical predictor, there are an infinite number of sets of contrast codes that could be used. The choice of codes to be used should be guided by some theoretical or substantive notions about how the groups defined by the categorical predictor variable are expected to differ on the dependent variable. For instance, in the

FIGURE 8.8 Codes for a four-level categorical predictor

	Level 1	*Level 2*	*Level 3*	*Level 4*
λ_{1k}	–3	1	1	1
λ_{2k}	0	–2	1	1
λ_{3k}	0	0	–1	1

illustration at hand, we expected subjects in the Failure condition to have lower scores than subjects in the other two conditions. Since, as we saw in the case of a categorical predictor with only two levels, the regression coefficient for a contrast code tells us about the relative mean difference between observations having different values on the contrast code, it makes sense to derive a code that will allow us to examine this prediction about mean differences on the dependent variable. In other words, given that we want to see whether the observations in the Failure condition have lower scores than observations in the other two conditions, the first contrast code we gave in Set A of Figure 8.7 is one that we may well choose to examine.

As was the case with a single contrast-coded predictor that codes a categorical variable with two levels, the regression coefficient associated with a contrast-coded predictor in the case of a categorical variable with more than two levels tells us about mean differences among the various groups or levels of the categorical variable, according to the following formula:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

But this will be the case *only* if a complete set of $m - 1$ coded predictors are included in the model and *only* if the contrast codes used meet the orthogonality condition that we have just defined. At a later point in this chapter we will discuss estimation in the presence of nonorthogonally coded predictors. For now, the important point is that slopes tell us about coded mean differences among the categories only with a complete set of codes and only with orthogonality.

Based on theoretical considerations, then, we are interested in the comparison that is made by the first contrast code of Set A in Figure 8.7. With three levels of our categorical variable and one code chosen, the second code is constrained to be one that compares the means in the No Feedback and Success conditions, like the second code in Set A. In general, with m levels of a categorical variable and $m - 1$ contrast codes, the final code is constrained once the first $m - 2$ codes have been defined, in order to meet the orthogonality condition.

We can use these two codes to define two predictor variables, X_{1i} based on the codes –2, 1, 1 (for Failure, No Feedback, and Success respectively) and X_{2i} based on the codes 0, –1, 1, and then estimate a multiple regression model in which these are used as simultaneous predictors of Y_i. If we did this, we can exactly specify the mean differences estimated by the two resulting slopes using the formula for the slope of a contrast-coded predictor in the context of a model with a full set of orthogonal contrast-coded predictors:

$$b_{X_1} = \frac{\sum_k \lambda_{1k}\bar{Y}_k}{\sum_k \lambda_{1k}^2} = \frac{(-2)\bar{Y}_F + (1)\bar{Y}_{NF} + (1)\bar{Y}_S}{(-2)^2 + (1)^2 + (1)^2} = \frac{(\bar{Y}_{NF} + \bar{Y}_S) - 2\bar{Y}_F}{6} = \frac{\left(\frac{\bar{Y}_{NF} + \bar{Y}_S}{2}\right) - \bar{Y}_F}{3}$$

$$b_{X_2} = \frac{\sum_k \lambda_{2k}\bar{Y}_k}{\sum_k \lambda_{2k}^2} = \frac{(0)\bar{Y}_F + (-1)\bar{Y}_{NF} + (1)\bar{Y}_S}{(0)^2 + (-1)^2 + (1)^2} = \frac{\bar{Y}_S - \bar{Y}_{NF}}{2}$$

In general, such slopes will inform us about differences among category means following the codes used, with the numerator of the above expressions representing the mean difference, and the denominator representing a scaling factor. Notice that group means for levels of the categorical variable that are coded with a zero value of λ on a particular contrast-coded predictor drop out of the numerator of the slope and thus do not figure in the comparison that is made (i.e., the group mean for the Failure condition does not play a role in the slope of the second contrast-coded variable).

To show the impact of the scaling factor in the denominator of the slope expression, had we used fractional values for λ (–⅔, ⅓, ⅓ for X'_{1i}, and 0, –½, ½ for X'_{2i}) rather than those defined above, then the following would be the values of the slopes:

$$b_{X1}' = \frac{\sum_k \lambda_{1'k}\bar{Y}_k}{\sum_k \lambda_{1'k}^2} = \frac{(-\frac{2}{3})\bar{Y}_F + (\frac{1}{3})\bar{Y}_{NF} + (\frac{1}{3})\bar{Y}_S}{(-\frac{2}{3})^2 + (\frac{1}{3})^2 + (\frac{1}{3})^2} = \left(\frac{\bar{Y}_{NF} + \bar{Y}_S}{2}\right) - \bar{Y}_F$$

$$b_{X1}' = \frac{\sum_k \lambda_{2'k}\bar{Y}_k}{\sum_k \lambda_{2'k}^2} = \frac{(0)\bar{Y}_F + (-\frac{1}{2})\bar{Y}_{NF} + (\frac{1}{2})\bar{Y}_S}{(0)^2 + (-\frac{1}{2})^2 + (\frac{1}{2})^2} = \bar{Y}_S - \bar{Y}_{NF}$$

The advantage of such fractional codes is that their slopes will equal the mean differences rather than fractions of mean differences.

Without practice, it may seem difficult to come up with a set of orthogonal contrast codes, particularly when dealing with a categorical variable having more than three or so levels. Our advice is that one should initially create codes that represent mean comparisons one would like to make theoretically, and then derive the remainder of the codes so that orthogonality is preserved. One way to do this, once one or more initial codes have been defined, is to construct further contrast codes that compare category means that were tied (or received the same value of λ) on already used code(s). With some practice, deriving orthogonal codes becomes a relatively easy task.

In the absence of any motivated comparisons, one can always use a convention called Helmert codes, regardless of the number of levels. A simple algorithm generates such codes. If there are m levels of the categorical variable, one defines the first of $m - 1$ contrast codes by assigning the value of $m - 1$ to the first level and the value of –1 to each of the remaining $m - 1$ levels. For the second contrast code, the first level is given

FIGURE 8.9 Helmert contrast codes

	Category level						
Code	*1*	*2*	*3*	…	*m–2*	*m–1*	*m*
λ_{1k}	m–1	–1	–1	…	–1	–1	–1
λ_{2k}	0	m–2	–1	…	–1	–1	–1
λ_{3k}	0	0	m–3	…	–1	–1	–1
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
λ_{m-2k}	0	0	0	…	2	–1	–1
λ_{m-1k}	0	0	0	…	0	1	–1

the value of 0, the second level is given the value of $m - 2$, and all remaining levels are given the value of –1. For the third contrast code, the first two levels of the categorical predictor are assigned values of 0, the third level is given the value of $m - 3$, and the remaining levels are given the value of –1. One proceeds in this manner to define all $m - 1$ contrast codes, with the last one having values of 0 for all levels of the predictor variable except for the last two. These last two levels have values of 1 and –1. The resulting code values are presented in Figure 8.9.

Estimation and Inference with Multilevel Categorical Predictors

Using the data in Figure 8.6, we estimated the parameters of the following multiple regression model, with X_{1i} and X_{2i} as contrast-coded predictors, given the values of λ defined by Set A in Figure 8.7:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

The parameter estimates are:

$$\text{MODEL A: } \hat{Y}_i = 3.7083 + .3542 X_{1i} + .3125 X_{2i}$$

and the sum of squared errors is 23.375.

Unsurprisingly, the predicted values from this model are the means of the three categories (given in Figure 8.6) of the categorical independent variable:

$$\hat{Y}_F = 3.7083 + .3542(-2) + .3125(0) = 3.000$$

$$\hat{Y}_{NF} = 3.7083 + .3542(1) + .3125(-1) = 3.750$$

$$\hat{Y}_S = 3.7083 + .3542(1) + .3125(1) = 4.375$$

As we have said before, a model with a categorical independent variable will make predictions of the group or category level means whenever a complete set of $m - 1$ codes is used as predictors.

We have already discussed the interpretations of the two parameter estimates associated with the contrast-coded predictors in terms of the category means. Let us revisit these interpretations now that we have the numerical estimates:

$$b_{X_1} = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} = \frac{\left(\frac{\bar{Y}_{NF} + \bar{Y}_S}{2}\right) - \bar{Y}_F}{3} = \frac{\left(\frac{3.750 + 4.375}{2}\right) - 3.00}{3} = .3542$$

$$b_{X_2} = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2} = \frac{\bar{Y}_S - \bar{Y}_{NF}}{2} = \frac{4.375 - 3.750}{2} = .3125$$

And just as we found with a categorical predictor with two levels, the estimated intercept in this model equals the mean of the three category means:

$$b_0 = \frac{\sum_k \bar{Y}_k}{m} = \frac{3.000 + 3.750 + 4.375}{3} = 3.7083$$

Although these interpretations for the regression coefficients in models with contrast-coded predictors are typically the most useful, interpretations we gave earlier for parameter estimates in multiple regression models continue to be entirely appropriate. Thus slopes of a predictor can be interpreted as differences in $\hat{Y}_i$ values as the predictor increases by one unit, holding constant other predictors. In the case of the slope for X_{1i}, as we move from a score of –2 (for the Failure condition) to a score of +1 (for the No Feedback and Success conditions) the predicted values go from the mean of the Failure condition (3.000) to the means in the No Feedback and Success conditions (3.750 and 4.375). Thus, for a three-unit increase in X_{1i}, we go from a predicted value of 3.000 to one of 4.0625, meaning that the increase in predicted values for a one-unit increase in X_{1i} is .3542. And for X_{2i}, as we go from a score of –1 to 1, the predicted value goes from 3.75 to 4.375. Accordingly, per unit increase in X_{2i}, we predict a .3125 increase in $\hat{Y}_i$. And finally, the intercept equals the predicted value when both contrast-coded predictors equal zero. When do these predictors equal zero? From the first condition used to define contrast codes, the mean of each contrast code, across categories, equals zero. Accordingly, the intercept is the predicted value for the average of the categories.

There are, of course, many Model Cs with which we can compare this model to test various null hypotheses. One obvious comparison is with the single-parameter simplest model, estimating just the intercept:

MODEL C: $Y_i = \beta_0 + \varepsilon_i$

and predicting the grand mean, $\overline{Y}$, for all the observations. Since, in this example, each of the levels of the categorical variable has the same number of observations, the overall grand mean of the 24 observations is the same as the mean of the category means. Hence, it is the case that the estimated parameter in this Model C is identical to the intercept in the three-parameter Model A with which we are comparing it:

MODEL C: $\hat{Y}_i = 3.7083$

This estimated Model C has a sum of squared errors of 30.9583.

What exactly is the null hypothesis that is tested by this model comparison? Obviously it is that the two predictors have slopes of zero, that is, that using them as predictors does nothing to improve the quality of our predictions:

H_0: $\beta_1 = \beta_2 = 0$

But this null hypothesis can also be expressed in terms of the equality of the category means, since Model C predicts the grand mean, $\overline{Y}$, for every observation and Model A makes predictions that are conditional on category membership, predicting the category mean, $\overline{Y}_k$, for each observation. Accordingly, the null hypothesis can equivalently be expressed as:

$$H_0: \mu_{\mathrm{F}} = \mu_{NF} = \mu_S$$

where these are the true but unknown means of the three levels of the categorical independent variable.

The comparison of these two models yields the following values of PRE and F:

$$\mathrm{PRE} = \frac{30.9583 - 23.3750}{30.9583} = .245$$

$$F_{2,21} = \frac{\mathrm{PRE}/(\mathrm{PA} - \mathrm{PC})}{(1 - \mathrm{PRE})/(n - \mathrm{PA})} = \frac{.245/2}{(1 - .245)/21} = 3.406$$

$$F_{2,21} = \frac{\mathrm{SSR}/(\mathrm{PA} - \mathrm{PC})}{\mathrm{SSE(A)}/(n - \mathrm{PA})} = \frac{7.583/2}{23.375/21} = 3.406$$

And these come just short of beating the critical values for 2 and 21 degrees of freedom. Hence, we cannot reject the null hypothesis that there are no mean differences among these three categories or conditions.

This conclusion does not mean, of course, that we should accept the null hypothesis of no mean differences. And in this case, since we clearly had an expectation that the mean in the Failure condition would be less than the mean in the other two conditions, we should certainly proceed to directly test that hypothesis, which is the comparison made by the first contrast-coded predictor. Within the analysis of variance tradition, it is sometimes maintained that one should not test specific focused comparisons among category means unless the overall multiple-degree-of-freedom test that we have just conducted—that there are no mean difference among the categories—is rejected. We strongly disagree with this point of view. For reasons we have explained earlier, we are generally not enamored of model comparisons where PA – PC is > 1. One of the distinct advantages of a regression-based approach to traditional analysis of variance procedures is that one is forced to construct individual one-degree-of-freedom comparisons or contrasts among group means. Many traditional ANOVA programs automatically provide only the omnibus, multiple-degree-of-freedom test, and this, we think, is a distinct disservice.

As we have seen, the regression coefficients for X_{1i} and X_{2i} estimate particular differences among category means, the first comparing the Failure mean with the average of the No Feedback and Success means, and the second comparing the No Feedback and Success means. Hence, model comparisons that test whether these two parameters depart from zero are equivalently tests of mean differences among the three categories. Specifically, one model comparison is whether the parameter associated with X_{1i} equals zero:

MODEL A: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

MODEL C1: $Y_i = \beta_0 + \beta_2 X_{2i} + \varepsilon_i$

with the following equivalent null hypotheses:

H_0: $\beta_1 = 0$

H_0: $\mu_F = (\mu_{NF} + \mu_S)/2$

And the other model comparison tests whether the parameter associated with X_{2i} equals zero:

MODEL A: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

MODEL C2: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$

with the following equivalent null hypotheses:

H_0: $\beta_2 = 0$

H_0: $\mu_{NF} = \mu_S$

Model A for both of these comparisons is the same three-parameter augmented model that we estimated earlier, with a sum of squared errors of 23.375. Model C1 for the first comparison is estimated as follows:

MODEL C1: $\hat{Y}_i = 3.7083 + .3125 X_{2i}$

with a sum of squared errors of 29.396. And Model C2 for the second comparison is estimated as follows:

MODEL C2: $\hat{Y}_i = 3.7083 + .3542 X_{1i}$

with a sum of squared errors of 24.937. Note that the estimated intercept and slope in these models are unchanged from what they were in the Model A with both predictors. This results from the conjunction of two conditions: First, we have employed contrast-coded predictors, which by definition are orthogonal at the level of the three categories. Second, we have an equal number of observations in each of the three conditions. As a result of these two conditions, the contrast-coded predictors are uncorrelated with each other across the 24 individual observations. Their tolerance in Model A is 1.00.

The first model comparison, asking whether β_1 differs from zero, yields the following PRE and F statistics:

$$\text{PRE} = \frac{29.396 - 23.375}{29.396} = .205$$

$$F_{1,21} = 5.41$$

This F statistic exceeds the critical value of F with α at .05. Hence, we conclude that β_1 differs significantly from zero. Equivalently, we conclude that the mean value of Y_i in the Failure condition is significantly different from the average of the mean values in the Success and No Feedback conditions. Since the sample mean in the Failure condition is less than the average of the other two sample means, we conclude that Failure feedback in this study decreases subsequent performance relative to Success and No Feedback.

The test of the second null hypothesis, that β_2 equals zero, yields the following PRE and F statistics:

$$\text{PRE} = \frac{24.937 - 23.375}{24.937} = .063$$

$$F_{1,21} = 1.40$$

Since this F does not exceed its critical value, we conclude that β_2 does not differ significantly from zero. Equivalently, we cannot conclude that the mean performance under the Success condition is different from that under No Feedback.

Earlier in this chapter we gave a general formula for the SSR due to a contrast-coded predictor expressed in terms of the category means:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

This expression for the SSR of a contrast-coded predictor continues to apply in the case of categorical variables with more than two levels, as long as a full set of $m - 1$ contrast-coded predictors is included in Model A. Thus, in the present case, we have seen that the SSR for the Model A/Model C1 comparison that tested whether β_1 equaled zero was equal to:

$$\text{SSR}_{X_1} = \text{SSE(C1)} - \text{SSE(A)} = 29.396 - 23.375 = 6.021$$

This can be obtained equivalently in terms of the category means as:

$$\frac{((-2)3.00 + (+1)3.750 + (+1)4.375)^2}{(-2)^2/8 + (+1)^2/8 + (+1)^2/8} = 6.021$$

Likewise, we saw that the SSR for the Model A/Model C2 comparison that tested whether β_2 equaled zero was equal to:

$$\text{SSR}_{X_2} = \text{SSE(C2)} - \text{SSE(A)} = 24.937 - 23.375 = 1.562$$

This can be obtained equivalently in terms of the category means as:

$$\frac{((0)3.00 + (-1)3.750 + (+1)4.375)^2}{(0)^2/8 + (-1)^2/8 + (+1)^2/8} = 1.562$$

We have now done three different tests comparing the augmented model, which includes both contrast-coded predictors, with three different compact ones. The results of these three tests are presented in Figure 8.10. Notice that we have given labels, in parentheses, for each of these tests to indicate the questions they are examining in terms of the group means. The two-degree-of-freedom test, done first, comparing Model A to a Model C that predicted the grand mean for all observations, was an omnibus test of any group mean differences. The second, comparing models with and without X_{1i} as a predictor, examined whether the mean in the Failure condition differed from the average of the two in the other conditions. And the third, comparing models with and without X_{2i} as a predictor, examined whether the means in the No Feedback and Success conditions differed. We want to emphasize again that even though the two-degree-of-

FIGURE 8.10 Summary source table

Source	*b*	*SS*	*df*	*MS*	*F*	*PRE*
Model (between conditions)		7.583	2	3.792	3.406	.245
X_1 (Failure vs. No Feedback, Success)	.3542	6.021	1	6.021	5.409	.205
X_2 (No Feedback vs. Success)	.3125	1.562	1	1.562	1.403	.063
Error		23.375	21	1.113		
Total		30.958	23			

freedom test did not prove to be significant, we did find a significant mean difference when we tested the more focused contrast question represented by X_{1i}. As always, we strongly encourage focused PA – PC = 1 model comparisons.

As the sums of squares in this source table show, the SSRs for the individual predictor variables sum to the SSR for the first model comparison, where the overall augmented model was compared to a compact single-parameter model, predicting the grand mean for all observations. As we saw in Chapter 6, this will be the case whenever predictors are completely nonredundant, with tolerances of 1.0. In the present case, this results from the conjunction of two conditions: the use of contrast-coded predictors, which are by definition orthogonal at the level of the groups or categories; and the fact that each category contains the same number of observations.

A Quick Look at Alternative Contrast Codes

Earlier we stated that there were many, many possible sets of contrast codes that could be used to code a categorical predictor. Let us examine the same data that we have been focusing on using a different set of codes. Suppose we now define our contrast codes as follows:

	Failure	*No Feedback*	*Success*
λ_{1k}	–1	0	1
λ_{2k}	–1	2	–1

To differentiate these codes from the earlier set, we define Z_{1i} and Z_{2i} as contrast-coded predictors, assigning individuals the indicated values to represent category membership. We then regress Y_i on Z_{1i} and Z_{2i} with the estimated parameters:

$$\hat{Y}_i = 3.7083 + .6875Z_{1i} + .0208Z_{2i}$$

While the parameter estimates for the two contrast-coded predictors in this model are quite different from those that we estimated using the earlier set, in a deeper sense this model is equivalent to the model we developed under the old set of codes. Substituting for the values of Z_{1i} and Z_{2i}, we see that the group or condition means continue to be the predictions made by the model for all observations:

$$\hat{Y}_F = 3.7083 + .6875(-1) + .0208(-1) = 3.000$$

$$\hat{Y}_{NF} = 3.7083 + .6875(0) + .0208(+2) = 3.750$$

$$\hat{Y}_S = 3.7083 + .6875(+1) + .0208(-1) = 4.375$$

Since the model makes the same predictions for all observations as the model with the previous set of contrast-coded predictors, the sum of squared errors is identical to what it was before, that is, 23.375.

The regression coefficients for the contrast-coded predictors have changed since the new contrast codes are making different comparisons among condition means from the comparisons made by the old set of codes. The contrast-coded predictor Z_{1i} is now comparing the means in the Success and Failure conditions. The value of its regression coefficient equals half the difference between these two group means. The second contrast-coded predictor, Z_{2i}, compares the mean in the No Feedback condition with the average of the means of the other two conditions. Its regression coefficient equals one-third of the difference between the mean in the No Feedback condition and the average of the other two means. These values for the regression coefficients are easily derived using the formula we gave earlier for the regression coefficient for a contrast-coded predictor. They also follow immediately once we realize the comparisons made by the contrasts and the number of units that separate observations in the various conditions on Z_{1i} and Z_{2i}. The intercept has not changed in value as a result of the new set of codes. It still equals the mean of the three condition means, as it will whenever a full set of contrast-coded predictors is used.

Since the change in codes has not changed the predicted values or the sum of squared errors for this model, a test of the null hypothesis that both β_1 and β_2 equal zero produces the same values of PRE and F as it did under the old set of codes. The compact single-parameter model is:

MODEL C: $\hat{Y}_i = 3.7083$

with a sum of squared errors of 30.958. PRE continues to equal .245, which converts to an F of 3.406 with 2 and 21 degrees of freedom. Thus, a test of the omnibus null hypothesis—that all of the condition means equal one another—reaches the same conclusion regardless of our choice of contrast codes.

Single-degree-of-freedom tests of whether β_1 or β_2 equals zero, however, reach rather different conclusions than they did before. These regression coefficients now estimate different comparisons between the condition means from those estimated with the earlier set of contrast codes. A test of whether β_1 equals zero is now equivalent to a test of whether the means in the Failure and Success conditions are equal to each other. The compact model for this test is:

MODEL C1: $\hat{Y}_i = 3.7083 + .0208Z_{2i}$

with a sum of squared errors of 30.938. The resulting PRE equals:

$$\frac{30.938 - 23.375}{30.938} = .244$$

which converts to an F statistic of 6.793 with 1 and 21 degrees of freedom. Since this exceeds the critical value, our test reveals both that β_1 is significantly greater than zero and that the mean in the Success condition is significantly greater than the mean in the Failure condition. As before, the sum of squares reduced associated with this contrast-coded predictor can be computed as a function of the relevant condition means that are being compared:

$$SSR_{Z_1} = \frac{[(-1)3.00 + (0)\ 3.750 + (+1)4.375]^2}{(-1)^2/8 + (0)^2/8 + (+1)^2/8} = 7.562$$

The test of whether β_2 differs from zero is equivalently a test of whether the No Feedback mean is significantly different from the average of the Failure and Success means. The compact model for this test is:

MODEL C2: $\hat{Y}_i = 3.7083 + .6875Z_{1i}$

with a sum of squared errors of 23.396. The resulting PRE equals .001, which converts to an F of 0.019 with 1 and 21 degrees of freedom. Clearly, the difference between the No Feedback mean and the average of the means in the other two conditions is not significant. As before, the SSR for this comparison can be directly calculated from the condition means:

$$SSR_{Z_2} = \frac{[(-1)3.00 + (+2)3.750 + (-1)4.375]^2}{(-1)^2/8 + (+2)^2/8 + (-1)^2/8} = 0.021$$

Since these two new contrast-coded predictors are nonredundant, just as were the earlier two, the two SSRs explained by each predictor over and above the other can be added to equal the SSR explained by them both as a set. The three tests we have just conducted can be summarized in Figure 8.11. Note that the only changes in this source table compared to the one using the earlier set of contrast codes (Figure 8.10) occur in the two rows of the table testing the specific comparisons made by the coefficients associated with Z_{1i} and Z_{2i}. All we have done is divide up the sum of squares between conditions (7.583) differently here, focusing on a different set of contrasts. As we have said before, with m category levels, there are only $m - 1$ orthogonal contrasts or comparisons that can be used. And whether we use these codes or those we discussed earlier is only a matter of theoretical preference.

FIGURE 8.11 Summary source table

Source	*b*	*SS*	*df*	*MS*	*F*	*PRE*
Model (between conditions)		7.583	2	3.792	3.406	.245
Z_1 (Failure vs. Success)	.6875	7.562	1	7.562	6.793	.244
Z_2 (No Feedback vs. Failure, Success)	.0208	0.021	1	0.021	0.019	.001
Error		23.375	21	1.113		
Total		30.958	23			

Problems and Pitfalls in Using Nonorthogonal Codes

Suppose we were interested in asking the following two questions of these data, coded by the following set of codes:

	Failure	*No Feedback*	*Success*
λ_{1k}	–2	+1	+1
λ_{2k}	–1	0	+1

The first question is whether the Failure mean differs from the average of the means in the other two conditions. This was the first of the codes that we used in our first set, used to create the predictor X_{1i}. The second is whether the means in the Failure and Success conditions differ. This was the first of the codes that we used in our second set, used to create the predictor Z_{1i}. As we have seen, these are perfectly legitimate questions that we might want to ask of these data, but they are not orthogonal questions about the mean differences, as revealed by the fact that the second condition for contrast codes is not met by these two codes if we use them simultaneously:

$$\sum_k \lambda_{1k}\lambda_{2k} = (-2)(-1) + (+1)(0) + (+1)(+1) = +3$$

It is for this reason that we have not called them "contrast codes" when we think about them as a set. But, of course they are perfectly valid questions to ask of the data, albeit nonorthogonal.

If we were interested in these two questions, we might be tempted to use the two resulting predictors, X_{1i} and Z_{1i}, created with these codes to simultaneously predict Y_i, even though the codes themselves are not orthogonal. The resulting model would be:

$$\hat{Y}_i = 3.7083 + .0417\, X_{1i} + .6250 Z_{1i}$$

This model, even with these nonorthogonal codes, is the same in a deep sense as the earlier model, in that it makes the same predictions of the condition means for every observation:

$$\hat{Y}_F = 3.7083 + .0417(-2) + .6250(-1) = 3.000$$

$$\hat{Y}_{NF} = 3.7083 + .0417(+1) + .6250(0) = 3.750$$

$$\hat{Y}_S = 3.7083 + .0417(+1) + .6250(+1) = 4.375$$

And, as a result, it has the same sum of squared errors, 23.375. A model comparison between it, as Model A, and the single-parameter Model C, making a constant prediction for all observations, thus still provides the omnibus two-degree-of-freedom test about whether there are any differences among the category means. The resulting test of the overall model is summarized in the first row of the source table given in Figure 8.12 (PRE = .245, $F(2,21) = 3.406$). Both this row and the final two rows of the table are identical to what they were in the earlier source tables that we presented from these data. However, when we estimate the model in a regression program and examine the regression coefficients for the individual predictor variables and their respective SSRs, as given in the source table of Figure 8.12, they are not the values that we might expect.

FIGURE 8.12 Summary source table using nonorthogonal codes

Source	*b*	*SS*	*df*	*MS*	*F*	*PRE*
Model (between conditions)		7.583	2	3.792	3.406	.245
X_1	.0417	0.021	1	0.021	0.019	.001
Z_1	.6250	1.562	1	1.562	1.403	.063
Error		23.375	21	1.113		
Total		30.958	23			

The coefficient for X_{1i}, when it was embedded in an orthogonal set of contrast-coded predictors (i.e., with X_{2i}), equaled .3542, which was shown to be:

$$\frac{\frac{\bar{Y}_{NF} + \bar{Y}_S}{2} - \bar{Y}_F}{3}$$

and its associated SSR was 6.021, which was shown to equal:

$$\frac{((-2)\bar{Y}_F + (+1)\bar{Y}_{NF} + (+1)\bar{Y}_S)^2}{(-2)^2/8 + (+1)^2/8 + (+1)^2/8}$$

Now, however, when it is used to predict Y_i along with the nonorthogonal predictor Z_{1i}, its estimated coefficient equals .0417 and its SSR equals 0.021.

Similarly, the coefficient for Z_{1i} in this model equals .625, whereas its coefficient when used in the orthogonal set with Z_{2i} equaled .687, which was half the difference between $\bar{Y}_S$ and $\bar{Y}_{NF}$. And now its SSR equals 1.562, whereas earlier, when used in the orthogonal set with Z_{2i}, its SSR was 7.562, which was shown to equal:

$$\frac{((-1)\bar{Y}_F + (0)\bar{Y}_{NF} + (+1)\bar{Y}_S)^2}{(-1)^2/8 + (0)^2/8 + (+1)^2/8}$$

The important point is that predictors that code the levels of a categorical independent variable will not yield coefficients that equal the expected mean differences and their associated SSRs unless a full set of orthogonal contrast-coded predictors is used. In other words, the following important equalities will *not* hold, unless a full set of contrast-coded predictors is included in the augmented model:

$$b = \frac{\sum_k \lambda_k \bar{Y}_k}{\sum_k \lambda_k^2}$$

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)}$$

Serious interpretative errors can ensue if one thinks that a given categorical predictor codes a particular mean difference when it is embedded in a nonorthogonal set. Orthogonality here depends solely on meeting the second defining condition of contrast codes—that the sum of the products of their coded values across category levels equals zero. As we will show, unequal numbers of observations can result in redundant contrast-coded predictors, redundant across observations, but this creates no interpretative problems as long as the codes are themselves orthogonal contrasts.

Given this, how might one proceed if one really were interested in testing the nonorthogonal questions of whether the Failure mean differs from the average of the

No Feedback and Success means (i.e., the question implicit in the X_{1i} codes) and of whether the Failure and Success means differ (i.e., the question implicit in the Z_{1i} codes)? Obviously, one could test these sequentially by specifying two models, one using both X_{1i} and X_{2i} as predictors and the other using Z_{1i} and Z_{2i} as predictors, just as we did in the earlier sections. Alternatively, one could simply rely on two bits of knowledge to test the mean differences implied by these contrasts without doing both estimations. First, as we have shown, the SSE(A) for a Model A that incorporates any complete set of contrast-coded predictors will be the same regardless of the specific set of such predictors used. Second, if a given contrast-coded predictor were included in a complete set of contrast-coded predictors, its SSR would be given by the following formula:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)}$$

Accordingly, in the case at hand, had one simply estimated the model that included X_{1i} and X_{2i} as predictors, one would have known that the SSE(A) for a model that used any full set of contrast-coded predictors would equal 23.375 with 21 degrees of freedom. One then could calculate the SSR associated with Z_{1i} if it were embedded in a full set of contrast-coded predictors, that is:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2/n_k)} = \frac{((-1)\bar{Y}_F + (0)\bar{Y}_{NF} + (+1)\bar{Y}_S)^2}{(-1)^2/8 + (0)^2/8 + (+1)^2/8} = 7.562$$

Then one could calculate the values of PRE and F that would result if one tested Z_{1i} in the context of a complete set of contrast-coded predictors:

$$\text{PRE} = \frac{7.562}{23.375 + 7.562} = .244$$

$$F_{1,21} = \frac{7.562/1}{23.375/21} = 6.793$$

When doing this, one should recognize that the questions represented by these two contrast codes are not independent. The question of whether the Failure mean differs from the No Feedback and Success means is not entirely independent of the question of whether the Failure and Success means differ from each other. The answer to one is partially informative about the answer to the other.

Dummy Codes

There is another coding convention, known as dummy coding, that is widely used in some of the literature. Under this convention, one of the groups defined by the categorical variable is given values of zero on all $m - 1$ codes, and the other groups are given values

of zero on all but one of the codes. So, for instance, in the case at hand the following codes are consistent with this convention:

	Failure	*No Feedback*	*Success*
λ_{1k}	0	+1	0
λ_{2k}	0	0	+1

Obviously dummy codes do not meet either of the conditions that define contrast codes. As a result, it takes care to interpret exactly what is examined when one uses such codes to form predictor variables. One might think that a predictor that uses the first codes above would be comparing the No Feedback group to the other two groups and that a predictor that uses the second code above would be comparing the Success group with the other two groups, but in fact the regression coefficients for these two coded predictors, if included simultaneously, would each be asking whether the group coded zero on both codes differs from the group coded with a 1 for the predictor that is being examined. So, the predictor with the first code would examine the Failure–No Feedback difference and the predictor with the second code would examine the Failure–Success difference.

Because of the fact that interpretive mistakes can follow from the use of dummy codes, unless one is thoroughly familiar with them, we strongly recommend that researchers adopt the contrast-coding convention that we have explicated and that we will use in the remainder of this book.

CONTRAST CODES WITH UNEQUAL CELL SIZES

Historically, the procedure of ANOVA to detect mean differences was developed for data from experimental designs in which there were equal numbers of observations in every cell or condition of the design. In this sense, it was developed as an arithmetic shortcut, based on the assumption that predictors would be nonredundant. Of course, with the wide availability of computer programs that permit the estimation of linear regression models with partially redundant predictors, the assumption of nonredundant predictors is no longer necessary and, as we have just shown, ANOVA is easily implemented within general purpose multiple regression procedures, even with unequal numbers of observations in the various conditions.

In this section we present a new example with four levels of a categorical variable having unequal numbers of observations in each level or category. The bottom line is that, as long as estimation is done with a full set of contrast-coded predictors, all interpretations that we have previously given continue to be applicable, even though with unequal n values those predictors will be partially redundant across observations.

Let us assume that you are in a Psychology Department of a major university in which there are four PhD programs to which students are admitted every year: Clinical, Developmental, Experimental, and Social. The number of students who are admitted varies across the programs. Your question is whether there are mean differences in the verbal Graduate Record Examinations (GREs) of admitted students across the four programs. The data for a given year are presented in Figure 8.13, along with the

FIGURE 8.13 Hypothetical GRE scores, group means, and contrast codes

	Program			
	Clinical	*Developmental*	*Experimental*	*Social*
	750	700	640	690
	730	630	660	720
	710	620	710	750
	690		620	670
	670			650
	770			
$\bar{Y}_k$	720	650	657.5	696
n_k	6	3	4	5
λ_{1k}	3	−1	−1	−1
λ_{2k}	0	2	−1	−1
λ_{3k}	0	0	1	−1

four group means and the contrast-coded predictors that we will use. We simply use the Helmert coding convention here to derive these codes, as we have no strong expectations about where mean differences might be found.

We create three contrast-coded predictors, X_{1i}, X_{2i}, and X_{3i}, using these codes. We then proceed to estimate a Model A in which these three are used to predict verbal GRE scores:

$$\hat{Y}_i = 680.875 + 13.042X_{1i} - 8.917X_{2i} - 19.250X_{31}$$

This model has a sum of squared errors of 21,595 and makes predictions of the category means for all observations.

The model's parameter estimates can be interpreted as we have done previously, based on the fact that we used a full set of contrast-coded predictors, even though those predictors are now partially redundant because of the unequal numbers of observations (i.e., the tolerances of all three predictors are less than 1.0).

The intercept, 680.875, is the mean of the four category means. It is not the mean of all 19 observations, which equals 687.778. The values of the three slopes equal differences among the category means, according to the following formulas:

$$b_1 = \frac{\sum_k \lambda_{1k}\bar{Y}_k}{\sum_k \lambda_{1k}^2} = \frac{(3)\bar{Y}_C + (-1)\bar{Y}_D + (-1)\bar{Y}_E + (-1)\bar{Y}_S}{12} = \frac{\bar{Y}_C - \dfrac{\bar{Y}_D + \bar{Y}_E + \bar{Y}_S}{3}}{4} = 13.042$$

$$b_2 = \frac{\sum_k \lambda_{2k}\bar{Y}_k}{\sum_k \lambda_{2k}^2} = \frac{(0)\bar{Y}_C + (2)\bar{Y}_D + (-1)\bar{Y}_E + (-1)\bar{Y}_S}{6} = \frac{\bar{Y}_D - \dfrac{\bar{Y}_E + \bar{Y}_S}{2}}{3} = -8.917$$

$$b_3 = \frac{\sum_k \lambda_{3k}\bar{Y}_k}{\sum \lambda_{3k}^2} = \frac{(0)\bar{Y}_C + (0)\bar{Y}_D + (1)\bar{Y}_E + (-1)\bar{Y}_S}{2} = \frac{\bar{Y}_E - \bar{Y}_S}{2} = -19.250$$

FIGURE 8.14 Summary source table for analysis of GRE scores

Source	*b*	*SS*	*df*	*MS*	*F*	*PRE*
Model (between groups)		14,516.00	3	4838.70	3.14	.40
X_1 (C vs. D, E, S)	13.04	10,727.00	1	10,727.00	6.95	.33
X_2 (D vs. E, S)	–8.92	1605.00	1	1605.00	1.04	.07
X_3 (E vs. S)	–19.25	3293.89	1	3293.89	2.14	.13
Error		21,595.00	14	1542.5		
Total		36,111.00	17			

Figure 8.14 presents the source table that results from a comparison of this Model A with four different Model Cs. The first row of the table is a comparison between this Model A and the single-parameter Model C that predicts the grand mean for all observations. The null hypothesis for this comparison is that all the group means are equal to each other. The second row of the table gives the model comparison between this Model A and a Model C that omits the X_{1i} predictor. The null hypothesis here is that β_1 equals zero or, equivalently, that the mean of the Clinical group (C) is equal to the average of the means of the other three groups. The third row of the table gives the model comparison between this Model A and a Model C that omits the X_{2i} predictor. The null hypothesis here is that β_2 equals zero or, equivalently, that the mean of the Developmental group (D) is equal to the average of the means of the Experimental (E) and Social (S) groups. And the fourth row of the table gives the model comparison between this Model A and a Model C that omits the X_{3i} predictor. The null hypothesis here is that β_3 equals zero or, equivalently, that the mean of the Experimental group is equal to the mean of the Social group. All the resulting SSRs for these single-degree-of-freedom comparisons can be expressed in terms of the means, according to the formula we have frequently used before:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

In short, all interpretations and computations explicated in this chapter for the analysis of a categorical independent variable apply regardless of whether there are equal numbers of observations across the levels of the categorical variable, as long as (*once again*) a full set of contrast-coded predictors is employed. The only thing that differs in this example from those presented earlier, as a function of the unequal values of n_k, is that the sums of squares for the individual single-degree-of-freedom tests in the above source table cannot be added up to equal the overall SSR for the model as a whole. In this case, the sum of the SSRs for the single-degree-of-freedom comparisons equals 15,625.89, while the overall test of the model yields an SSR of 14,516. This difference is due to the fact that across observations the predictors are now somewhat redundant. But all model comparisons and interpretations remain as they have been all along throughout the chapter.

ORTHOGONAL POLYNOMIAL CONTRAST CODES

Sometimes the observations fall into discrete categories on an independent variable of interest even though the underlying variable itself can be thought of as a continuum. For instance, suppose we were interested in age differences among elementary school children in their performance on a standardized arithmetic test. We take children from three different elementary school classes, those in the fourth, fifth, and sixth grades, and give them the standardized test. We then want to know if there are class mean differences. Our conceptual independent variable of interest is the children's age, but what we measure is their year in school and observations are clearly in three distinct categories on this measured variable.

FIGURE 8.15 Hypothetical scores on standardized arithmetic test

	Grade		
	Fourth	*Fifth*	*Sixth*
	68	68	80
	72	75	75
	76	68	78
	65	72	
	70	65	
		80	
		69	
$\bar{Y}_k$	70.20	71.00	77.67
n_k	5	7	3

Imagine that we had data from the 14 children given in Figure 8.15. In cases like this one, there is a special set of contrast codes that are sometimes useful for assessing trends in category means. These special codes are really just regular contrast codes, but they have the special name of "orthogonal polynomials" when values on the categorical independent variable can be ordered on some underlying continuum, as they clearly can in this case. In Figure 8.16 we present orthogonal polynomial contrast codes for categorical predictors having up to five levels.

As we indicate there, these codes have names that refer to the trend in the category means, across levels of the categorical independent variable, that they examine. So, with a two-level categorical variable, we can only examine whether the means go up or down, in essence fitting a linear function to the category means. With three levels, we can fit both a linear trend to the three category means and also ask whether the mean of the middle level is higher or lower than it ought to be given a simple linear ordering

FIGURE 8.16 Orthogonal polynomial contrast codes

Trend	*Category*				
	1	2			
Linear	–1	1			
	1	2	3		
Linear	–1	0	1		
Quadratic	–1	2	–1		
	1	2	3	4	
Linear	–3	–1	1	3	
Quadratic	1	–1	–1	1	
Cubic	–1	3	–3	1	
	1	2	3	4	5
Linear	–2	–1	0	1	2
Quadratic	2	–1	–2	–1	2
Cubic	–1	2	0	–2	1
Quartic	1	–4	6	–4	1

(the quadratic trend). With four levels, we can fit not only linear and quadratic trends, but also a cubic one, having two bends rather than one. And so forth.

With the three-level categorical variable in our data, let us use the two codes from the orthogonal polynomials to fit the linear and quadratic trends to these data. We create two contrast-coded predictors, X_{1i} and X_{2i}, using the codes specified in Figure 8.16 for a three-level categorical variable (4th grade is category level 1, etc.). The estimated model, using these to predict the standardized test scores, is:

$$\hat{Y}_i = 72.96 + 3.73X_{1i} - .98X_{2i}$$

with a sum of squared errors of 237.47. Of course this model, with a full set of contrast codes, exactly predicts the group means:

$$\overline{Y}_{4\text{th}} = 72.96 + 3.73(-1) - 0.98(-1) = 70.20$$

$$\overline{Y}_{5\text{th}} = 72.96 + 3.73(0) - 0.98(+2) = 71.00$$

$$\overline{Y}_{6\text{th}} = 72.96 + 3.73(+1) - 0.98(-1) = 77.67$$

And as always the parameter estimates can be interpreted in terms of the group means. The intercept, 72.96, is the mean of the three means, 3.73 is half the difference between $\overline{Y}_{6th}$ and $\overline{Y}_{4th}$ and –0.98 is one-third of the difference between $\overline{Y}_{5th}$ and the average of the other two group means. The source table that summarizes the analysis of these data is given in Figure 8.17.

So far, there is nothing new about this model or its interpretations. So why do we refer to the codes we have used as orthogonal polynomials? The reason is that the slope of the first one is the slope that results if we were to fit a straight line to the three group means, going from the fourth grade up to the sixth grade. And the slope of the second contrast-coded predictor estimates the degree to which the group mean for the fifth grade does not lie on that prediction line, that is, the degree to which that prediction function deviates from a straight line if it is to predict all three group means. Given the significance of the coefficient associated with X_{1i}, we can conclude that there is a linear increase in performance on the standardized test as we go up from children in the fourth grade to children in the sixth grade.

An obvious question is how these results would differ from what would be obtained if we simply regressed Y_i on Grade itself, treated as a continuous variable, numerically coded as 4, 5, and 6. Such a model would be a simple regression model, asking if there is a linear relationship between Grade and test performance. It is estimated as:

$$\hat{Y}_i = 55.62 + 3.38Grade_i$$

with a sum of squared errors of 268.62. A test of whether there is a linear relationship between grade and performance yields an SSR of 88.31, PRE = .25, and $F = 4.27$ with 1 and 13 degrees of freedom. In this model, the slope associated with the Grade predictor variable informs us about the degree to which predicted performance on the standardized math test increases as grade goes up by one unit, that is, from fourth to fifth and from fifth to sixth.

While these are obviously different models in a variety of ways, the important conceptual difference is that in the one using the orthogonal polynomial contrast codes we are predicting the group means and asking about differences among those group means. And in the simple linear regression model, using Grade as our predictor, we are

FIGURE 8.17 Summary source table for analysis of arithmetic scores

Source	*b*	*SS*	*df*	*MS*	*F*	*PRE*
Model (between groups)		119.47	2	59.73	3.019	.34
X_1	3.73	104.506	1	104.506	5.28	.31
X_2	–0.98	31.17	1	31.17	1.57	.12
Error		237.47	12	19.79		
Total		356.93	14			

simply fitting a linear function to all the individual observations, rather than to the group means. Given that the group sizes are very unequal, modeling the group means and modeling the individual observations yield different results.

TYPE I ERROR RATES IN TESTING MEAN DIFFERENCES

Earlier we discussed the general strategy for testing any mean difference that was of interest in designs with multiple levels (m) of a categorical independent variable. One first derives a full set of contrast codes ($m - 1$ of them) and uses them to estimate an augmented model, which predicts the category means. This model provides tests of the specific mean differences that were used as codes for the predictors. Importantly, it also provides the SSE(A) and the mean square error for any model that used a full set of codes, regardless of which set was used. Then, for any additional mean comparison of interest, one calculates the SSR associated with that contrast as if it were used as a predictor in a complete set of orthogonal contrast-coded predictors:

$$\text{SSR} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k (\lambda_k^2 / n_k)}$$

Dividing this SSR by the MSE from the estimated model yields the F statistic associated with the mean comparison of interest, and the PRE for that comparison can be calculated as:

$$\frac{\text{SSR}}{\text{SSE(A)} + \text{SSR}}$$

When there are more than two or three levels of the categorical variable, the number of potential mean differences that might be tested can become very large. For instance, with only four levels, one could in theory test the means of individual groups against each other, the means of all pairs of groups against each other, the mean of each triad of groups against the mean of the remaining group, and so forth. A problem that arises in this case, where many mean differences might be tested, is that the probability of a Type I statistical error may become unacceptably large. While α may be set at .05 for any one mean comparison, across many such comparisons the probability that somewhere a Type I error has been committed can quickly become quite large. For instance, if we

had four groups and we asked whether each mean differed from each other mean, that would be six contrasts tested (in addition to others included in the orthogonal set of codes used initially to generate Model A). If on each test α was set at .05, across the six tests, the probability that we would make at least one Type I error is equal to:

$$1 - (1 - .05)^6 = 1 - .95^6 = .265$$

In other words, even if the null hypothesis were true and all the true group means were equal to each other, at least one of the six mean comparisons would be significant more than a quarter of the time. In the ANOVA literature many different procedures have been developed for dealing with this issue. We focus on only two of them, and the crucial difference between these two is whether the mean comparison that is tested is a *planned comparison* or a *post hoc comparison*.

Planned comparisons

Planned comparisons are those that the researcher had theoretical or substantive reasons for examining *before* conducting the experiment. In other words, the researcher specified all the planned comparisons of interest before collecting or examining the data. Ideally, as many of these planned contrasts as possible would be included in the set of orthogonal contrast-coded predictor used to generate the initial Model A. Regardless, one adds up the number of comparisons that one intends to examine, both in the initial model and in the other models of theoretical interest. Let us say that number is c. To keep α at .05 across all c tests, one wants to compare each obtained F to a critical value of F, using α/c to determine the critical F. So, for instance, if there are six tests to be conducted, one would use a critical value of F at .05/6 = .0083 rather than at .05.[1]

Post hoc comparisons

Post hoc comparisons are those that do not occur to us until *after* we have examined the data. Often when looking at the data certain comparisons that we did not anticipate appear to be interesting. It is natural to want to test those interesting, unanticipated contrasts. However, it is impractical to use the above procedure for planned comparisons because when looking at the data we are implicitly doing many, many comparisons—all those that do not strike us as interesting—that ought to be included in c, the total number of comparisons made. Instead of trying to count all those implicit comparisons, standard practice is to compare F to the following critical value developed by Scheffé (1959):

$$(m - 1)F_{crit;m-1,n-PA;\alpha}$$

There are two important features of using the Scheffé adjusted critical value. First, the overall probability of making at least one Type I error will remain at α no matter how many contrasts are evaluated using the adjusted critical value. Thus, the researcher can do as much snooping and exploring with contrasts as desired without undue risk of making Type I errors. Second, there will be at least one contrast whose F exceeds the Scheffé adjusted critical value if and only if the omnibus test, comparing Model A to a Model C that predicts the grand mean for all observations, is statistically significant. Thus, if the omnibus test is not significant, then there is no point in evaluating any post hoc contrasts using the Scheffé criterion.

POWER ANALYSIS FOR ONE-WAY ANOVA

Estimating Statistical Power

The advantage of consistently adhering to the model comparison approach is that all that we have learned before still applies as we consider new types of models. Thus, the methods for estimating statistical power presented in Chapter 6 for multiple regression apply unaltered to one-way ANOVA. In particular, you can use prior research to estimate the value of η^2 (the expected proportional reduction in error) for either the omnibus test of any mean differences or specific one-degree-of-freedom contrasts. As before, you should consider adjusting empirical estimates of η^2 based on the number of observations and parameters. This is especially important for experiments in which the number of observations is often small relative to the number of model parameters. Also, as before, you may use either Cohen's values for small, medium, and large effects for your power analyses, or values from the literature of your substantive research topic. For example, if we wanted to re-do the SAT coaching study (with which we began this chapter) with a larger number of observations, we would start our power analysis by adjusting the PRE of .196 reported in that study:

$$\hat{\eta}^2 = 1 - (1 - PRE)\left[\frac{n - PC}{n - PA}\right] = 1 - (1 - .196)\left[\frac{20 - 1}{20 - 2}\right] = .15$$

A quick check of the power associated with this value suggests that about 50 participants would be required to have an 80% chance of detecting an effect of this magnitude.

Earlier we also saw how estimates of the parameter values—the regression slopes and the variance of the error—can provide estimates of the effect size for power analysis. However, in one-way ANOVA, rather than having prior ideas about the values of regression parameters, a researcher more commonly has notions about the values of the group means that determine the regression slopes. Hence, it is worth examining how we can estimate the expected effect size by beginning with expectations of the cell means. We begin with our usual definition of PRE:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{\text{SSR}}{\text{SSE(C)}}$$

We have noted before that SSE(C) = SSE(A) + SSR (i.e., the error for the compact model includes all the error of the augmented model plus the error that was reduced by the addition of the extra parameters in the augmented model). Hence:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE(A)} + \text{SSR}} = \frac{1}{(\text{SSE(A)}/\text{SSR}) + 1}$$

To obtain a definition of η^2, we simply calculate SSE(A) and SSR using the true parameter values (β_0, β_1, etc., ε^2) instead of the estimated parameters (b_0, b_1, etc., s^2). For one-way ANOVA, we can start with our expectations about what the true group means, μ_k, might be, and use those in the formula for the SSR for a contrast-coded predictor:

$$\text{SSR for a contrast} = \frac{[\Sigma \lambda_k \mu_k]^2}{\Sigma \lambda_k^2 / n_k}$$

For a complete model, SSE(A) depends only on σ^2, the within-cell variance. Specifically:

$$\text{SSE(A)} = \left(\sum_k n_k\right)\sigma^2 = n\sigma^2$$

where n is the grand total of observations. We multiply by n instead of by (n – PA) because we have not estimated any parameters from data in the calculations of SSE(A). Substituting these values for SSR and SSE(A) calculated from the presumed true parameters into the formula for PRE yields the following formula for η^2:

$$\hat{\eta}^2 = \left(\frac{n\sigma^2(\Sigma \lambda_k^2/n_k)}{[\Sigma \lambda_k \mu_k]^2} + 1\right)^{-1}$$

If n_k, the number of observations in each group, is equal for all groups and if m is the number of groups so that $n = mn_k$, then the above formula reduces to:

$$\hat{\eta}^2 = \left(\frac{m\sigma^2(\Sigma \lambda_k^2)}{[\Sigma \lambda_k \mu_k]^2} + 1\right)^{-1}$$

which does not depend on either the total number of observations or the number of observations in each group. Therefore, to find a value for η^2 to use in our power calculations, we need only specify the values for σ^2 and for μ_k that we expect to obtain in our study. And, of course, we need to specify the λ_k values for the contrast code for which we want to estimate the statistical power.

As an example of this direct approach for estimating η^2, let us again consider the feedback study described earlier in this chapter. Suppose the researcher on the basis of prior research had expected values of 3, 4, and 4 for the means of the Failure, No Feedback, and Success groups, respectively, and a within-cell variance of about 1.5. Then for the {–2,1,1} contrast the expected effect size is:

$$\hat{\eta}^2 = \left(\frac{3(1.5)6}{[(-2)3 + 4 + 4]^2} + 1\right)^{-1} = .13$$

From this we estimate that about 60 observations, 20 per group, would be necessary to provide a statistical power of .8.

Power and Research Design

When planning experiments, researchers must choose how many groups to use and how many observations should be in each group. It is important to consider the consequences that such choices have for statistical power. We first consider the design implications for the omnibus test of whether there are any differences among the means and then for the power of specific contrasts.

The statistical power of the omnibus test of whether there are any mean differences among the groups is maximized when there are an equal number of observations in each group. In that case, the omnibus F statistic equals the average of all the F statistics for

a set of orthogonal contrasts. This highlights a common mistake in research design: If too many groups are used, then there are many contrasts where no differences are expected, which in turn lowers the omnibus *F*. Sometimes using too many groups cannot be avoided. For example, biopsychological researchers must sometimes use multiple control groups (e.g., handling the animal, injecting with a drug vehicle, injecting with a placebo) in comparison to a single treatment group. If there are no differences expected among the different control groups, then the possible magnitude of the omnibus *F* statistic is reduced. In this case, the omnibus *F* is often misleading and researchers should simply focus on the treatment versus controls contrast.

Researchers using ordered category levels over which they expect polynomial trends often use too many groups. For example, researchers expecting linear and perhaps quadratic trends sometimes err in using as many as five levels. With equal numbers of observations at each level, the omnibus *F* is reduced because it is the average of not only the expected linear and quadratic contrasts but also the not-expected cubic and quartic contrasts. Even the power of the separate tests of the linear and quadratic contrasts is reduced because some of the study's valuable resources—the observations—have been allocated to test for the cubic and quartic effects. At the same time, testing more polynomial trends than expected increases the chances of making Type I errors.

For specific contrasts, allocating an equal number of observations to all groups effectively gives equal importance to all contrasts. If one or two contrasts are more important to the research purpose than the other contrasts, one may want to allocate the number of observations unequally across the groups so as to maximize the statistical power of the contrasts of greater importance. Power for a contrast is maximized when the allocation of observations to groups is proportional to the absolute values of the contrast weights. For example, for the *t*-test where the weights are {–1,+1}, statistical power is maximized with an equal number of observations in each group. However, for the quadratic contrast for three groups with weights {–1,+2,–1}, power is maximized by allocating a quarter of the observations to each of the extreme categories and the remaining half of the observations to the middle category. The biopsychology researcher comparing the treatment group to four control groups could maximize the power of that comparison by allocating half the total observations to the treatment group and one-eighth of the observations to each of the four control groups.

SUMMARY

In this chapter we have considered models with a single categorical predictor having two or more levels or categories. Such predictors need to be numerically coded and we have given our strong preference that the codes to be used are orthogonal contrast codes. A full set of such codes means that there are $m - 1$ contrast-coded predictors with m category levels. Additionally, given that the sum of the codes for any predictor equals one, orthogonality is assured by codes where the sum of all pairs of crossproducts equals zero.

Assuming that a full set of codes is used to predict the data variable, then the predicted values from such a model will equal the category means of that data variable. Accordingly, inferences about slopes are then equivalently inferences about difference among category means. And these models are then equivalent to traditional independent sample *t*-tests (given $m = 2$) and one-way ANOVA models (given $m > 2$). Although

one-way ANOVA has traditionally emphasized omnibus tests to examine whether there are any mean differences among the groups or categories, our strong preference is for single-degree-of-freedom model comparisons, testing specific focused contrasts that are of theoretical interest. We encourage analysts to ask about mean differences that are of interest to them, even when those differences are not themselves orthogonal. We discuss appropriate procedures for asking about such nonorthogonal differences, given the constraint that a full set of codes be orthogonal. Finally, we discuss procedures for avoiding inflated α levels when conducting many mean comparisons, whether a priori or post hoc.

Note

1 The use of α/c instead of α is based on the *Bonferroni inequality* and so some manuals for statistical programs refer to this as the Bonferroni method of multiple comparisons. Sometimes this is also referred to as the Dunn method.