Tutorial 3

Linear Models

The Relation Between ANOVA and Regression

There is a close relation between the regression models presented in such standard regression textbooks as Cohen, Cohen, West, and Aiken (2002); Darlington and Hayes (2017); Judd and McClelland (2017); and Pedhazur (1997), and the ANOVA models we present in this book. We touch upon this relation in the tutorial on regression, but that tutorial is intended primarily as a brief primer on multiple regression for readers who have not already encountered it previously. This tutorial, on the other hand, is intended primarily for readers who already have considerable background in regression but who are interested in developing a deeper understanding of the relationship between ANOVA and regression. Specifically, our purpose in this tutorial is threefold: (1) to illustrate some examples of the relation between regression and ANOVA, (2) to explain the place of regression models and ANOVA models in a broader methodological framework, and (3) to explain why we have chosen to focus primarily on ANOVA models in this book, although ANOVA is often regarded as a special case of regression.

THE RELATION BETWEEN REGRESSION AND ANOVA MODELS

Statistical models are usually defined in terms of parameters, which specify the form of relationship between variables. It is often the case that two models can be written in different forms with parameters that appear to be different from one another, and yet the two models are in fact equivalent to one another. We shall take "equivalent" to mean that even though the specific parameter estimates of the two models may not necessarily be the same, the sum of squared errors of the two models will always be equal to one another.

Single-Factor Between-Subjects Design

For example, consider how to write a full model for the single-factor between-subjects design of Chapter 3. One possible model is the following cell means model:

$$Y_{ij} = \mu_j + \varepsilon_{ij_{\mathrm{F}}}. \qquad \text{(3.47, repeated)}$$

However, Chapter 3 also presents the effects model as an alternative full model:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}. \qquad \text{(3.59, repeated)}$$

Chapter 3 shows that even though these two models have different forms, their sum of squared errors E_F will always equal one another. This equivalence holds even though the effects model has one more parameter than the cell means model, because both models have *a* independent parameters (it is not the case, however, throughout the realm of statistical models that any two models having the same number of parameters will necessarily be equivalent models).

Just as it is possible to write more than one ANOVA model to represent the data and yet discover that the two models are equivalent, there is also an entire collection of regression models that are also equivalent to these ANOVA models. To consider some examples of such equivalent models, we will begin with the general form of a linear model for one dependent variable, which Chapter 3 shows can be written as

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \cdots + \beta_p X_{p_i} + \varepsilon_i. \qquad \text{(3.1, repeated)}$$

Let’s consider the relationship between this general regression model and the ANOVA effects model:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}. \qquad \text{(3.59, repeated)}$$

To develop the relationship between these two models, we will start with the simplest case of two groups. Notice that we could write the ANOVA effects model in this case as

$$\text{For Group 1: } Y_{i1} = \mu + \alpha_1 + \varepsilon_{i1}, \qquad (1)$$

$$\text{For Group 2: } Y_{i2} = \mu + \alpha_2 + \varepsilon_{i2}. \qquad (2)$$

We can mimic this same result using regression analysis with three *X* variables:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \qquad (3)$$

where we define the *X* variables in the following way:

X_{0i} = 1 for all individuals,
X_{1i} = 1 for individuals in Group 1 and 0 for individuals in Group 2,
X_{2i} = 1 for individuals in Group 2 and 0 for individuals in Group 1.

Notice that with this definition of our 3 *X* variables, we can write our regression model from Equation 3 as

$$\text{For Group 1: } Y_i = \beta_0(1) + \beta_1(1) + \beta_2(0) + \varepsilon_i, \qquad (4)$$

$$\text{For Group 2: } Y_i = \beta_0(1) + \beta_1(0) + \beta_2(1) + \varepsilon_i. \qquad (5)$$

Simplifying Equations 4 and 5 yields

For Group 1: $Y_i = \beta_0 + \beta_1 + \varepsilon_i$, (6)
For Group 2: $Y_i = \beta_0 + \beta_2 + \varepsilon_i$. (7)

Notice that Equation 6 is identical to Equation 1, where $\beta_0 = \mu$ and $\beta_1 = \alpha_1$. Similarly, Equation 7 is identical to Equation 2, where $\beta_0 = \mu$ and $\beta_2 = \alpha_2$. In other words, by defining X variables in regression this way, we obtain a regression model whose parameters are identical to those of the ANOVA effects model.

This relationship between the ANOVA effects model and regression holds not only for two groups but also more generally for a groups. In this more general case, we continue to define $a + 1$ X variables. The general formulation is to define X variables according to the following rule:

$X_{0i} = 1$ for all individuals,
$X_{ji} = 1$ for individuals in Group j and 0 for all other individuals.

Notice that the X_0 variable, which we will refer to as a "unit variable" (because it assumes a constant value of 1 for all subjects), allows for a grand mean. There are then a additional X variables, one for each group.

If it were not for one complication, it might be possible to end this tutorial immediately. However, as we mentioned briefly in Chapter 3, there is a difficulty in working with either the ANOVA effects model or the comparable regression model. The problem is that we have $a + 1$ parameters, but only a groups. To understand why this is a problem, let's return to our simple example of only two groups. We will formulate the problem in terms of the regression models of Equations 6 and 7, but keep in mind that these two equations are formally identical to Equations 1 and 2, so, ultimately, we must deal with the problem we are about to explain regardless of whether we adopt an ANOVA or a regression approach to the data.

Equations 6 and 7 stipulate that scores on Y can be explained in terms of three parameters: β_0, β_1, and β_2. Specifically, our model specifies that (1) the mean Y score in Group 1 can be expressed as $\beta_0 + \beta_1$, and (2) the mean Y score in Group 2 can be expressed as $\beta_0 + \beta_2$. We can write these two statements more formally as

$$\mu_1 = \beta_0 + \beta_1, \tag{8}$$

$$\mu_2 = \beta_0 + \beta_2. \tag{9}$$

The problem is that we are allowing ourselves three parameters to explain two population means. Even if the two means are different, we will never need as many as three parameters to reproduce the population means. To see why, suppose that $\mu_1 = 40$ and $\mu_2 = 60$. What are the proper values of β_0, β_1, and β_2? It turns out that there are infinitely many possible values of these three parameters, all of which succeed perfectly in reproducing the population means. We don't have space to list all of the possibilities (listing them all would require an infinite number of pages!), but here are some examples:

$$\beta_0 = 50, \beta_1 = -10, \text{ and } \beta_2 = 10,$$

$$\beta_0 = 0, \beta_1 = 40, \text{ and } \beta_2 = 60,$$

$$\beta_0 = 60, \beta_1 = -20, \text{ and } \beta_2 = 0,$$

$$\beta_0 = 100, \beta_1 = -60, \text{ and } \beta_2 = -40.$$

Although these four possibilities may look very different from one another, they all share an important property. Namely, in all four cases, the sum of β_0 and β_1 equals 40 and the sum of β_0 and β_2 equals 60. Thus all four sets of parameter values imply the correct population means of our two groups. The problem is that we have more parameters than we need. With two groups, we need only two parameters, not three. The problem is essentially the same with more than two groups. In general, with *a* groups, our effects model has $a + 1$ parameters, but we need only *a* parameters. Also notice that on a practical note, we have considered the problem in terms of population parameters, but we must confront the exact same problem in a sample. For example, if we have two sample means of 40 and 60, we still do not need three parameter estimates to explain these two values. The same four sets of values shown earlier when used as parameter estimates would all perfectly reproduce the sample means of 40 and 60, so it would be impossible to identify unique values for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. If we cannot find unique values for these parameter estimates, it becomes impossible to interpret the meaning behind the parameters. Without being able to interpret the parameter estimates, we cannot hope to understand what properties of our data are reflected in these parameters. Thus we must find some way to get around this problem.

To solve this problem, we need to obtain unique values for parameters (and in a sample, unique values for parameter estimates). The solution is to reduce the number of parameters by 1. The way we do this is by establishing a "side condition," as we mentioned in Chapter 3. It turns out that there are a variety of possible side conditions we might impose. Which one we choose does not affect the overall fit of the model, but does affect the meaning of the specific parameters of the model. To explore how this works, we will consider three different side conditions we might impose on our parameters. As we will explain later, each of these three types of side conditions has its own advantages and disadvantages, so there is some value in becoming familiar with all three. As before, we will continue to examine these side conditions from the perspective of regression analysis, in particular the β parameters of the regression model.

Reference Cell Model

The first side condition we will consider is to constrain β_a, the parameter associated with the last group, to be zero. Intuitively, the idea is that we need to reduce the number of parameters by 1, so why not simply eliminate the last parameter. Stated differently, we have $a + 1$ X variables in our model, but we need only *a* variables, so it seems reasonable to drop the final variable. It turns out that this is exactly what we are doing when we constrain β_a to equal zero.

To understand what is happening here, let's reconsider the simple case of two groups. Recall from Equations 8 and 9 that in the case of two groups, we can express the relationship between cell means and regression parameters as

$$\mu_1 = \beta_0 + \beta_1, \tag{8, repeated}$$

$$\mu_2 = \beta_0 + \beta_2. \tag{9, repeated}$$

However, we will now impose our constraint that the parameter associated with the last group equals zero. In the case of two groups, our constraint would be that $\beta_2 = 0$, because $a = 2$. Thus we can simplify Equations 8 and 9 as

$$\mu_1 = \beta_0 + \beta_1, \tag{10}$$

$$\mu_2 = \beta_0. \tag{11}$$

Equations 10 and 11 show us that we now have unique values and thus unique interpretations for β_0 and β_1, the two parameters in our model. Namely, Equation 11 says that β_0 is the population mean of Group 2. Notice that we have to be careful here to realize that the intercept of our regression analysis does not represent the grand mean, but instead represents the mean of Group 2. This happens because in general the intercept is the population value of Y when all X variables equal 0, and in this case, X_1 (our only predictor variable with two groups) is zero for individuals in Group 2. The meaning of β_1 becomes clear if we rewrite Equation 10 as

$$\beta_1 = \mu_1 - \beta_0. \tag{12}$$

Because we know from Equation 11 that $\mu_2 = \beta_0$, we can further rewrite Equation 12 as

$$\beta_1 = \mu_1 - \mu_1. \tag{13}$$

Equation 13 reveals one of the advantages of this particular side condition—namely, that β_1, the parameter associated with our X_1 predictor, represents the difference between the population means of Groups 1 and 2. This is especially convenient because if we want to test a null hypothesis that $\mu_1 = \mu_2$, we can simply test a null hypothesis that $\beta_1 = 0$ in our regression model. Similarly, if we want to form a confidence interval for $\mu_1 - \mu_2$, we can form a confidence interval for β_1 using regression analysis.

Another advantage of this choice of side condition is that it is trivial to implement. By saying that β_a equals zero, we are effectively omitting X_a as a predictor in our model. To see why, let's look again at Equation 3 for two groups:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \tag{3, repeated}$$

Imposing the side condition that $\beta_2 = 0$ (recall that β_a is the same as β_2 when $a = 2$) implies that we can write the model as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + (0)X_{2i} + \varepsilon_i, \tag{14}$$

which further simplifies to

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \tag{15}$$

because X_{2i} drops out when multiplied by zero (notice also that X_{0i} has disappeared in Equation 15 because we have substituted 1 for X_{0i}, knowing that $X_{0i} = 1$ for all individuals). Notice that the only difference between Equation 15 and Equation 3 is that the X_2 variable has been omitted from Equation 15. Thus, in a practical sense, we can impose the side condition that $\beta_2 = 0$ by omitting X_2 from our model. In other words, we fit a regression model where we intentionally omit the variable representing Group 2. Instead of including two predictor variables in our regression model, when we have two groups, we include only one predictor (although keep in mind that we have also included the unit variable X_0 in our model).

What about the more general case of a groups? Exactly the same logic holds here as well. We simply omit X_a as a predictor. As a consequence, β_0, the intercept term in the model, becomes equal to the population mean of the final group—i.e,

$$\beta_0 = \mu_a. \tag{16}$$

In addition, the β parameter associated with a given predictor variable becomes equal to the difference between the population mean of that group and Group a. In symbols,

$$\beta_j = \mu_j - \mu_a. \tag{17}$$

Notice then that each β_j parameter compares one of the first $a - 1$ group means to the mean of the last group. For this reason, the model resulting from this side condition is often referred to as a "reference cell" model, where Group a is the reference group against which all other groups are compared. In some situations, such as when one group is a control group and all other groups are experimental groups, the parameters have a very natural interpretation. However, in other situations, it may seem artificial to single out one group and compare it to all of the others. Even so, it turns out that this specific side condition offers additional advantages, which is why it forms the basis of the approach used by many general linear model routines, such as the parameterization resulting from a CLASS statement in SAS PROC GLM and SAS PROC MIXED. We will return to this issue later in this tutorial. We should also mention that this method of coding variables to represent group membership is often referred to in the literature as "dummy coding."

One additional point regarding the reference cell model will prove to be helpful for understanding its extension to factorial designs. Remember that the regression intercept β_0 in this model equals μ_a. Implicitly, this leads to a new definition of μ in ANOVA notation. Instead of automatically interpreting μ as the grand mean averaged over all groups, μ effectively becomes μ_a, the mean of Group a, in the reference cell model. As usual, we can still define an effect for group j to be of the form

$$\alpha_j = \mu_j - \mu.$$

However, the meaning of an "effect" from this perspective is clearly different than it is in the ANOVA effects model. In the reference cell model, the effect of group j is the mean difference between it and Group a, not the mean difference between group j and the grand mean. This definition has two important implications. First, the effect of Group a is now necessarily equal to zero. Notice that this is literally the constraint we have imposed on the effect parameters in order to reduce the number of parameters to be estimated from $a + 1$ to a. Second, the sum of the effects typically no longer equals zero. Both of these points will become important when we examine the reference cell model for a factorial design.

Cell Means Model

The second side condition we will consider is to constrain β_0 to equal zero. In a moment, we will describe how to operationalize this constraint, but first, we will examine its meaning in the context of two groups. Recall from Equations 8 and 9 that we can express the relationship between cell means and regression parameters as

$$\mu_1 = \beta_0 + \beta_1, \tag{8, repeated}$$

$$\mu_2 = \beta_0 + \beta_2. \tag{9, repeated}$$

However, if we impose a constraint that $\beta_0 = 0$, we can simplify these equations as

$$\mu_1 = \beta_1, \tag{18}$$

$$\mu_2 = \beta_2. \tag{19}$$

Notice that with this constraint in place, each of the remaining β parameters simply becomes equal to a population mean. In the more general case of a groups, there would be a β parameters, each of which would equal a population mean. For this reason, we say that imposing a side condition where $\beta_0 = 0$ results in a “cell means” model.

Using the cell means model from a regression perspective requires us to constrain β_0 to equal zero. The simplest way to do this is to realize that by saying that β_0 equals zero, we are effectively omitting X_0 as a predictor in our model. To see why, let’s look again at Equation 3:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \qquad (3, \text{repeated})$$

Imposing the side condition that $\beta_0 = 0$ implies that we can write the model as

$$Y_i = (0)X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \qquad (20)$$

which further simplifies to

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \qquad (21)$$

Notice that the only difference between Equation 21 and Equation 3 is that the X_0 variable has been omitted from Equation 21. Thus, in a practical sense, we can impose the side condition that $\beta_0 = 0$ by omitting X_0 from our model. In other words, we fit a regression model where we intentionally omit the intercept term. Although this may seem odd, most regression routines in standard statistical packages provide omission of the intercept term as an option.

An advantage of the cell means model is that each parameter of the model has a clear interpretation. We can directly interpret any given β_j as the population mean of group j. When applied to sample data, estimated regression coefficients, β_j, are simply sample means $\bar{Y}_j$ for group j. Despite this ease of interpretation, the cell means model suffers from two disadvantages when viewed from a regression perspective. First, although each individual parameter has a clear interpretation, it can be tedious to examine differences among parameters. In other words, our usual questions are not literally about single population means, but instead about differences between population means. To answer these questions in the cell means model, we must form differences of parameters, whereas other choices of side conditions result in parameters that are already expressed as mean differences. This is not an insurmountable difficulty, but essentially requires the availability of a regression program that allows tests and confidence intervals for linear combinations of regression coefficients. Second, as a minor point, we have already noted that the cell means formulation requires that we omit the intercept term from our model. This also is not insurmountable, but it is unconventional, so it requires care in implementing.

Effects Model

The third side condition we will consider is already familiar from Chapter 3. You may recall that the side condition we imposed there was that the sum of effects added across all groups equal zero, which we can write in symbols as

$$\sum_{j=1}^{a} \alpha_j = 0 . \qquad (3.60, \text{repeated})$$

We can rewrite this side condition in a notation for regression by remembering that each regression coefficient β_j equals a corresponding ANOVA effect α_j (as we saw in our comparison of Equations 1 and 6, as well as Equations 2 and 7). With this substitution, Equation 3.60 becomes

$$\sum_{j=1}^{a} \beta_j = 0\,. \tag{22}$$

To understand this third type of side condition, let's begin with the two-group case. Recall from Equations 8 and 9 that in the special case of two groups, we can express the relationship between cell means and regression parameters as

$$\mu_1 = \beta_0 + \beta_1, \tag{8, repeated}$$

$$\mu_2 = \beta_0 + \beta_2. \tag{9, repeated}$$

However, we will now impose our constraint that the sum of the β_j parameters equals zero. In the case of two groups, our constraint would be

$$\beta_1 + \beta_2 = 0, \tag{23}$$

which we can rewrite as

$$\beta_2 = -\beta_1. \tag{24}$$

By substituting $-\beta_1$ for β_2, we can simplify Equations 8 and 9 as

$$\mu_1 = \beta_0 + \beta_1, \tag{25}$$

$$\mu_2 = \beta_0 - \beta_1. \tag{26}$$

We now need to consider the interpretation of our two parameters (i.e., β_0 and β_1) in this formulation. We will begin with β_0. Suppose we were to add together Equations 25 and 26. The result would be

$$\mu_1 + \mu_2 = 2\beta_0, \tag{27}$$

which we can simply rewrite as

$$\beta_0 = (\mu_1 + \mu_2)\,/\,2. \tag{28}$$

The right side of Equation 28 is just the grand mean in the population, which we have previously designated as μ. Thus we can express Equation 28 as

$$\beta_0 = \mu, \tag{29}$$

which in words means that with this side condition, we can interpret β_0 in our regression model as the grand mean μ from an ANOVA perspective.

Notice that the only remaining parameter in our model is β_1. How can we interpret it? To answer this question, we can rewrite Equation 25 as

$$\beta_1 = \mu_1 - \beta_0. \tag{30}$$

Because we now know from Equation 29 that $\beta_0 = \mu$, we can substitute μ for β_0 in Equation 30, yielding

$$\beta_1 = \mu_1 - \mu. \tag{31}$$

This reveals that we can interpret β_1 as the difference between the population mean of Group 1 and the population grand mean. In other words, β_1 is literally the “effect” associated with Group 1, where we understand that “effect” refers to a deviation from the grand mean. Rewriting Equation 26 shows us that we can also use β_1 to find the “effect” associated with Group 2 because

$$-\beta_1 = \mu_2 - \mu \tag{32}$$

so that by changing the sign of β_1 we obtain the “effect” of Group 2—i.e., the deviation of the population mean of Group 2 from the grand mean. Most importantly, Equations 29 and 31 show us that we now have unique values and thus unique interpretations for β_0 and β_1, the two parameters in our model. Furthermore, Equation 31 reveals one of the advantages of this particular side condition, namely that the β_1 parameter has a straightforward interpretation as the “effect” associated with Group 1. In a moment, we will extend this side condition to the general case of *a* groups, but first let’s see how to implement this side condition in a regression model for 2 groups.

To see how to operationalize this third form of side condition in regression analysis with two groups, let’s return to Equation 3:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \tag{3, repeated}$$

Imposing the side condition that $\beta_2 = -\beta_1$ (from Equation 24) allows us to express Equation 3 as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} - \beta_1 X_{2i} + \varepsilon_i, \tag{33}$$

which we can then rewrite as

$$Y_i = \beta_0 X_{0i} + \beta_1 (X_{1i} - X_{2i}) + \varepsilon_i. \tag{34}$$

Implicit in Equation 34 is that we have redefined the predictor variable to be included in our model. Instead of including X_1 and X_2 as two separate predictor variables, Equation 34 tells us that we now should include a single predictor variable where the score for an individual on this new variable is literally the difference between that individual’s score on our original X_1 and our original X_2. To figure out what this means, we need to return to the way we defined our original X variables. Remember that we originally defined our X variables in the following way:

X_{0i} = 1 for all individuals,
X_{1i} = 1 for individuals in Group 1 and 0 for individuals in Group 2,

X_{2i} = 1 for individuals in Group 2 and 0 for individuals in Group 1.

According to Equation 34, we should continue to include in our model an X_0 variable coded as 1 for each individual. However, instead of separate X_1 and X_2 variables, we should form a single predictor created as X_1 minus X_2. Now, here is the key point from a practical perspective. Notice that a predictor defined in this way will have the following property:

$X_{1i} - X_{2i} = 1 - 0$ for individuals in Group 1,
$X_{1i} - X_{2i} = 0 - 1$ for individuals in Group 2.

In words, we need to code our single predictor so that it has a value of 1 for individuals in Group 1 but a value of −1 for individuals in Group 2. Using this method of coding imposes a constraint on our original parameters so that the sum of the ANOVA "effects" equals zero.

We are now ready to generalize this third type of side condition to the case of *a* groups. The same logic continues to apply in the more general case, but the details become slightly more complicated. Recall that with *a* groups, we can write our constraint as

$$\sum_{j=1}^{a} \beta_j = 0 . \qquad \text{(3.60, repeated)}$$

In general, we know that

$$\mu_j = \beta_0 + \beta_j. \qquad (35)$$

To understand the meaning of the β_0 parameter in our regression model, we can sum both sides of Equation 35 across groups, which results in

$$\sum_{j=1}^{a} \mu_j = a\beta_0 + \sum_{j=1}^{a} \beta_j \ . \qquad (36)$$

It immediately follows from Equation 3.60 that Equation 36 simplifies to

$$\sum_{j=1}^{a} \mu_j = a\beta_0 , \qquad (37)$$

which we can rewrite as

$$\beta_0 = \frac{\sum_{j=1}^{a} \mu_j}{a} = \mu . \qquad (38)$$

Thus the β_0 parameter in our regression model is the population grand mean. To interpret the remaining parameters in the regression model, we can simply substitute μ for β_0 in Equation 35, and move β_j to the left side of the equation, yielding

$$\beta_j = \mu_j - \mu, \qquad (39)$$

which shows that the remaining β_j parameters are ANOVA "effect" parameters.

To understand how we need to code our predictor variables in the general case of *a* groups, recall that the general form of a linear model is given by

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_p X_{pi} + \varepsilon_i. \qquad \text{(3.1, repeated)}$$

When we have a groups, we will have a predictors plus X_0, so we can rewrite Equation 3.1 as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_a X_{ai} + \varepsilon_i\,. \qquad (40)$$

However, we know from Equation 3.60 that

$$\beta_a = -\sum_{j=1}^{a-1} \beta_j\,,$$

which means that we can express Equation 40 as

$$Y_i = \beta_0 X_{0_i} + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \cdots - \left(\sum_{j=1}^{a-1} \beta_j \right) X_{a_i} + \varepsilon_i\,. \qquad (41)$$

We can now rewrite Equation 41 as

$$Y_i = \beta_0 X_{0i} + \beta_1 (X_{1i} - X_{ai}) + \beta_2 (X_{2i} - X_{ai}) \ + \beta_3 (X_{3i} - X_{ai}) \ + \cdots + \beta_{a-1} (X_{a-1i} - X_{ai}) + \varepsilon_i. \quad (42)$$

Equation 42 reveals two important points: (1) our regression model now contains $a - 1$ predictors along with an intercept term, and (2) each predictor is an original predictor minus X_a.

To understand the practical implications of this second point, let's consider the coding of $X_{1i} - X_{ai}$ for individuals in each group. This predictor will have the following property:

$X_{1i} - X_{ai} = 1 - 0$ for individuals in Group 1,
$X_{1i} - X_{2i} = 0 - 0$ for individuals in Group 2 through a – 1,
$X_{1i} - X_{ai} = 0 - 1$ for individuals in Group a.

In words, we need to code our new predictor so that it has a value of 1 for individuals in Group 1, a value of 0 for individuals in all other groups except the last group, and a value of −1 for individuals in the last group. More generally, we could say that we need to code predictor variable X_j so that individuals in Group j receive a value of 1, individuals in Group a receive a value of −1, and individuals in all other groups receive a score of 0. This method of coding ensures us that the resulting regression parameters will reflect ANOVA "effect" parameters.

Numerical Example

At this point it may be helpful to provide a numerical example to illustrate the theoretical developments we have presented. Table T3.1 reproduces the data originally shown in the tutorial on regression. The sample size in this example is smaller than should be used in an actual study, but it allows ease of calculation and interpretation. Table T3.2 illustrates the three different coding methods we have presented. For example, consider the reference cell model. It contains a unit variable, where each individual receives a score of 1. The remaining $a - 1$ predictor variables are all scored either zero or one. In particular, a "1" appears on variable X_j if and only if an individual belongs to Group j. Finally, notice that we have included only $a - 1$ predictor variables (not counting the unit variable), because we have omitted the predictor for the final group. Next, consider the cell means model. Notice that it does not contain a unit variable. Otherwise, the coding is identical to that for the reference cell model except that here we also include a predictor for the final group. The final third of the table illustrates coding for the effects model. Notice that like the reference cell model, here we include a unit variable. In fact, the only difference from the reference cell model is that in the effects model, the individuals in the final group receive a value of −1 on every predictor variable (except of course they receive a score of 1 on the unit variable, by definition).

TABLE T3.L
DATA TO ILLUSTRATE REGRESSION
APPROACH TO ANOVA

Group			
1	2	*3*	*4*
11	13	15	17
13	15	17	19

TABLE T3.2
THREE CODING SCHEMES FOR TABLE T3.L DATA

	Reference Cell Model				
Group	Y	X_0	X_1	X_2	X_3
1	11	1	1	0	0
1	13	1	1	0	0
2	13	1	0	1	0
2	15	1	0	1	0
3	15	1	0	0	1
3	17	1	0	0	1
4	17	1	0	0	0
4	19	1	0	0	0
	Cell Means Model				
Group	Y	X_1	X_2	X_3	X_4
1	11	1	0	0	0
1	13	1	0	0	0
2	13	0	1	0	0
2	15	0	1	0	0
3	15	0	0	1	0
3	17	0	0	1	0
4	17	0	0	0	1
4	19	0	0	0	1
	Effects Model				
Group	Y	X_0	X_1	X_2	X_3
1	11	1	1	0	0
1	13	1	1	0	0
2	13	1	0	1	0
2	15	1	0	1	0
3	15	1	0	0	1
3	17	1	0	0	1
4	17	1	−1	−1	−1
4	19	1	−1	−1	−1

Table T3.3 summarizes parameter estimates and sums of squared errors for the various ANOVA and regression models. We encourage readers to duplicate these results, especially for those with access to ANOVA and regression procedures in statistical packages.

Several points illustrated in Table T3.3 deserve special mention. First, notice that all six of these models are equivalent to one another in the sense that the sum of squared errors for each of these full models equals 8. Second, the reason for this equivalence is revealed in the column entitled "Predicted Cell Means." Each and every one of these models ultimately results in the same predicted score for any individual. In other words, regardless of how the model is formulated, the predicted score for someone in Group 1 will be 12, the predicted score for someone in Group 2 is 14, and so forth. Because the models all make the same predictions, they are equivalent to one another. Third, notice that the specific parameter estimates are not necessarily equal to one another. Although for any given ANOVA model there is always a regression model that has the same parameter estimates, the estimates of the effects model, cell means model, and reference cell model are different from one another. These differences reflect the fact that the meaning of specific parameters depends on the specific way in which we have chosen to formulate the model. However, there are certain patterns in the parameter estimates that manifest themselves across all these models. For example, consider the difference between the parameter estimate associated with Group 1 and the corresponding estimate associated with Group 2. In all six cases, the difference between these two parameter estimates is −2. Such a difference is called an "estimable function" and is meaningful because its value and interpretation remain the same regardless of how we chose to parameterize the model. We will say more about estimable functions later, but for now we will simply say that any contrast of the means (or, correspondingly, of the effects) whose coefficients sum to zero will necessarily be an estimable function. Thus when we are interested in estimating and testing differences among group means, such as by formulating contrasts, we will eventually reach the same conclusion regardless of how we originally parameterized the model. Fourth, it is extremely important to realize that the numerical values of the coded predictor variables do not in general directly represent contrast coefficients. For example, consider X_1 for effects coding as shown in Table T3.2. The values of this variable are 1 for Group 1, 0 for Group 2, 0 for Group 3, and −1 for Group 4. A naive interpretation would be that this variable reflects the difference between the means of Groups 1 and 4 (i.e., that β_1 equals $\mu_1 - \mu_4$). However, we have already seen that the actual meaning of this variable is that it indicates the "effect" of Group 1 so that in reality it reflects the difference between the mean of Group 1 and the grand mean. In other words, this variable corresponds to a contrast that compares Group 1 to the average of all (other) groups, which we could represent as a contrast with coefficients of 1, −1/3, −1/3, and−1/3 for Groups 1, 2, 3, and 4, respectively. Notice that there is no direct relationship between the X values of 1, 0, 0, −1, and the coefficients of the contrast represented by this variable. This lack of a direct relationship opens the door for possible confusion and misinterpretation, so we now feel the need to consider the topic of contrasts in some additional detail.

Contrasts in Single-Factor Between-Subjects Designs

So far we have verified that regression can duplicate ANOVA for the omnibus null hypothesis in a single-factor between-subjects design. This next section investigates the connection between ANOVA and regression for contrasts. Specifically, we will see that although the connection is sometimes more difficult to operationalize, once again it is possible to form regression models that are equivalent to ANOVA models for testing contrasts. Our emphasis here is on hypothesis testing, but the results also apply to confidence intervals.

Two fundamentally different but ultimately equivalent approaches exist to test contrasts with regression models. The first of these approaches involves coding predictors so as to obtain tests of specific contrasts directly as tests of specific regression coefficients. The second approach relies not on coding, but instead expresses any contrast of interest as a specific linear combination of the regression coefficients. We will briefly illustrate each of these approaches, and will touch upon some of the complexities that can arise, especially in the first approach.

TABLE T3.3
PARAMETER ESTIMATES AND SUM OF SQUARED ERRORS FOR VARIOUS ANOVA AND REGRESSION MODELS

	E_F	*Parameter Estimates*	*Predicted Cell Means*
		ANOVA Models	
Effects	8	$\hat{\mu} = 15,\ \hat{\alpha}_1 = -3,\ \hat{\alpha}_2 = -1, \hat{\alpha}_3 = 1$	$\hat{Y}_1 = 12, \hat{Y}_2 = 14, \hat{Y}_3 = 16, \hat{Y}_4 = 18$
Cell Means	8	$\hat{\mu}_1 = 12,\ \hat{\mu}_2 = 14, \hat{\mu}_3 = 16, \hat{\mu}_4 = 18$	$\hat{Y}_1 = 12, \hat{Y}_2 = 14, \hat{Y}_3 = 16, \hat{Y}_4 = 18$
Reference Cell	8	$\hat{\mu} = 18,\ \hat{\alpha}_1 = -6,\ \hat{\alpha}_2 = -4, \hat{\alpha}_3 = -2$	$\hat{Y}_1 = 12, \hat{Y}_2 = 14, \hat{Y}_3 = 16, \hat{Y}_4 = 18$
		Regression Models	
Effects	8	$\hat{\beta}_0 = 15, \hat{\beta}_1 = -3, \hat{\beta}_2 = -1, \hat{\beta}_3 = 1$	$\hat{Y}_1 = 12, \hat{Y}_2 = 14, \hat{Y}_3 = 16, \hat{Y}_4 = 18$
Cell Means	8	$\hat{\beta}_1 = 12, \hat{\beta}_2 = 14, \hat{\beta}_3 = 16, \hat{\beta}_4 = 18$	$\hat{Y}_1 = 12, \hat{Y}_2 = 14, \hat{Y}_3 = 16, \hat{Y}_4 = 18$
Reference Cell	8	$\hat{\beta}_0 = 18, \hat{\beta}_1 = -6, \hat{\beta}_2 = -4, \hat{\beta}_3 = -2$	$\hat{Y}_1 = 12, \hat{Y}_2 = 14, \hat{Y}_3 = 16, \hat{Y}_4 = 18$

We will begin by considering how appropriate coding of predictor variables can be used in regression to test contrasts of interest. You may be surprised to learn that various regression books suggest different strategies for testing contrasts. In part, this variety reflects the impressive flexibility of the regression model, but it may also reflect a trade-off between simplicity and generality. Authors are faced with a choice of either presenting simple approaches that are often but not always appropriate, or else presenting more complicated approaches that are always appropriate.

We have chosen to present one approach for pairwise comparisons and a second approach for complex comparisons. In principle, the approach we describe for complex comparisons could also be used for pairwise comparisons, so it fulfills the criterion of generality but does so at the expense of increased complexity. As such, we should stress that the choice between these two approaches and yet other possibilities described in other sources is ultimately a subjective decision.

We will continue to use the data from Table T3.1 to illustrate contrasts. Recall that we have equal *n* in this example. Certain methods of testing contrasts work fine with equal *n*, but produce erroneous results with unequal *n*. In order to demonstrate that the approaches we present here produce accurate results with unequal *n*, we have chosen to modify the data shown in Table T3.1 In particular, we will suppose now that there are a total of four observations in the first group. For simplicity, we will assume that these four individuals have scores of 11, 11, 13, and 13. Notice that the sample mean for the first group is still 12, as it was for the original data. For future reference, we will refer to this unequal *n* version of our data as our "augmented" data set to distinguish it from our original equal *n* example.

Pairwise Contrasts via Coding

Suppose a researcher wants to use the augmented data set to test whether Groups 1 and 2 have equal population means. From an ANOVA perspective, Equation 4.11 shows that the sum of squares for this contrast equals 5.33. The value of MS_w is 1.67, so from Equation 4.16 the F statistic for this contrast equals 3.20.

How can we obtain the correct F statistic for this contrast using regression? There is more than one way to accomplish this goal, but the method we will present relies on reference cell coding. Specifically, we can compare Groups 1 and 2 simply by making either group the reference cell in our coding scheme. The top portion of Table T3.4 shows a coding scheme where we have chosen Group 1 as our reference group. The middle portion of Table T3.4 shows an excerpt from the output using PROC REG to analyze our data with the aforementioned coding scheme. Notice that the estimated coefficient for X_1 is 2.0, which is simply the difference between the sample means for Groups 1 and 2. The t value for this coefficient is 1.79, which is the square root of our earlier F value of 3.20. As further corroboration, the bottom of Table T3.4 shows output from PROC GLM, where we have used an ESTIMATE command to estimate $\mu_2 - \mu_1$. The estimated value is 2.0, which is simply the difference in sample means, and the corresponding t value is 1.79, once again confirming that the regression analysis produced the desired result. Thus, even with unequal n, we can use the reference cell coding method to compare Groups 1 and 2.

What if we also wanted to compare additional pairs of groups? The output from our regression analysis shows that we already have information about the comparisons of Groups 1 and 3 as well

TABLE T3.4
PAIRWISE CONTRASTS VIA CODING

	Reference Cell Coding for Augmented Data				
Group	Y	X_0	X_1	X_2	X_3
1	11	1	0	0	0
1	11	1	0	0	0
1	13	1	0	0	0
1	13	1	0	0	0
2	13	1	1	0	0
2	15	1	1	0	0
3	15	1	0	1	0
3	17	1	0	1	0
4	17	1	0	0	1
4	19	1	0	0	1

Output from SAS PROC REG Parameter Estimates

Parameter Estimates					
Variable	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr* > |*t*|
Intercept	1	12.00000	0.64550	18.59	<.0001
Xl	1	2.00000	1.11803	1.79	.1238
X2	1	4.00000	1.11803	3.58	.0117
X3	1	6.00000	1.11803	5.37	.0017

Output from SAS PROC GLM for Comparing First Two Groups

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr* >|*t*|
mul vs mu2	2.00000000	1.11803399	1.79	0.1238

as Groups 1 and 4. However, just as obvious is that we have learned nothing of direct interest about any of the pairwise differences that do not involve Group 1. Unfortunately, we cannot simply do one regression analysis with all $a(a - l)/2$ predictor variables (except in the uninteresting case where a = 2) because that would give us more than $a - 1$ predictors in one equation. Instead, what we must do is cycle through $a - 1$ choices of which group is used as the reference group. If α is very large, this is clearly a tedious process, which is one reason in practice that there are obvious advantages to using software that does not require us to code our own predictor variables for a regression analysis.

Complex Contrasts Via Coding

Suppose a researcher wants to consider a complex comparison. For example, suppose he or she wants to use the augmented data set to test the difference between the mean of Group 4 and the average of the other three groups. Once again, there is more than one way to do this in regression, but we will present a general approach that works with either equal or unequal *n*.

We will describe this approach in terms of four steps:

1. Write the contrast to be tested in terms of its coefficients, as in Chapter 4. For example, in our case, the coefficients would be −1/3, −1/3, and −1/3, and 1 for Groups 1, 2, 3, and 4, respectively.
2. Create $a - 2$ additional contrasts orthogonal to the contrast of interest. Whenever $a > 3$, there are multiple sets of orthogonal contrasts. A simple pair in our case would be one contrast with coefficients of −1, 1, 0, and 0 and another contrast with coefficients of −1, −1, 2, and 0. Even if only one contrast is of ultimate interest, it is necessary to include $a - 1$ predictors in the regression analysis, especially with unequal *n*.
3. Each contrast becomes a predictor variable in the regression analysis. For example, scores on X_1 are based on the coefficients of the first contrast, scores on X_2 are based on the coefficients of the second contrast, and so forth. In particular, consider scores on X_1. Each member of Group 1 is assigned a score on X_1 equal to the coefficient for Group 1 in the first contrast. In our example, this score would be −1/3. Similarly, each member of Group 2 is assigned a score on X_1 equal to the coefficient for Group 2 in the first contrast. In our example, this score is once again −1/3. This process proceeds until scores have been assigned for each individual on each predictor variable.
4. One more step is necessary if we want our regression coefficients to be interpretable as differences in means. To do so, we need to consider two values: (1) the value of our negative coefficient and (2) the value of the positive coefficient (our approach presumes that every group receiving a negative coefficient has the same value for that coefficient, and the same is true for the positive coefficients, although the absolute value of the negative and positive coefficients can be different). We need to scale each predictor variable so that the difference between the negative score and the positive score equals 1. For example, so far, our values for X_1 are −1/3, −1/3, −1/3, and 1. The difference between the negative value of −1/3 and the positive value of 1 is 4/3. Thus we need to divide each value by 4/3 (i.e., multiply by 3/4). Doing so yields our final codes for X_1: −1/4, −1/4, −1/4, and 3/4 for individuals in Groups 1, 2, 3, and 4, respectively. We need to make three additional points. First, the reason we need to rescale in this way in order to interpret each coefficient as a mean difference is that a regression coefficient depicts the change in Y divided by the change in X. If we code X so that its change is 1, then the coefficient will simply reflect the change in Y. Here the change in X is simply the difference between the negative value and the positive value, so we need to scale our predictors so that this difference is 1. Second, this algorithm for rescaling works only if there are at most three distinct values of coefficients: a single negative value applied

to one or more groups, a single positive value applied to one or more groups, and a value of 0 applied to one or more groups. A counterexample would be coefficients of 1/3, 2/3, −1, and 0. Rescaling weighted averages such as this requires more complex methods beyond the scope of this tutorial. Third, keep in mind that rescaling is unnecessary if you only want to test hypotheses. Rescaling becomes relevant only when you want to estimate values of mean differences, often in conjunction with confidence intervals.

The top portion of Table T3.5 shows the end result of applying these four steps to our augmented data. The middle of the table shows an excerpt of the output from the corresponding regression analysis. Notice that the estimated value for the contrast of interest is 4.0, as reflected by the estimated value for the coefficient associated with X_1. We can tell this is exactly what it should be by realizing that the mean of Group 4 is 18, while the average of the means of the other three groups is 14 (namely, the unweighted average of 12, 14, and 16; a different approach is needed in the less typical case where we want a weighted average). The t value for the contrast equals 3.88, which implies a p value of .0082. Notice that the regression results for X_3 here agree exactly with what we saw earlier for the comparison of Groups 1 and 2. Finally, the bottom portion of the table shows an excerpt of results obtained from PROC GLM. As before, this output confirms the results we obtained using regression analysis.

A Good Idea Gone Bad

Let's return to our example where the only question of interest is a comparison of Groups 1 and 2. A simple method of coding suggests itself for a regression approach to this question. It seems natural to create a predictor variable coded as follows:

$$X_{1i} = \begin{cases} 1 & \text{if the individual belongs to Group 1} \\ -1 & \text{if the individual belongs to Group 2}\,. \\ 0 & \text{otherwise} \end{cases}$$

Because we are interested only in the contrast of Groups 1 and 2, we could form a regression model that uses only X_1 as a predictor. Will this yield the same conclusion about the difference between the means of Groups 1 and 2 as the more complicated model that includes all three coded predictors?

As long as we have equal n, things are fairly straightforward. It turns out that this single predictor variable accounts for a sum of squares equal to 4 in the equal n version of our data (i.e., the original data shown in Table B.l), just as in the ANOVA approach. Furthermore, the estimated coefficient for X_1 is −1, just as we would hope when the sample mean of the first group is 12, the sample mean of the second group is 14, and the two groups have been coded as 1 and −1, respectively. However, one immediate complication arises here because we must consider the nature of our full model. The ANOVA full model for these data has four parameters and allows each group to have its own mean. However, if we consider the regression model with only X_1 to be the full regression model, we will not obtain the same result we did in ANOVA. Specifically, for these data, the ANOVA full model has a sum of squared errors of 8 and degrees of freedom of 4. The regression model with only X_1 as a predictor, on the other hand, has a sum of squared errors of 44 and degrees of freedom of 7. As a consequence, the F statistic for regression here equals 0.64, considerably smaller than the ANOVA value of 2.00. However, regression can duplicate the ANOVA value of 2.00 by reformulating the full model as one that contains all four parameters. One way of accomplishing this goal is to form two

TABLE T3.5
COMPLEX CONTRASTS VIA CODING

	Complex Contrast Coding for Augmented Data				
Group	*Y*	X_0	X_1	X_2	X_3
1	11	1	−1/4	−1/3	−1/2
1	11	1	−1/4	−1/3	−1/2
1	13	1	−1/4	−1/3	−1/2
1	13	1	−1/4	−1/3	−1/2
2	13	1	−1/4	−1/3	1/2
2	15	1	−1/4	−1/3	1/2
3	15	1	−1/4	2/3	0
3	17	1	−1/4	2/3	0
4	17	1	3/4	0	0
4	19	1	3/4	0	0

Output from SAS PROC REG Parameter Estimates

Parameter Estimates							
Variable	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr>* |*t*|	*Type ISS*	*Type II SS*
Intercept	1	15.00000	0.42696	35.13	<.0001	2,073.60000	2,057.14286
Xl	1	4.00000	1.03190	3.88	.0082	32.40000	25.04348
X2	1	3.00000	1.07044	2.80	.0311	16.66667	13.09091
X3	1	2.00000	1.11803	1.79	.1238	5.33333	5.33333

Output from SAS PROC GLM

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr>*|*t*|
Four vs. other three	4.00000000	1.03189865	3.88	0.0082

separate regression models, one of which uses only X_1 as a predictor and the other of which uses $a - 1$ predictors, such as in the effects or reference cell full models. Alternatively, a single analysis with $a - 1$ predictors can be undertaken as long as the additional $a - 2$ predictors form an orthogonal set with the X_1 predictor of interest. The choice between these two approaches is a matter of convenience and personal preference because both will yield the desired *F* value of 2.00 for these data.

Both regression alternatives are somewhat cumbersome, but still relatively straightforward. Unfortunately, however, things become more complicated when sample sizes are unequal. For example, let's now see what happens in our augmented data set with unequal *n*. Recall that we have seen earlier from Equation 4.18 that the *F* statistic for this contrast equals 3.20. Now suppose that we use the same natural coding we used previously to compare Groups 1 and 2. In the augmented data, the sum of squares accounted for by X_1 equals 13.83, more than twice as large as the value obtained from ANOVA (the interested reader is encouraged to verify the value of 13.83, either by hand or more likely by the regression procedure in a statistical package). The corresponding *F* value in the regression analysis is 8.30, based on the correct MS_W value of 1.67. In addition, the estimated coefficient for X_1 is no longer −1.00, but instead has become −1.57, even though all four sample means are the same in the augmented data as they were in the original data.

Five points must be stressed here. First, the regression approach has produced an incorrect estimate of the mean difference, an incorrect sum of squares for the contrast, and an incorrect F value, and thus an incorrect test of the desired contrast. The problem is that the simple method of coding that works fine when sample sizes are equal does not test the contrast it might appear intuitively to test when sample sizes are unequal. Instead, the X_1 variable can be shown to be testing an entirely different hypothesis. In particular, by simplifying the general expression for a slope in simple regression, it can be shown that the coefficient for X_1 will be given by

$$\hat{\beta}_1 = \frac{n_1(\bar{Y}_1 - \bar{Y}) - n_2(\bar{Y}_2 - \bar{Y})}{n_1(1-((n_1 - n_2)/N))^2 + n_2(1+((n_1 - n_2)/N))^2 + (n_3 + n_4)((n_1 - n_2)/N)^2} .$$

Substituting $n_1 = 4$, $n_2 = n_3 = n_4 = 2$, $\bar{Y}_1 = 12$, and $\bar{Y}_2 = 14$ into this expression for $\hat{\beta}_1$ yields a value of −1.57, just as we obtained in our regression analysis. In this sense, regressing Y on X_1 by itself produces the right answer to the wrong question. Although it may not be obvious, with equal sample sizes, the expression for $\hat{\beta}_1$ given earlier simplifies greatly to become equal to $(\bar{Y}_1 - \bar{Y}_2)/2$, which then yields the desired result. So this simple approach works fine with equal n, but it does not truly test the difference between Groups 1 and 2 with unequal n. Second, it would be a mistake to infer that there is some inherent problem here with regression. Instead, the problem arises from a mistaken belief that the simple method of coding will test the hypothesis of interest. As we have already pointed out, there is not necessarily a direct relationship between the values of a coded variable and the coefficients of a contrast being tested by that variable. Third, regression can be used to test the correct hypothesis, but this necessitates a fundamentally different approach from simply regressing Y on X_1. The most important point is that in general it is necessary to regress Y on an entire set of $a - 1$ predictors (plus the unit variable) in order to obtain the correct test of the desired contrast, even though this may be the only contrast of interest. In fact, for pairwise comparisons, we can always rely on the reference cell model, where we simply define the reference cell to be one of the two groups to be compared. Or more generally, we can use the method presented earlier in the tutorial for complex contrasts. Yet other possible options are described in such sources as Cohen et al. (2002), Darlington and Hayes (2017), Pedhazur (1997), and Serlin and Levin (1985).

Contrasts via Linear Combinations of Regression Coefficients

We have just seen that creating appropriate X variables to test contrasts of interest through regression is often tedious at best and can produce erroneous results at worst. In fact, in our judgment, this is one of the main limitations of adopting a regression approach for teaching and learning analysis of variance. However, there is yet another approach that avoids these problems. Instead of attempting to identify appropriate coding methods, this alternative approach relies on forming linear combinations of the parameters of the model. Although this approach may be slightly less intuitive than directly coding contrast variables, it has the advantage of being much more general in that it produces correct values even when sample sizes are unequal and/or contrasts are nonorthogonal. Because of its generality, this approach has been implemented in the general linear model procedure of many popular statistical packages, such as SAS and SPSS.

Most packages that include the capability of forming linear combinations of the parameters are based on the reference cell model, so that is the approach we will illustrate here. The key to understanding how this approach works is to recall the relationship between population means and regression parameters. Recall that this relationship can be written as

$$\beta_0 = \mu_a \text{ for Group } a,$$
$$\beta_j = \mu_j - \mu_a \text{ for each Group } j.$$

Notice that the regression parameter β_a for the final group is necessarily zero, so it need not (indeed, cannot) be estimated from the data. Now consider how this approach goes about testing a pairwise contrast of the difference between the first two means. What needs to be done here is to express μ_1 and μ_2 in terms of the betas—i.e., the regression parameters. Notice that we can accomplish this goal by substituting β_0 for μ_a in the bottom equation and then placing μ_j on the left side of the equation, yielding

$$\mu_j = \beta_0 + \beta_j \text{ for each group } j.$$

We can now rewrite our contrast μ_1 minus μ_2 in terms of the betas as

$$\mu_1 - \mu_2 = (\beta_0 + \beta_1) - (\beta_0 + \beta_2).$$

The β_0 term obviously drops out of this expression, which can then be written more simply as

$$\mu_1 - \mu_2 = \beta_1 - \beta_2.$$

What have we accomplished with all of this? We have learned that we can test our pairwise contrast by using the regression model to test whether β_1 and β_2 are equal to one another. We can also form a confidence interval for the mean difference $\mu_1 - \mu_2$ by forming a confidence interval for $\beta_1 - \beta_2$. The same logic applies to any other contrast, pairwise or complex.

This approach has a very important advantage over the direct coding method—namely, that it is completely general. It provides correct answers with equal or unequal sample sizes, pairwise or complex contrasts, and with orthogonal or nonorthogonal sets of contrasts. In theory, the only disadvantage is that it necessitates finding the relationship between the ANOVA parameters μ_j and the regression parameters β_j. In practice, however, this work is often done by the statistical package behind the scenes. For example, the general linear model procedures in SAS and SPSS allow you to specify your hypothesis of interest in terms of ANOVA parameters. The computer program then translates the hypothesis into the regression formulation and performs the appropriate analysis.

The only serious disadvantage of this approach is that it requires specifying the standard error of the linear combination of regression parameters. The bad news is that the general form of this specification involves matrix algebra. The good news is that SAS, SPSS, and other packages perform these calculations for you, so computational burden need not be an issue.

Numerical Example

We have already used a numerical example to illustrate complications that can emerge in attempting to use direct coding of predictor variables to test contrasts. Thus, in this section, we will briefly illustrate how testing linear combinations of parameters in the reference cell model circumvents these difficulties.

We will return to the augmented data set we introduced earlier in this tutorial. Table T3.6 shows an excerpt of the output from PROC GLM in SAS. The top portion of Table T3.6 shows parameter estimates and corresponding information from the reference cell model (as of this writing, this output is obtained from PROC GLM by specifying SOLUTION as a desired option). Several points are especially noteworthy. First, the intercept β_0 is estimated to be 18, because as we discussed, the intercept in the reference cell model represents the mean of the final group, and the sample mean for the fourth group here is in fact 18 (i.e., $\bar{Y}_4 = 18$). Second, notice that β_1 is estimated to be −6, which is simply the difference between the sample mean of the first group ($\bar{Y}_1 = 12$) and the sample mean of the fourth group. Third, the estimates for β_2 and β_3 follow the same logic. Fourth, the estimate for β_4 is zero, because that is the constraint imposed by the reference cell model. Fifth, each estimate is accompanied by its standard error, *t*-statistic, *p* value, and confidence interval. Sixth, the program includes a warning that serves as a reminder that the estimates obtained here reflect a specific choice of constraints we have placed on the parameters.

The next portion of Table T3.6 shows the result of including a CONTRAST statement in the command syntax. Specifically, this output results from requesting a contrast of the first two groups. As you may recall, we previously calculated the sum of squares for this contrast using Equation 4.11 and obtained a value of 5.33, in agreement with the value shown in the table. We also found that the observed *F* value for this contrast is 3.20, once again agreeing with the value shown in the table. While this may not seem like a big deal, keep in mind that obtaining these values with direct coding necessitated tedious creation of $a - 1$ coded predictor variables. As we have stated, however, forming linear combinations of regression coefficients allows SAS PROC GLM and similar programs to duplicate the results we presented in Chapter 4 and elsewhere in the book.

The final portion of Table T3.6 shows the result of including an ESTIMATE statement in the command syntax, once again for comparing the first two groups. The information provided by ESTIMATE largely duplicates information obtained from CONTRAST. However, ESTIMATE offers an advantage in that it expresses results in terms of a confidence interval unlike CONTRAST.

Two-Way Between-Subjects Factorial Designs

As in single-factor designs, models for factorial designs can be expressed in terms of either ANOVA or regression, and in terms of effects, cell means, or a reference cell. While the logic we have developed for the single-factor design extends to factorial designs, the issues become more complicated. Thorough consideration of these complications is beyond the scope of this book.

Instead of attempting to present a variety of models for the two-way design and discussing their interrelationships, we have chosen to focus on a demonstration of the reference cell model in this design. We have chosen this model because as we have already mentioned, it often underlies linear models procedures in current statistical software packages. Our presentation here is intentionally less detailed in the single-factor design and is intended simply to give readers a general sense for how concepts we developed in the single-factor design can be extended to the factorial design.

We will focus our attention on how the reference cell model can be written as a regression model in order to analyze the data shown in Table 7.9 of the body of the text. As a reminder, the data in this table represented scores for 36 individuals who had received various combinations of biofeedback and drug therapy for treatment of hypertension. The resultant design was 2 × 3, with two levels of biofeedback (present or absent) and three types of drugs. For convenience, the cell means and marginal means for these data are reproduced here as Table T3.7.

The regression approach based on a reference cell model requires appropriate coding of predictor variables. While any specific cell could serve as the reference cell, we will illustrate the

TABLE T3.6
OUTPUT FROM SAS PROC GLM FOR AUGMENTED DATA SET

The GLM Procedure

Dependent Variable: Y

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*95% Confidence*	*Interval*
Intercept	18.00000000 B	0.91287093	19.72	<.0001	15.76628530	20.23371470
Group 1	−6.00000000 B	1.11803399	−5.37	.0017	−8.73573062	−3.26426938
Group 2	−4.00000000 B	1.29099445	−3.10	.0212	−7.15894962	−0.84105038
Group 3	−2.00000000 B	1.29099445	−1.55	.1723	−5.15894962	1.15894962
Group 4	0.00000000 B					

Note. The X’X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter ‘B’ are not uniquely estimable.

Contrast	*DF*	*Contrast SS*	*Mean Square*	*F Value*	*Pr> F*
mul vs. mu2	1	5.33333333	5.33333333	3.20	0.1238

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr> |t|*	*95% Confidence*	*Interval*
mul vs. mu2	−2.00000000	1.11803399	−1.79	0.1238	−4.73573062	0.73573062

TABLE T3.7
CELL MEANS AND MARGINAL MEANS FOR 2 × 3 DESIGN

			B(Drug)		
		1(X)	*2(Y)*	*3(Z)*	*Marginal Means*
A(Biofeedback)	1 (Present)	168	204	189	187
	2 (Absent)	188	200	209	199
	Marginal Means	178	202	199	193

typical default of allowing the final cell in the design to serve as the reference cell. In this 2 × 3 design, this “final” cell will be the cell in row 2 and column 3. Thus, we will need predictors to distinguish other rows and columns from row 2 and column 3. In theory, there are many different ways we could parameterize such a model, and if we are clever enough we could figure out how to form linear combinations of these parameters to test whatever types of effects we decide to test, but we will proceed to illustrate how predictors are typically coded so as to set the stage to test effects of interest in factorial designs, such as main effects, interactions, and simple effects.

At the outset we need to remember that the reference cell model includes an intercept term equal to one for every individual. As usual we will represent this predictor variable as X_0, which reminds us that this term is already included in the regression model by default, so we need not actively create this variable ourselves. Before attempting to define additional predictor variables, it may be helpful to anticipate how many predictors we will need to include in our full model. The full model should include a parameter for each population mean, so in a 2 × 3 design, we should have a total of six parameters. The intercept counts as one of these parameters, so in our case, we should define five additional X variables. The general form of the regression model here can thus be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i,$$

where X_1 through X_5 are coded so as to represent differences among the groups. With two rows and three columns, one of these variables will typically reflect a type of row difference, two will reflect a type of column difference, and the remaining two will reflect the interaction.

We can now proceed to define a predictor that will represent a type of row effect (we will see shortly exactly what type of row effect this variable represents). Specifically, we can define X_1 as follows:

$$X_{1i} = \begin{cases} 1 & \text{if the individual belongs to row 1 (i.e., biofeedback present)} \\ 0 & \text{otherwise} \end{cases}.$$

Similar logic applies for the columns, except that with three columns, we need to create two predictors to represent column differences. A typical choice would be

$$X_{2i} = \begin{cases} 1 & \text{if the individual belongs to column 1 (i.e., drug X)} \\ 0 & \text{otherwise} \end{cases}$$

and

$$X_{3i} = \begin{cases} 1 & \text{if the individual belongs to column 2 (i.e., drug Y)} \\ 0 & \text{otherwise} \end{cases}.$$

The final two predictors then represent the interaction. Each interaction predictor is the product of a row predictor with a column predictor. Because there is one row predictor and there are two column predictors, there will be two interaction predictors. Notice that in general there will be $a - 1$ row variables and $b - 1$ predictor variables. Thus $(a - 1)(b - 1)$ interaction variables will result from multiplying each row variable by each column variable, just as we would expect, because $(a - 1)(b - 1)$ is the numerator degrees of freedom for the interaction effect in the two-way design. In our specific 2 × 3 design, the predictors are defined as follows:

$$X_{4i} = \begin{cases} 1 & \text{if the individual belongs to row 1 and column 1 (i.e., the} \\ & \text{combination of biofeedback present and drug X)} \\ 0 & \text{otherwise} \end{cases},$$

and

$$X_{5i} = \begin{cases} 1 & \text{if the individual belongs to row 1 and column 2 (i.e., the} \\ & \text{combination of biofeedback present and drug Y)} \\ 0 & \text{otherwise} \end{cases}.$$

Table T3.8 shows the result of regressing Y on these five X variables for these 36 individuals, as obtained from PROC REG in SAS. In particular, the table shows each parameter estimate, as well as corresponding standard errors, t values, and p values. As we will see in more detail shortly, we have to be very careful not to misinterpret these parameter estimates. For example, it would be tempting to assume that X_1 represents the row main effect, in which case we might infer that the p value for the row main effect in these data is .0059. In reality, however, X_1 does not represent the main effect of the row factor, but instead the simple effect of row within the third column. We can understand this by returning to our coding scheme. Notice that individuals in the row 2, column 3 cell (i.e., the combination of biofeedback absent and drug Z) have been assigned a score of 0 on all five X variables. Thus the regression model for scores in this specific cell simplifies to

$$Y_i = \beta_0 + \varepsilon_i.$$

TABLE T3.8
REGRESSION ANALYSIS OF DATA FROM 2 × 3 FACTORIAL DESIGN

Parameter Estimates					
Variable	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr>|t|*
Intercept	1	209.00000	4.77144	43.80	<.0001
X1	1	−20.00000	6.74784	−2.96	0.0059
X2	1	−21.00000	6.74784	−3.11	0.0041
X3	1	−9.00000	6.74784	−1.33	0.1923
X4	1	0	9.54289	0.00	1.0000
X5	1	24.00000	9.54289	2.51	0.0175

Similarly, individuals in row 1, column 3 (i.e., the combination of biofeedback present and drug Z) have been assigned a score of 0 on all variables except for X_1, where they have received a score of 1. Thus the regression model for scores in this specific cell simplifies to

$$Y_i = \beta_0 + \beta_1 + \varepsilon_i.$$

Comparing these two equations shows us that β_1 represents the difference between scores in these two cells. We can now see that the difference between these two cells is simply the simple effect of row in the third column. Indeed, the parameter estimate of −20 shown in Table T3.8 equals the difference between the sample means of these cells shown in Table T3.7. Namely, the value of −20 equals 189 minus 209, as it must. Thus, although we coded X_1 so that all individuals in row 1 received a score of 1 and all individuals in row 2 received a score of 0, β_1 does not represent the row main effect. Instead, β_1 represents the simple effect of row in the third column. One way of understanding this is to realize that the five X variables we have created are correlated, even with equal n. Thus, β_1 reflects the relationship between Y and X_1 when the remaining predictor variables are "held constant." What it means to hold the remaining predictors constant here turns out to correspond to the simple effects test. The other regression coefficients also must be interpreted in this light. Although this is no problem whatsoever mathematically, it does show once again that intuitive interpretations of regression parameters may be incorrect.

We can immediately see that the parameters in the regression model do not necessarily have the meaning we might naively assume them to have. As a consequence, interpreting the output shown in Table T3.8 is not as straightforward as one might hope. While we could proceed to explain the proper interpretation of each parameter estimate, we will instead take a different approach. Although we could use PROC REG or a similar regression procedure to calculate estimates such as those shown in Table T3.8 and then test hypotheses of interest by forming appropriate linear combinations, the availability of GLM procedures in SAS, SPSS, and other packages provides a simpler alternative. At this moment, we can well imagine that you may be asking why we have bothered to introduce the confusion surrounding estimates from regression if we are really going to use GLM all along. However, keep in mind that the purpose of this section is to explain the relationship between regression and ANOVA models. Even though the practical import of our discussion may be to convince you always to use GLM and to avoid the complications of interpreting REG, nevertheless it is useful to understand how they are related to one another. Of course, we do not mean to imply that GLM avoids all potential problems of misinterpretation among uninformed researchers. For example, GLM does not automatically center continuous variables, so when continuous and categorical variables are allowed to interact with one another, what appears to be a main effect may in actuality be a simple effect of the categorical variable conditional on a value of 0 for the continuous variable (see West, Aiken, & Krull, 1996, for more on this topic).

Our plan at this point is to reconsider this factorial design from the perspective of GLM. We will illustrate a few tests one might perform using GLM and then consider how to duplicate these tests using REG. We want to stress that from a purely practical perspective, we could simply use GLM and stop. There is no need to use REG here at all. However, because a major purpose of this tutorial is to show the relationship between ANOVA and regression, we will perform a few selected tests both ways.

In particular, we will use both GLM and REG to perform three tests: (1) the row main effect, (2) the simple row effect within the first column, and (3) the interaction contrast within the first two columns. We have chosen these three tests simply because they provide a variety of types of effects one might wish to test. We do not mean to imply in any fashion that they are necessarily the tests that we would really perform on these data; a description of how we might choose a meaningful set of tests was presented in Chapter 7 when we originally presented these data. Also we should be clear that although all of the effects we have chosen to test here are single degree of freedom effects, the same logic applies to multiple degree of freedom tests.

We will begin by considering how to test these three effects in GLM. We can use ESTIMATE and/or CONTRAST to perform these tests. We have chosen to illustrate ESTIMATE because it also provides an estimate of the contrast, which we can then compare to the table of cell means shown in Table T3.7. To provide the proper commands to GLM, we must return to the ANOVA formulation of the model. In the case of a two-way between-subjects design, we can write the model as

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}. \tag{7.6, repeated}$$

To use ESTIMATE or CONTRAST, we need to write the corresponding model for each specific cell mean of the design. In the case of a 2 × 3 design, this yields six equations, one for each cell mean:

$$\begin{aligned}
\mu_{11} &= \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11},\\
\mu_{12} &= \mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12},\\
\mu_{13} &= \mu + \alpha_1 + \beta_3 + (\alpha\beta)_{13},\\
\mu_{21} &= \mu + \alpha_2 + \beta_1 + (\alpha\beta)_{21},\\
\mu_{22} &= \mu + \alpha_2 + \beta_2 + (\alpha\beta)_{22},\\
\mu_{23} &= \mu + \alpha_2 + \beta_3 + (\alpha\beta)_{23}.
\end{aligned}$$

The next step is to express the effect of interest in terms of the population cell means. For example, the row main effect compares the average of the three cells in the first row to the average of the three cells in the second row:

$$\tfrac{1}{3}(\mu_{11} + \mu_{12} + \mu_{13}) - \tfrac{1}{3}(\mu_{21} + \mu_{22} + \mu_{23}) .$$

We now must re-express this linear combination in terms of the ANOVA effect parameters. Substituting from the six previous equations and simplifying yields

$$\alpha_1 - \alpha_2 + \tfrac{1}{3}(\alpha\beta)_{11} + \tfrac{1}{3}(\alpha\beta)_{12} + \tfrac{1}{3}(\alpha\beta)_{13} - \tfrac{1}{3}(\alpha\beta)_{21} - \tfrac{1}{3}(\alpha\beta)_{22} - \tfrac{1}{3}(\alpha\beta)_{23} .$$

The final step of the process is to write the appropriate syntax to tell the computer program to estimate this linear combination. As of this writing, the appropriate syntax to estimate this effect in PROC GLM of SAS would have the following general form, where instead of writing coefficients as fractions, integer values are given along with the “divisor” that is to be used as the denominator of all coefficients:

```
estimate ‘rowl vs. row2’ row 3 − 3 row*column 1 1 1 −1 −1 −1 / divisor = 3.
```

In certain cases such as this one, users are allowed to omit the interaction effects, because SAS will automatically include them when the user has not done so him or herself; that is, the following syntax would here produce the same result:

```
estimate ‘rowl vs. row2’ row 1–1.
```

Instead of belaboring the point by deriving similar expressions for the other two effects of interest, the top portion of Table T3.9 shows the end results of writing these questions in terms of ANOVA model parameters. The middle of Table T3.9 shows the equivalent SAS syntax. The bottom of Table T3.9 shows output from PROC GLM for estimating these three contrasts in our data.

Two goals yet remain. First, we will briefly verify that the results shown in Table T3.9 are correct. Second, we still need to return to our regression model so we can see how to duplicate these results using regression.

We can easily verify that the results shown in Table T3.9 are correct. First, consider the row main effect. Notice that the estimated value of the contrast is −12. Returning to Table T3.7 shows that the difference between the row marginal means is indeed −12, because the marginal mean for the first row is 187 and the marginal mean for the second row is 199. In addition, squaring the *t* value of −3.08 shown in Table T3.9 yields a value of 9.49, which is the *F* value reported for the row main effect in Chapter 7. Second, consider the simple row effect within the first column. The estimated value of the contrast, −20, equals the difference between 168 and 188 from Table T3.7. Once again, squaring the *t* value of −2.96 shown in Table T3.9 yields a value of 8.76, which is the same (except for rounding error) as the *F* value reported for the simple effect in Chapter 7. Finally, consider the interaction contrast within the first two columns. The estimated value of the contrast, −24, equals (168 + 200) minus (188 + 204) from Table T3.7. As it must, squaring the *t* value of −2.51 shown in Table T3.9 yields a value of 6.30, which once again agrees (within rounding error) with the *F* value reported for this interaction contrast in Chapter 7 (see page 346).

Our final challenge is to see how these tests we have just performed using PROC GLM can be duplicated in PROC REG. How did PROC GLM test our contrasts, and how does this relate to PROC REG? Unbeknownst to the user, PROC GLM has created the reference cell model we used to perform our regression analysis. Specifying the SOLUTION option in GLM produces the same estimates, standard errors, *t* values, and ρ values shown in Table T3.8 for the analysis we conducted using PROC REG. However, PROC GLM has an extremely useful advantage over

TABLE T3.9
ESTIMATES USING PROC GLM

Effects in Terms of Model Parameters

(1) Row main effect

$$\alpha_1 - \alpha_2 + 1/3(\alpha\beta)_1 + 1/3(\alpha\beta)_{12} + 1/3(\alpha\beta)_{13} - 1/3(\alpha\beta)_{21} - 1/3(\alpha\beta)_{22} - 1/3(\alpha\beta)_{23}$$

(2) Simple effect of row within first column

$$\alpha_1 - \alpha_2 + (\alpha\beta)_{11} - (\alpha\beta)_{21}$$

(3) Interaction contrast within first two columns

$$(\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22}$$

Corresponding SAS PROC GLM Syntax

(1) Row main effect

estimate ‘rowl vs. row2’ row 3 −3 row*column 1 1 1 −1 −1 −1/ divisor=3

(2) Simple effect of row within first column

estimate ‘mu11 vs. mu21’ row 1 –1 row*column 1 0 0 –1 0 0

(3) Interaction contrast within first two columns

estimate interaction contrast’ row*columns 1 –1 0 –1 1 0

SAS Output for Data in Table T3.7

		Standard			*95%*	
Parameter	*Estimate*	*Error*	*t Value*	*Pr>|t|*	*Confidence*	*Interval*
rowl vs. row2	−12.0000000	3.89586676	−3.08	0.0044	−19.9564214	−4.0435786
mull vs. mu21	−20.0000000	6.74783916	−2.96	0.0059	−33.7809261	−6.2190739
interaction contrast	−24.0000000	9.54288566	−2.51	0.0175	−43.4891725	−4.5108275

PROC REG from the user’s perspective. PROC GLM, unlike PROC REG, allows the user to specify questions of interest in terms of the ANOVA model parameters. However, PROC REG is unaware of any concept of ANOVA model parameters, and requires the user to specify all questions of interest in terms of X_1 through X_5, the predictors we created as indicator codes. In reality, PROC GLM has created these same five predictors, but it does much of the work for the user by translating between the ANOVA effects and the regression parameters.

To understand how PROC GLM does this translation, we will once again consider the same three contrasts we estimated using GLM. However, now we will see how we can estimate and test these contrasts directly using PROC REG. The key here is to realize the relationship between the regression parameters and the ANOVA parameters. At the outset, it is crucial to consider the constraints we have placed on parameters by adopting the reference cell approach. For example, consider α_1 and α_2, the effect parameters for the first and second row, respectively. It seems most natural to think of these parameters in terms of the ANOVA effects model, in which case we impose the constraint that α_1 and α_2 must sum to zero. However, we must remember that the constraint in the reference cell model is different. Recall that in our discussion of reference cell coding in the one-way model, we saw that the effect for the final group is set equal to zero. Similarly, in the two-way model, instead of constraining α_1 and α_2 to sum to zero, the reference cell model sets α_2 equal to zero. More generally, instead of imposing any constraints about sums, the reference cell model simply sets an appropriate number of parameters equal to zero.

The top portion of Table T3.10 shows the resulting correspondence between ANOVA parameters and regression parameters using reference cell coding for our 2 × 3 design. We must remember that the meaning of the specific ANOVA parameters is contingent on the constraints we have imposed. For example, the meaning (and therefore the value) of α_1 in the reference cell model will generally be different from the meaning and the value of α_1 in the effects model, even though the same symbol appears in both models. Before abandoning all hope, however, there are certain types of effects whose meaning and value do not depend on our choice of constraint. These effects are called estimable functions and include such effects as main effects, interactions, marginal mean comparisons, simple effects, interaction contrasts, and cell mean comparisons. All three contrasts we estimated using GLM are examples of estimable functions. Further discussion of estimable functions is beyond the scope of our presentation (for further details, see Green, Marquis, Hershberger, Thompson, and McCollam (1999) or Littell, Freund, & Spector, 1991), so we will simply see how we can use the correspondence shown in the top portion of Table T3.10 to estimate and test the three contrasts in which we are interested.

To use our regression model to estimate and test contrasts of interest, we must write each contrast in terms of the parameters of the model (i.e., in terms of β_1 through β_5). This proves to be straightforward once we have written the contrast in terms of ANOVA model parameters, as in Table T3.9. For example, consider the row main effect. We saw that this effect can be written in terms of ANOVA model parameters as

$$\alpha_1 - \alpha_2 + \tfrac{1}{3}(\alpha\beta)_{11} + \tfrac{1}{3}(\alpha\beta)_{12} + \tfrac{1}{3}(\alpha\beta)_{13} - \tfrac{1}{3}(\alpha\beta)_{21} - \tfrac{1}{3}(\alpha\beta)_{22} - \tfrac{1}{3}(\alpha\beta)_{23}.$$

TABLE T3.10
REGRESSION MODEL PARAMETERS AND CORRESPONDING REGRESSION ANALYSIS OF TABLE T3.4 DATA

Relationship of ANOVA and Regression Parameters

ANOVA	*Regression*
μ	β_0
α_1	β_1
α_2	0
β_1	β_2
β_2	β_3
β_3	0
$(\alpha\beta)_{11}$	β_4
$(\alpha\beta)_{12}$	β_5
$(\alpha\beta)_{13}$	0
$(\alpha\beta)_{21}$	0
$(\alpha\beta)_{22}$	0
$(\alpha\beta)_{23}$	0

Re-expression of contrasts in terms of regression parameters

(1) Row main effect
ANOVA: $\alpha_1 - \alpha_2 + 1/3(\alpha\beta)_{11} + 1/3(\alpha\beta)_{12} + 1/3(\alpha\beta)_{13} - 1/3(\alpha\beta)_{21} - 1/3(\alpha\beta)_{22} - 1/3(\alpha\beta)_{23}$
Regression : $\beta_1 + 1/3\beta_4 + 1/3\beta_5$
(2) Simple effect of row within first column
ANOVA : $\alpha_1 - \alpha_2 + (\alpha\beta)_{11} - (\alpha\beta)_{21}$
Regression : $\beta_1 + \beta_4$
(3) Interaction contrast within first two columns
ANOVA : $(\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22}$
Regression : $\beta_4 - \beta_5$

Output from PROC REG

(1) Row main effect					
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	1	1,296.00000	9. 49	0.0044
	Denominator	30	136.60000		
(2) Simple effect of row within first column					
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	1	1,200.00000	8.78	0.0059
	Denominator	30	136.60000		
(3) Interaction contrast within first two columns					
	Source	DF	Mean Square	F Value	Pr > F
	Numerator	1	864.00000	6.33	0.0175
	Denominator	30	136.60000		

All that remains now is to rewrite this linear combination in terms of the regression model parameters. The top portion of Table T3.10 allows us to make a simple substitution, yielding

$$\beta_1 - 0 + \tfrac{1}{3}\beta_4 + \tfrac{1}{3}\beta_5 + 0 - 0 - 0 - 0\ ,$$

which can obviously be written more simply as

$$\beta_1 + \tfrac{1}{3}\beta_4 + \tfrac{1}{3}\beta_5\,.$$

What this tells us is that we can use our regression model to estimate and test the row main effect by estimating and testing this admittedly rather strange looking linear combination of regression parameters. The middle of Table T3.10 shows the corresponding expressions for the row simple effect in the first column and for the interaction contrast in the first two columns. The bottom of Table T3.10 shows the result of testing each of these three linear combinations of regression parameters using the TEST option of PROC REG in SAS. That the regression results are in fact equivalent to the results shown earlier in Table T3.9 using GLM can be seen most easily by comparing the three *p*-values of the two approaches. Of course, the squared *t* values shown in Table T3.9 are also equal (within rounding error) to the *F* values shown in Table T3.10.

The equivalence of results in Tables T3.9 and T3.10 illustrates how it is possible to use regression models with reference cell coding to duplicate results obtained from ANOVA models in two-way factorial designs. Similar logic extends this relationship to more complex designs.

THE RELATION OF ANOVA AND REGRESSION TO OTHER STATISTICAL MODELS

The fact that regression can (1) literally duplicate the ANOVA effects model, (2) literally duplicate the ANOVA cell means model, and (3) also include continuous predictor variables makes it an extremely flexible and valuable statistical method. Indeed, if these ANOVA models are special cases of the regression model, it seems reasonable to ask why the majority of this book formulates models in terms of ANOVA parameters instead of regression parameters. We will answer this question in two ways: (1) by considering where ANOVA and regression fit into a broader structure of statistical methods and (2) by introducing some examples where we believe that the ANOVA formulation has important advantages.

Researchers who understand the relationship between ANOVA and regression can benefit by formulating models that best address their scientific questions instead of forcing their questions to fit a model that may not really be appropriate. Thus we wholeheartedly endorse books that emphasize the relationship between ANOVA and regression throughout their presentation. Nevertheless, we have chosen not to follow this route and believe that the reader deserves some explanation for our choice to emphasize ANOVA models.

From one perspective, our choice can be understood in terms of levels of generality. For example, we believe that a major strength of the model comparison approach we employ throughout the book is that the concept of comparing a full model to a restricted model plays a central role in many other types of statistical models. However, a case could be made that our approach is narrow because it focuses so heavily on ANOVA models. Why not present data analysis from a broader perspective, such as regression analysis?

Although there are some advantages associated with presenting material in its most general form, there are frequently associated disadvantages. For example, one issue that must be confronted is how general one should be. Although the regression model provides considerable flexibility, it is hardly the most general statistical model one might contemplate. Instead, the regression model can itself be regarded as a special case of a number of other models.

Figure T3.1 provides a graphical depiction of the relationships among some common statistical models. Models nearer the top of the figure are more general than models nearer the bottom. Specifically, when two models are connected with a line, the model at the bottom is a special case of the model at the top. Thus, for example, ANOVA is depicted as a special case of MRA, multiple regression analysis. Multiple regression models are more general than ANOVA models in that they can include not only categorical (i.e., nominal, or “class”) predictor variables but also continuous predictors. Similarly, ANOVA is a special case of MANOVA, because ANOVA models allow only a single dependent variable, while MANOVA allows one or more dependent variables. Chapters 13 and 14 illustrate one use of MANOVA as a method for analyzing data from repeated measures designs. Notice that multiple regression analysis and MANOVA are shown on the same level of the figure, and neither is directly connected to the other. These two models are shown this way because neither is a special case of the other. MANOVA is not a special case of regression because regression (with the exception of multivariate multiple regression) allows only one dependent variable. However, regression is also not a special case of MANOVA because MANOVA allows only categorical predictors.

Both regression and MANOVA can be thought of as special cases of the general linear model (GLM), which is sometimes called the multivariate general linear model. This model extends MANOVA by allowing continuous as well as categorical predictors. In this sense it is like regression. However, the GLM is more general than regression in that it allows multiple dependent variables. Yet one other distinction is sometimes made between GLM and regression. For example, Darlington and Hayes (2017) describe the ability to create a set of indicator variables automatically with a single command as a fundamental difference between GLM computer programs and regression computer programs.

The GLM can itself be extended in any of several ways. For example, the generalized linear model (GLIM) extends the GLM by allowing the dependent variable to take on a distributional form other than normality. This flexibility makes GLIM especially useful for modeling data such as proportions and counts. The general linear mixed model (GLMM) provides yet another example of an extension to the general linear model. The GLMM, which we refer to as the mixed-effects model in Chapters 15 and 16, expands the general linear model by allowing predictors to be random as well as fixed. Chapters 15 and 16 provide an introduction to some examples of this type of model. The final model shown in the figure is the structural equation model (SEM), also known as the LISREL model, which is the name of the computer package that helped popularize these methods. SEMs expand on the GLM in two ways. First, SEMs allow relationships to be examined for latent as well as manifest variables. Latent variables, as the name implies, are variables that are not directly observed; instead, manifest variables are included in the model as imperfect manifestations of presumed underlying latent variables. Even without directly measuring latent variables, it is possible to study their relationships under certain conditions and assumptions. In this respect, latent variables are much like factors in traditional factor analysis. Second, SEMs also allow a variable to serve as both an independent variable and a dependent variable. For example, a child's eagerness to learn might be dependent (at least in part) on certain parental characteristics, while at the same time the child's eagerness to learn has some influence on the child's academic progress. This set of relationships could not be directly examined in the GLM, because eagerness would need to be either an X variable or a Y variable in the model. However, structural equation modeling allows eagerness to serve as an X variable in one equation and as a Y variable in another equation.

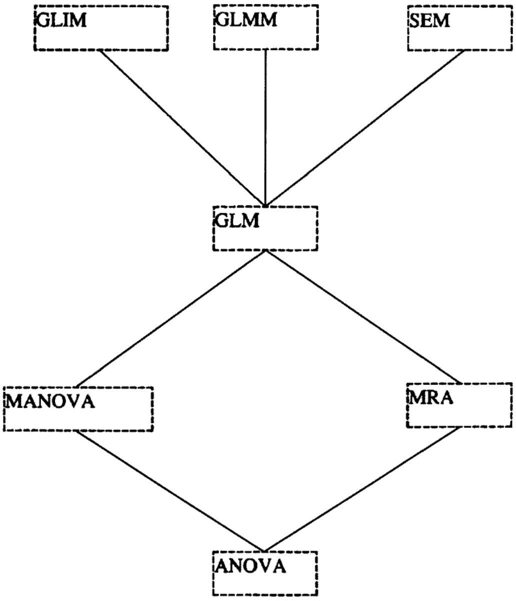

FIGURE T3.1 The relation of ANOVA and regression to other statistical models

ANOVA "VERSUS" REGRESSION MODELS

Some individuals have suggested that ANOVA should always be conceptualized in terms of regression because ANOVA is simply a special case of regression. However, if the fact that ANOVA is a special case of regression is reason enough by itself to suggest that ANOVA should always be thought of in terms of regression, it would seem to follow logically that we should not stop at the level of regression. As Figure T3.1 shows, regression is itself simply a special case of yet other methods. Thus this logic would seem to imply that both ANOVA and regression should always be thought of in terms of the GLM, or perhaps in yet higher terms such as structural equation modeling.

Our goal in this book is to present concepts in such a way that they are generally applicable. For example, our basic philosophy of comparing models is applicable throughout the hierarchy shown in Figure T3.1 Even though specific test statistics may take on different forms, the ideas remain much the same. Nevertheless, we have chosen to present these ideas primarily in the context of ANOVA models. We made this choice for several related reasons. First, the analysis of data from certain types of designs is much more natural in the ANOVA framework. For example, Chapters 10, 11, and 12 (along with Chapters 15 and 16) show that designs with random effects usually require consideration of multiple error terms, which does not fit easily in the regression framework. Similarly, the multivariate approach to repeated measures presented in Chapters 13 and 14 cannot be accommodated in (univariate) multiple regression. Even in purely between-subjects designs with only fixed effects, we would suggest that more attention needs to be paid to assumptions underlying pooled error terms, especially when forming confidence intervals or testing hypotheses for contrasts. Consideration of separate error terms is much clearer from the perspective of ANOVA instead of regression. Second, from a purely practical perspective it is much easier to test differences in group means using GLM procedures in statistical packages than using regression procedures. As Darlington and Hayes (2017) point out, a distinguishing characteristic of GLM procedures is that they allow the user to specify questions of interest in terms of ANOVA effect parameters. Thus researchers will almost certainly rely on an ANOVA conceptualization in using statistical packages to analyze their data. While in theory regression could be used to obtain the same results, we have seen that exclusive reliance on regression can become confusing and tedious even in rather simple designs. Third, ANOVA models provide the most natural and convenient way of thinking about group differences. If your questions are about means, it seems sensible that the parameters of your model should themselves be means, or perhaps differences in means. While parameters of other models such as regression models may also be means or mean differences, the precise meaning of such parameters is often much less clear than in ANOVA models. Fourth, in a similar vein, ANOVA models directly reflect the fact that an investigator's questions pertain to multiple populations. Regression models, on the other hand, do not necessarily reflect multiple populations. Instead, they reflect the relationship between a dependent variable and one or more independent variables in a single population. Of course, we can code indicator variables so as to represent different populations, but even here the method itself addresses the extent to which these indicator variables are related to the dependent variable. Nowhere in the model itself is there a direct expression that multiple populations exist. Fifth, although we believe that ANOVA formulations offer significant advantages, nevertheless it is crucial that researchers not force all of their questions into an ANOVA framework. For example, as numerous authors including ourselves (e.g., Cohen, 1983; Maxwell & Delaney, 1993; Vargha, Rudas, Delaney, & Maxwell, 1996) have shown, artificially categorizing a continuous variable can badly distort the true nature of relationships between variables. Contrary to conventional wisdom, results based on artificial dichotomies are not necessarily conservative. Instead, categorizing continuous variables can sometimes create spurious findings and inflate Type I error rates dramatically, so researchers cannot rely on an excuse that they simply opted for a more conservative way of analyzing their data when they categorized a continuous variable. Thus with rare exceptions continuous variables should be modeled in their continuous forms, necessitating an extension beyond ANOVA to ANCOVA or the GLM. However, even here we would suggest that any categorical predictors are best understood in terms of ANOVA parameters, not regression coefficients. The end result is that researchers need to be flexible and able to formulate models that truly reflect their hypotheses and data structures, instead of always attempting to force every analysis into the same model or computerized procedure. Sixth, in summary we agree with the view espoused by Mallows and Tukey (1982) that ANOVA should not be regarded as "simply" a special case of regression. They stated, "It is sometimes said that 'analysis of variance is just a form of regression.' The common occurrence of ANOVAs with several distinct error terms and the absence of both theory and technology for regressions involving two or more error terms makes the failure of this statement, at least with today's notion of regression, quite clear."

REFERENCES

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Darlington, R. B., & Hayes, A. F. (2017). *Regression and linear models: Concepts, applications, and implementation*. New York: Guilford.

Green, S. B., Marquis, J. G., Hershberger, S. L., Thompson, M. S., & McCollam, K. M. (1999). The overparameterized analysis of variance model. *Psychological Methods*, *4*, 214–233.

Judd, C. M., & McClelland, G. H. (2017). *Data analysis: A model-comparison approach to regression, ANOVA, and beyond (3rd ed.)*. New York: Taylor & Francis.

Littell, R. C., Freund, R. J., & Spector, P. C. (1991). *SAS system for linear models* (3rd ed.). Cary, NC: SAS Institute.

Mallows, C. L., & Tukey, J. W. (1982). An overview of techniques of data analysis, emphasizing its exploratory aspects. In J. T. deOliveira & B. Epstein (Eds.), *Some recent advances in statistics* (pp. 111–172). London, England: Academic Press.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181–190.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Orlando, FL: Harcourt Brace.

Serlin, R. C., & Levin, J. R. (1985). Teaching how to derive directly interpretable coding schemes for multiple regression analysis. *Journal of Educational Statistics*, *10*, 223–238.

Vargha, A., Rudas, T., Delaney, H. D., & Maxwell, S. E. (1996). Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics*, *21*, 264–282.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, *64*, 1–48.