10
More on Multicategorical Regressors

This chapter builds on the discussion of multicategorical regressors in Chapter 9 by introducing several additional methods of coding groups. Two of these methods, sequential and Helmert coding, are particularly useful when the multicategorical regressor is categorical and ordinal. We also discuss statistical tests of complex contrasts of means, both with and without covariates.

We described in Chapter 9 how a categorical variable representing $g \geq 3$ groups can be used as a regressor in a linear model if it is represented with $g - 1$ indicator variables. The gth group does not require its own indicator because it would not contain any information about group membership not already contained in the $g - 1$ indicators in the model. When using indicator coding, the group not given an indicator serves as the reference group, and regression coefficients for the indicators quantify the difference in Y between the group coded with an indicator and the reference group.

Using this system of coding groups, regression analysis can be used to compare g group means either with or without additional variables in the model serving as covariates. In this chapter, we discuss some other methods of coding groups that produce mathematically identical models, in that they fit just as well and produce the same estimates of Y, yet yield regression coefficients with different interpretations. We also introduce some methods for conducting complex contrasts between group means, formed by combining group means together in various ways to test whether one set of group means, when aggregated, differ from another group mean or set of means.

TABLE 10.1. Age and Willingness to Self-Censor

ID	Age cohort	*cohort*	*wtsc* (*Y*)
1	Baby boomer	3	2.75
2	Pre-baby boomer	4	3.50
3	Pre-baby boomer	4	2.75
4	Baby boomer	3	2.25
5	Generation X	2	4.00
⋮	⋮	⋮	⋮
457	Baby boomer	3	2.87
458	Pre-baby boomer	4	2.75
459	Generation X	2	3.00
460	Baby boomer	3	2.50
461	Generation Y	1	2.75

10.1 Alternative Coding Systems

Indicator coding is only one of many ways of representing a multicategorical variable in a linear regression model. Two of these alternatives, *sequential coding* and *Helmert coding*, are particularly useful when the groups can be ordered relative to each other on the variable used to define the groups (though these two coding methods can be used for strictly nominal categories as well). Another alternative called *effect coding* is similar to indicator coding but changes the reference against which the groups are compared.

We rely on a data set containing the responses of 461 people living in the United States and the United Kingdom to a set of questions on a survey administered through the Internet. An excerpt from the data file (named WTSC and downloadable from this book's web page at *www.afhayes.com*) can be found in Table 10.1. The variable in the column labeled *wtsc* is scores on an instrument called the Willingness to Self-Censor Scale (Hayes, Glynn, & Shanahan, 2005). This instrument measures how reluctant versus willing a person is to express his or her opinion publicly when the person believes others hold a different opinion. Higher scores reflect a greater willingness to self-censor one's opinion expression. This is the dependent variable *Y* in all analyses in this chapter.

The data set also contains an ordinal categorical variable coding a respondent's age named *cohort*. The data were returned from the data collection company with each respondent classified into one of four age

TABLE 10.2. Willingness to Self-Censor in Four Age Cohorts

Group (j)	Age cohort	$\overline{Y}_j$	SD_{Y_j}	n_j
1	Generation Y (born after 1985)	3.201	0.494	38
2	Generation X (born 1966 – 1985)	3.111	0.622	149
3	Baby boomer (born 1945 – 1965)	2.857	0.468	173
4	Pre-baby boomer (born before 1945)	2.802	0.454	101

cohorts. Ordinally lowest in age is "Generation Y," the youngest group and born after 1985, and is coded cohort = 1 in the data. The data collection occurred in 2009, so all Generation Y respondents were 23 years old or younger (no one under 18 participated in the study). Following Generation Y is Generation X, coded cohort = 2, a group containing people born between 1966 and 1985 and thus between the ages of 24 and 43. Next comes the baby boomers born between 1945 and 1965 (cohort = 3, between 44 and 64 years old). The ordinally highest group in age is the pre-baby boomers. They were born before 1945, all at least 65 years old, and coded cohort = 4. Thus, in terms of age, pre-baby boomers > baby boomers > Generation X > Generation Y.

Each group's mean willingness to self-censor can be found in Table 10.2. As can be seen, it appears that the relationship between age and willingness to self-censor is negative, as successive increments up the ordinal age scale correspond to a lower mean willingness to self-censor.

Let's regress willingness to self-censor on age cohort using the indicator coding system introduced in Chapter 9 at the top of Table 10.3. This system codes Generation Y, Generation X, and baby boomers with D_1, D_2, and D_3, and pre-baby boomers are the reference category. The resulting model can be found at the top of Table 10.4. You can verify for yourself that this model generates the group means as its estimates for Y for the four groups. A test of the null hypothesis that ${}_TR = 0$ can be rejected, $F(3, 457) = 12.207, p < .001$. That is, the four age groups differ in their average willingness to self-censor.

10.1.1 Sequential (Adjacent or Repeated Categories) Coding

Sequential coding would most typically be used when the groups can be ordered on the variable that defines them and interest is in examining

TABLE 10.3. Four Ways of Coding Age Cohort and the Group Means Defined in Terms of the Regression Coefficients and Regression Constant

Age cohort by increasing age	D_1	D_2	D_3	Mean of Y
Indicator coding				
Generation Y	1	0	0	$\overline{Y}_1 = b_0 + b_1$
Generation X	0	1	0	$\overline{Y}_2 = b_0 + b_2$
Baby boomer	0	0	1	$\overline{Y}_3 = b_0 + b_3$
Pre-baby boomer	0	0	0	$\overline{Y}_4 = b_0$
Sequential coding				
Generation Y	0	0	0	$\overline{Y}_1 = b_0$
Generation X	1	0	0	$\overline{Y}_2 = b_0 + b_1$
Baby boomer	1	1	0	$\overline{Y}_3 = b_0 + b_1 + b_2$
Pre-baby boomer	1	1	1	$\overline{Y}_4 = b_0 + b_1 + b_2 + b_3$
Helmert coding				
Generation Y	−3/4	0	0	$\overline{Y}_1 = b_0 - \frac{3}{4}b_1$
Generation X	1/4	−2/3	0	$\overline{Y}_2 = b_0 + \frac{1}{4}b_1 - \frac{2}{3}b_2$
Baby boomer	1/4	1/3	−1/2	$\overline{Y}_3 = b_0 + \frac{1}{4}b_1 + \frac{1}{3}b_2 - \frac{1}{2}b_3$
Pre-baby boomer	1/4	1/3	1/2	$\overline{Y}_4 = b_0 + \frac{1}{4}b_1 + \frac{1}{3}b_2 + \frac{1}{2}b_3$
Effect coding				
Generation Y	1	0	0	$\overline{Y}_1 = b_0 + b_1$
Generation X	0	1	0	$\overline{Y}_2 = b_0 + b_2$
Baby boomer	0	0	1	$\overline{Y}_3 = b_0 + b_3$
Pre-baby boomer	−1	−1	−1	$\overline{Y}_4 = b_0 - b_1 - b_2 - b_3$

TABLE 10.4. Estimating Willingess to Self-Censor from Age Cohort Using the Coding Systems in Table 10.3

		Coeff.	*SE*	*t*	*p*
Indicator coding					
(pre-baby boomers as reference)					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.802	0.052	53.933	< .001
D_1	b_1	0.399	0.099	4.019	< .001
D_2	b_2	0.310	0.067	4.603	< .001
D_3	b_3	0.055	0.065	0.846	.398
Sequential coding					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	3.201	0.085	37.797	< .001
D_1	b_1	−0.090	0.095	−0.944	.346
D_2	b_2	−0.254	0.058	−4.361	< .001
D_3	b_3	−0.055	0.065	−0.846	.398
Helmert coding					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.993	0.029	103.898	< .001
D_1	b_1	−0.278	0.089	−3.133	.002
D_2	b_2	−0.282	0.054	−5.240	< .001
D_3	b_3	−0.055	0.065	−0.846	.398
Effect coding					
(pre-baby boomers uncoded)					
$R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.993	0.029	103.898	< .001
D_1	b_1	0.208	0.066	3.133	.002
D_2	b_2	0.119	0.042	2.841	.005
D_3	b_3	−0.136	0.040	−3.376	.001

TABLE 10.5. Sequential Coding of g Categories

Group	D_1	D_2	D_3	$\cdots$	D_{g-1}
1	0	0	0	$\cdots$	0
2	1	0	0	$\cdots$	0
3	1	1	0	$\cdots$	0
4	1	1	1	$\cdots$	0
$\vdots$					
g	1	1	1	$\cdots$	1

how $\overline{Y}_j$ changes as the ordinal predictor variable increases by one step. Like indicator coding, sequential coding relies on dummy variables. When using sequential coding with g groups, we set D_j to 1 for cases that are members of a group ordinally higher than position j on the variable defining groups; otherwise, we set to D_j to 0.

Table 10.5 provides a general representation of sequential coding with g ordered groups, and Table 10.3 provides the sequential codes for coding four groups as in this example. In this case, we set D_1 to 1 for anyone older than Generation Y, and Generation Y gets $D_1 = 0$. Moving up the ordinal age scale, D_2 is set to 1 for anyone older than Generation X, and Generations X and Y receive $D_2 = 0$. Finally, anyone older than the baby boomers (the pre-baby boomers) receives a code of $D_3 = 1$, and all others get 0 on D_3.

Regressing willingness to self-censor on the set of three sequential codes yields the regression model in Table 10.4. As can be seen, the model is

$$\hat{Y} = 3.201 - 0.090D_1 - 0.254D_2 - 0.055D_3$$

and has exactly the same R (and thus the same $SS_{residual}$ and other measures of fit) as when indicator coding was used. The outcome of the test as to whether ${}_TR = 0$ is the same as well, with the same F-ratio, degrees of freedom, and p-value. We can reject the null hypothesis and conclude that the groups differ in their average willingness to self-censor. Furthermore, plugging values of D_1, D_2, and D_3 into the model generates the four group means, just as does the model based on indicator coding:

$$\begin{aligned}
\hat{Y}_1 &= 3.201 - 0.090(0) - 0.254(0) - 0.055(0) = 3.201 = \overline{Y}_1 \\
\hat{Y}_2 &= 3.201 - 0.090(1) - 0.254(0) - 0.055(0) = 3.111 = \overline{Y}_2 \\
\hat{Y}_3 &= 3.201 - 0.090(1) - 0.254(1) - 0.055(0) = 2.857 = \overline{Y}_3 \\
\hat{Y}_4 &= 3.201 - 0.090(1) - 0.254(1) - 0.055(1) = 2.802 = \overline{Y}_4
\end{aligned}$$

So mathematically, this model is the same as the model based on indicator coding of groups. It produces the same estimates of Y, it fits identically, and it yields the same p-value when testing the null hypothesis that ${}_TR = 0$.

But there is an obvious difference between the two models in the regression coefficients and the constant. This is because these now quantify something different. Recall that with indicator coding, b_0 is $\overline{Y}$ for the reference group, and b_j is the mean difference in Y between the group receiving 1 on D_j and the reference group. But in sequential coding, b_0 is $\overline{Y}$ for the group ordinally lowest on the variable defining the groups, and b_j is the mean difference in Y between the group in ordinal position j and the group one ordinal position *lower*. In other words, b_j is the difference in means between categories that are ordinally adjacent on the variable defining groups. And the t- and p-value tests the null hypothesis that these two means are equal.

To see how this works, consider that for Generation Y,

$$\begin{aligned}
\overline{Y}_1 &= b_0 + b_1 0 + b_2 0 + b_3 0 \\
\overline{Y}_1 &= b_0
\end{aligned}$$

and for Generation X,

$$\begin{aligned}
\overline{Y}_2 &= b_0 + b_1 1 + b_2 0 + b_3 0 \\
\overline{Y}_2 &= b_0 + b_1.
\end{aligned}$$

But $b_0 = \overline{Y}_1$ and so

$$\overline{Y}_2 = \overline{Y}_1 + b_1$$

which can be rewritten as

$$b_1 = \overline{Y}_2 - \overline{Y}_1$$

So b_1 is the mean difference in Y between the two groups that are ordinally lowest in age. In this example, $b_1 = -0.090$, which is indeed $\overline{Y}_2 - \overline{Y}_1 = 3.111 - 3.201$ (from Table 10.2). These means are not statistically different

from each other, $t(457) = -0.944, p = .346$. Generation X is no more or less willing to self-censor, on average, than Generation Y.

This same reasoning leads to the derivation that b_2 is the mean difference Y between the groups in the second and third ordinal position. In this case, this is the baby boomers versus Generation X. For baby boomers,

$$\overline{Y}_3 = b_0 + b_1 1 + b_2 1 + b_3 0$$
$$\overline{Y}_3 = b_0 + b_1 + b_2$$

but $b_0 + b_1 = \overline{Y}_2$, and so

$$\overline{Y}_3 = \overline{Y}_2 + b_2$$

and isolation of b_2 results in

$$b_2 = \overline{Y}_3 - \overline{Y}_2.$$

In this example, $b_2 = -0.254$, which is $\overline{Y}_3 - \overline{Y}_2 = 2.857 - 3.111$. Baby boomers are less willing to self-censor, on average, than Generation X, $t(457) = -4.361, p < .001$. Following this same logic leads to the conclusion that pre-baby boomers do not differ significantly from baby boomers, on average, in their willingness to self-censor, $b_3 = -0.055, t(457) = -0.846, p = .398$. Observe that $b_3 = \overline{Y}_4 - \overline{Y}_3 = 2.802 - 2.857$.

It should be apparent why sequential coding can also be called *adjacent categories* coding. It would be the coding system to use if you are interested in comparing how Y changes with incremental increases in the ordinal multicategorical predictor represented with the $g-1$ dummy variables. This could be especially useful when the g categories can be ranked on some a priori basis on some dimension such as cost or difficulty in implementation. For instance, perhaps five drugs differ in the amount they cost. Each b_j quantifies the increase in Y associated with each additional cost increase, and hypothesis tests formally examine whether the increase in Y associated with an additional step up in cost is statistically significant.

It might be apparent already to you that sequential coding does not require that the multicategorical variable be ordinal. It could be used for a nominal multicategorical variable as well if you strategically "ordered" the nominal categories in such a way that the regression coefficients that result quantify the mean differences of interest.

TABLE 10.6. Helmert Coding of Three or Four Ordinal Categories

Ordinal position (low to high)	D_1	D_2	D_3	Mean of Y
$g = 3$ groups				
1	−2/3	0	—	$\overline{Y}_1 = b_0 - (2/3)b_1$
2	1/3	−1/2	—	$\overline{Y}_2 = b_0 + (1/3)b_1 - (1/2)b_2$
3	1/3	1/2	—	$\overline{Y}_3 = b_0 + (1/3)b_1 + (1/2)b_2$
$g = 4$ groups				
1	−3/4	0	0	$\overline{Y}_1 = b_0 - (3/4)b_1$
2	1/4	−2/3	0	$\overline{Y}_2 = b_0 + (1/4)b_1 - (2/3)b_2$
3	1/4	1/3	−1/2	$\overline{Y}_3 = b_0 + (1/4)b_1 + (1/3)b_2 - (1/2)b_3$
4	1/4	1/3	1/2	$\overline{Y}_4 = b_0 + (1/4)b_1 + (1/3)b_2 + (1/2)b_3$

10.1.2 Helmert Coding

Sequential coding results in regression coefficients that quantify the difference between means for groups ordinally adjacent to each other on the variable defining groups. An alternative coding system useful for ordinal multicategorical variables is Helmert coding. This method of coding groups results in regression coefficients that quantify the difference in means between one group and the mean of the means of all groups ordinally higher on the multicategorical variable defining groups.

Table 10.6 shows a set of Helmert codes for three as well as four groups, and Table 10.7 provides the general algorithm for constructing codes for five or more groups. Regressing willingness to self-censor on D_1, D_2, and D_3 using the Helmert codes in Table 10.3, the resulting model is (see Table 10.4)

$$\hat{Y} = 2.993 - 0.278D_1 - 0.282D_2 - 0.055D_3$$

with the same R as when groups were coded with indicator or sequential codes, and the same F- and p-values for testing the null hypothesis that ${}_TR = 0$. And the model generates $\hat{Y}$ values that correspond to the groups means:

TABLE 10.7. Helmert Coding of g Categories, $g \geq 5$

Ordinal position (low to high)	D_1	D_2	D_3	$\cdots$	D_{g-1}
1	$-(g-1)/g$	0	0	$\cdots$	0
2	$1/g$	$-(g-2)/(g-1)$	0	$\cdots$	0
3	$1/g$	$1/(g-1)$	$-(g-3)/(g-2)$	$\cdots$	0
$\vdots$					
$g-1$	$1/g$	$1/(g-1)$	$1/(g-2)$	$\cdots$	$-1/2$
g	$1/g$	$1/(g-1)$	$1/(g-2)$	$\cdots$	$1/2$

$$\hat{Y}_1 = 2.993 - 0.278(-3/4) - 0.282(0) - 0.055(0) = 3.201 = \overline{Y}_1$$
$$\hat{Y}_2 = 2.993 - 0.278(1/4) - 0.282(-2/3) - 0.055(0) = 3.111 = \overline{Y}_2$$
$$\hat{Y}_3 = 2.993 - 0.278(1/4) - 0.282(1/3) - 0.055(-1/2) = 2.857 = \overline{Y}_3$$
$$\hat{Y}_4 = 2.993 - 0.278(1/4) - 0.282(1/3) - 0.055(1/2) = 2.802 = \overline{Y}_4$$

So mathematically, this model is no different than any other model of the groups means we have estimated in that it generates the same estimates of Y and fits exactly the same. But the regression coefficients are different, because they quantify different things now.

Tables 10.3 and 10.6 contains the formulas used to derive each group's mean from the regression model, and the computations above are completed for this example. As can be seen,

$$\overline{Y}_1 = b_0 - (3/4)b_1 \tag{10.1}$$

What cannot be seen quite as easily is that the mean of the three means for the groups ordinally higher than group 1 on the variable defining the groups is

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} = \frac{b_0 + b_1/4 - 2b_2/3}{3} + \frac{b_0 + b_1/4 + b_2/3 - b_3/2}{3} + \frac{b_0 + b_1/4 + b_2/3 + b_3/2}{3}$$

which, happily, reduces to a much simpler form:

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} = b_0 + (1/4)b_1 \tag{10.2}$$

Substraction of equation 10.1 from equation 10.2 yields

$$(b_0 + b_1/4) - (b_0 - 3b_1/4) = \frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1$$
$$b_1 = \frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1$$

and so b_1 quantifies the difference between $\overline{Y}_1$ and the average of $\overline{Y}_2$, $\overline{Y}_3$, and $\overline{Y}_4$. Indeed, observe in this example that

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1 = \frac{3.111 + 2.857 + 2.802}{3} - 3.201 = -0.278 = b_1$$

The t-statistic and p-value are used to test whether these two means are statistically different. In this example, we can conclude that generations older than Generation Y are less willing to self-censor on average than are members of Generation Y, $b_1 = -0.278, t(457) = -3.133, p = .002$.

Similar derivations lead to similar interpretations of b_2 and b_3:

$$b_2 = \frac{\overline{Y}_3 + \overline{Y}_4}{2} - \overline{Y}_2$$
$$b_3 = \overline{Y}_4 - \overline{Y}_3$$

which is indeed the case in this example:

$$b_2 = \frac{2.857 + 2.802}{2} - 3.111 = -0.282$$
$$b_3 = \overline{Y}_4 - \overline{Y}_3 = 2.802 - 2.857 = -0.055$$

Generations older than Generation X are less willing to self-censor on average than are members of Generation X, $b_2 = -0.282, t(457) = -5.240, p < .001$, but there is no statistically significant difference in average willingness to self-censor between baby boomers and pre-baby boomers, $b_3 = -0.055, t(457) = -0.846, p = .398$. Notice that b_3 is the same with Helmert coding as it was in section 10.1.1 when using sequential coding.

So when using Helmert coding, b_j, the regression coefficient for D_j, estimates the difference between $\overline{Y}_j$ and the unweighted mean of means for all groups ordinally higher than group j on the variable defining groups. The t- and p-values can be used to test the null hypothesis that these means are equal.

Thus far we have neglected the regression constant, b_0. In this example, $b_0 = 2.993$, which is equal to the mean of the four group means:

$$b_0 = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4}$$
$$b_0 = \frac{3.201 + 3.111 + 2.857 + 2.802}{4}$$
$$b_0 = 2.993$$

More generally, when using Helmert coding in this fashion (and assuming no other regressors are in the model, as in this case), the regression constant is the unweighted mean of all the group means.

A variation on Helmert coding is *reverse* Helmert coding. With reverse Helmert coding, the regression coefficient b_j quantifies the difference between the mean of Y for the group in ordinal position j on the variable defining groups and the unweighted average mean of Y for all groups ordinally *lower* than position j. Although reverse Helmert coding has a different name, there is no need here to provide detail about how the codes are constructed. This is because you can mimic reverse Helmert coding by using ordinary Helmert coding of the ordinal categories as in Tables 10.3, 10.6, and 10.7, but after first ordering the groups on the variable that defines them from high to low rather than low to high.

Helmert coding can be useful even when the multicategorical variable is nominal. For example, perhaps you have conducted an experiment with four conditions that consist of a control group and three experimental treatment conditions, with the treatment being a manipulation of a variable that is not quantitative in any sense of the word. If you use numerical codes for the four conditions strategically, then you can use Helmert coding (though it wouldn't generally be called this) to set up a set of comparisons between the mean of group 1 (say, the control group) versus the mean of the three treatment groups, the mean of the first treatment group versus the mean of the other two treatment groups, and the mean of the second treatment group versus the mean of the third treatment group.

10.1.3 Effect Coding

Effect coding is a minor variation on indicator coding, but the reference against which group means are compared changes. Recall that in indicator coding, one of the g groups receives a code of zero on all indicator variables, and that group ends up the reference group against which all other group means are compared.

With effect coding, a set of $g - 1$ variables D_j are constructed just as in indicator coding, but the group left "uncoded" is set to -1 on all D_j rather 0, as in Table 10.8. Thus, the $g - 1$ D_j variables are no longer dummy variables, as they contain three values (0, 1, or -1) rather than only two. This minor change in coding has an important effect on the interpretation of the regression coefficients and the constant relative to indicator coding.

When willingness to self-censor is regressed on D_1, D_2, and D_3 using the effect coding system for age in Table 10.3, the resulting model is (see Table 10.4)

$$\hat{Y} = 2.993 + 0.208D_1 + 0.119D_2 - 0.136D_3$$

with the same R as when groups were coded with indicator, sequential, or Helmert codes, and the same F- and p-value for testing the null hypothesis that ${}_TR = 0$. And the model generates $\hat{Y}$ values that equal the group means:

$$\hat{Y}_1 = 2.993 + 0.208(1) + 0.119(0) - 0.136(0) = 3.201 = \overline{Y}_1$$
$$\hat{Y}_2 = 2.993 + 0.208(0) + 0.119(1) - 0.136(0) = 3.111 = \overline{Y}_2$$
$$\hat{Y}_3 = 2.993 + 0.208(0) + 0.119(0) - 0.136(1) = 2.857 = \overline{Y}_3$$
$$\hat{Y}_4 = 2.993 + 0.208(-1) + 0.119(-1) - 0.136(-1) = 2.802 = \overline{Y}_4$$

So mathematically, this model is no different than any other model of the groups means we have estimated, in that it generates the same estimates of Y and fits exactly the same.

With indicator coding, b_j quantifies the difference between the mean of the group coded by D_j and the reference group. But with effect coding, b_j is the difference in the mean of group j and the mean of all g group means. That is,

$$b_j = \overline{Y}_j - \frac{\overline{Y}_1 + \overline{Y}_2 + \cdots + \overline{Y}_g}{g}$$

and the t- and p-values for each b_j tests the null hypothesis that $\overline{Y}_j$ equals the mean of all group means. Assuming no additional variables are in

TABLE 10.8. Effect Coding of g Categories

Group	D_1	D_2	$\cdots$	D_j	$\cdots$	D_{g-1}
1	1	0	$\cdots$	0	$\cdots$	0
2	0	1	$\cdots$	0	$\cdots$	0
$\vdots$						
j	0	0	$\cdots$	1	$\cdots$	0
$\vdots$						
$g-1$	0	0	$\cdots$	0	$\cdots$	1
g	−1	−1	$\cdots$	−1	$\cdots$	−1

the model, the regression constant is that unweighted mean of all g group means. For example, from Table 10.2

$$b_0 = \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = \frac{3.201 + 3.111 + 2.857 + 2.802}{4} = 2.993$$

and

$$b_1 = \overline{Y}_1 - \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = 3.201 - 2.993 = 0.208$$

$$b_2 = \overline{Y}_2 - \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = 3.111 - 2.993 = 0.119$$

$$b_3 = \overline{Y}_3 - \frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{4} = 2.857 - 2.993 = -0.136$$

all of which correspond to the model coefficients from Table 10.4. We can conclude that Generation Y is more willing to self-censor than average, $t(457) = 3.133, p = .002$, as is Generation X, $t(457) = 2.841, p = .005$. But baby boomers are less willing to self-censor than average, $t(457) = -3.376, p = .001$

Missing from this analysis is a comparison of the pre-baby boomers to the average. This finding is sacrificed by the requirement that only $g - 1$ variables coding group can be used in the model. But this comparison can be obtained by rerunning the analysis, setting a different group to receive the −1 codes on all D_j.

10.2 Comparisons and Contrasts

10.2.1 Contrasts

Many questions about differences between group means can be phrased as questions about *contrasts*. The simplest type of contrast is a *pairwise comparison*, which is the difference between two means. Some of the coding systems described in sections 9.1.1 and 10.1 produce regression coefficients and hypothesis tests that yield pairwise comparisons, such as the $g-1$ comparisons between each group mean and a reference group mean when using indicator coding or between group means for groups ordinally adjacent on the ordinal, multicategorical variable when using sequential coding.

More complex contrasts involve more than two means. For instance, perhaps an investigator is entertaining the efficacy of five different therapies for the treatment of depression. Perhaps methods 1, 2, and 3 are all based on principles of theory A about how people think and feel, and methods 4 and 5 are based on a second theoretical orientation B, perhaps one that also includes the use of medication. One might want to know whether clients who are treated with one of the theory A methods differ, on average, in depression 6 months later than clients treated with one of the theory B methods. In this example, the mean depression of those those treated by theory A could be expressed as $(\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3)/3$ and the mean depression of those treated by theory B would be $(\overline{Y}_4 + \overline{Y}_5)/2$. The difference between these,

$$\frac{\overline{Y}_1 + \overline{Y}_2 + \overline{Y}_3}{3} - \frac{\overline{Y}_4 + \overline{Y}_5}{2}$$

is a more complex comparison involving several group means rather than a simple pairwise comparison.

Any contrast, whether a pairwise comparison or more complex, can be expressed as a weighted sum of means of the form

$$\text{Contrast} = \sum_{j=1}^{g} c_j \overline{Y}_j \tag{10.3}$$

where c_j is the *contrast coefficient* for group j and $\sum c_j = 0$. For instance, the complex contrast above can be expressed as

$$\text{Contrast} = (1/3)\overline{Y}_1 + (1/3)\overline{Y}_2 + (1/3)\overline{Y}_3 + (-1/2)\overline{Y}_4 + (-1/2)\overline{Y}_5 \tag{10.4}$$

which is a weighted sum of means as in equation 10.3 with $c_1 = c_2 = c_3 = 1/3$ and $c_4 = c_5 = -1/2$.

There is no requirement that all contrast coefficients be nonzero, *so long as they sum to zero.* For example, a pairwise comparison among a set of five means that compares only the means of groups 1 and 2 can be written as

$$\text{Contrast} = (1)\overline{Y}_1 + (-1)\overline{Y}_2 + (0)\overline{Y}_3 + (0)\overline{Y}_4 + (0)\overline{Y}_5$$

which is in the form of equation 10.3 with $c_1 = 1$, $c_2 = -1$ and $c_3 = c_4 = c_5 = 0$. Observe that this simplifies to $\overline{Y}_1 - \overline{Y}_2$.

We can multiply contrast coefficients by a constant without affecting the results of statistical tests of contrasts. This can be especially convenient when a coefficient in fractional form cannot be expressed in decimal form without some rounding or loss of precision (such as $1/3 = 0.33333\ldots$). For instance, the complex contrast in equation 10.4 can be expressed as

$$\text{Contrast} = (2)\overline{Y}_1 + (2)\overline{Y}_2 + (2)\overline{Y}_3 + (-3)\overline{Y}_4 + (-3)\overline{Y}_5$$

which results from multiplying all contrast coefficients by six. The resulting contrast will be six times larger as a result, but the standard error as generated by the formula in section 10.2.2 will also be six times larger to compensate, so the p-value from a hypothesis test is unaffected. We'll see this illustrated in section 10.2.3.

To illustrate the computations, let's use contrast coefficients to generate a contrast of the average willingness to self-censor of Generation Y compared to everyone else. That is,

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} - \overline{Y}_1$$

We'll also construct a contrast comparing the average willingness to self-censor of Generation X and Generation Y relative to baby boomers and pre-baby boomers, which is

$$\frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2}$$

For the former contrast, we use contrast coefficients of $c_1 = -1$, $c_2 = 1/3$, $c_3 = 1/3$, and $c_4 = 1/3$, for Generation Y, Generation X, baby boomers, and

pre-baby boomers, respectively. Notice these add up to zero, as required for a proper contrast. Applying equation 10.3 yields

$$\begin{aligned}\text{Contrast} &= (1/3)\overline{Y}_2 + (1/3)\overline{Y}_3 + (1/3)\overline{Y}_4 - \overline{Y}_1 \\ &= (1/3)3.111 - (1/3)2.857 - (1/3)2.802) - (1)\overline{Y}_1 \\ &= -0.278\end{aligned}$$

which is interpreted to mean that Generation Y is estimated to be, on average, 0.278 units higher in willingness to self-censor than the average willingness to self-censor of Generation X, baby boomers, and pre-baby boomers. Note that this contrast is identical to b_1 from the regression model when using Helmert coding of groups (see Table 10.4).

The second contrast requires contrast coefficients of $c_1 = 1/2$, $c_2 = 1/2$, $c_3 = -1/2$, and $c_4 = -1/2$, respectively, which add to zero as required. Applying equation 10.3 produces

$$\begin{aligned}\text{Contrast} &= (1/2)\overline{Y}_1 + (1/2)\overline{Y}_2 + (-1/2)\overline{Y}_3 + (-1/2)\overline{Y}_4 \\ &= (1/2)3.201 + (1/2)3.111 - (1/2)2.857 - (1/2)2.802 \\ &= 0.327\end{aligned}$$

So Generations X and Y are estimated as, on average, 0.327 units higher in willingness to self-censor than baby boomers and pre-baby boomers.

10.2.2 Computing the Standard Error of a Contrast

For inference, we need an estimate of the standard error of the contrast. When regression analysis is used to emulate analysis of variance with g groups, some simple hand computations yield the standard error using only the contrast coefficients, group sample sizes, and $MS_{residual}$ from the regression. The formula is

$$SE(\text{contrast}) = \sqrt{MS_{residual} \sum_{j=1}^{g} \frac{c_j^2}{n_j}} \tag{10.5}$$

With the standard error for a contrast computed, a test of significance for the null hypothesis that the contrast equals zero can be conducted using

$$t = \frac{\text{Contrast}}{SE(\text{contrast})}$$

and generating a p-value using the $t(df_{residual})$ distribution. Alternatively, a confidence interval can be constructed in the usual way as the point estimate plus or minus t_{crit} standard errors, where t_{crit} is from a table of critical values of t for an interval corresponding to a certain degree of confidence (see Appendix C).

For the contrast comparing the mean of Generation Y against the mean of the other three group means, applying equation 10.5 using the means and group sample sizes in Table 10.2 and the $MS_{residual}$ from any of the regression models in Table 10.4 results in

$$\begin{aligned} SE(\text{contrast}) &= \sqrt{0.273\left(\frac{(-1)^2}{38} + \frac{(1/3)^2}{149} + \frac{(1/3)^2}{173} + \frac{(1/3)^2}{101}\right)} \\ &= 0.089 \end{aligned}$$

and so $t(457) = -0.278/0.089 = -3.133, p < .001$. This contrast of means is statistically significant. Notice that the standard error of this contrast is identical to the standard error of b_1 in the model of Y using Helmert coding. So clearly, given that this contrast is just a comparison of the ordinally lowest age group against all others, much work is saved conducting this contrast by just using Helmert coding and regressing willingness to self-censor on the Helmert codes.

None of the coding systems described in section 10.1 yield the second contrast results comparing the mean of the means of Generation X and Generation Y to the mean of the means of baby boomers and pre-baby boomers. The estimated standard error of this contrast is

$$\begin{aligned} SE(\text{contrast}) &= \sqrt{0.273\left(\frac{(1/2)^2}{38} + \frac{(1/2)^2}{149} + \frac{(-1/2)^2}{173} + \frac{(-1/2)^2}{101}\right)} \\ &= 0.058 \end{aligned}$$

and so $t(457) = 0.327/0.058 = 5.638$, which is statistically significant, $p < .001$.

10.2.3 Contrasts Using Statistical Software

These computations need not be conducted by hand if you have a statistics program capable of doing them. Most good programs can these days. You will probably find options for conducting contrasts in your program's ANOVA routine rather than its regression module, as historically it is in

Descriptives

WTSC: Willingness to Self-Censor

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Generation Y	38	3.2013	.49403	.08014	3.0389	3.3637	2.12	4.37
Generation X	149	3.1117	.62201	.05096	3.0110	3.2124	1.25	5.00
Baby boomer	173	2.8573	.46793	.03558	2.7871	2.9275	1.87	4.50
Pre baby boomer	101	2.8020	.45420	.04519	2.7123	2.8916	1.75	4.00
Total	461	2.9558	.54086	.02519	2.9063	3.0053	1.25	5.00

ANOVA

WTSC: Willingness to Self-Censor

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	9.983	3	3.328	12.207	.000
Within Groups	124.581	457	.273		
Total	134.564	460			

Contrast Coefficients

COHORT: Age cohort

Contrast	Generation Y	Generation X	Baby boomer	Pre baby boomer
1	-3	1	1	1
2	.5	.5	-.5	-.5

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
WTSC: Willingness to Self-Censor	Assume equal variances	1	-.8329	.26584	-3.133	457	.002
		2	.3269	.05762	5.674	457	.000
	Does not assume equal variances	1	-.8329	.25241	-3.300	44.897	.002
		2	.3269	.05551	5.888	125.437	.000

FIGURE 10.1. SPSS output from a one-way ANOVA with two contrasts.

the context of ANOVA that contrasts are usually introduced in textbooks and classrooms.

In SPSS, for example, the command below will conduct an analysis of variance testing for a difference in mean willingness to self-censor between the four age cohorts, while also conducting the two contrasts described earlier by specifying the appropriate contrast coefficients following the **`contrast`** option.

```
oneway wtsc by cohort/contrast -3 1 1 1/contrast 0.5 0.5 -0.5 -0.5
   /statistics descriptive.
```

In this command, the coefficients for the first contrast were multiplied by 3. SPSS's ONEWAY module does not allow fractions such as "1/3" in the contrast line, and 1/3 cannot be represented in decimal form exactly.

Observe from the output in Figure 10.1 that the resulting contrast is three times larger than when computed using fractional coefficients, but the standard error is also three times larger. As a result, the t-ratio and p-value are the same as when fractional coefficients are used. Because 1/2 can be represented exactly in decimal form, there is no need to multiply the coefficients by a constant for the second contrast.

Comparable code for SAS is

```
proc glm data=wtsc;
   class cohort;model wtsc=cohort;means cohort;
   contrast '1 vs 2 3 4' cohort -3 1 1 1;
   contrast '1 2 vs 3 4' cohort 0.5 0.5 -0.5 -0.5;
run;
```

In SAS you must provide a name for the contrast in quotes, as above, prior to listing the coefficients. SAS will produce the contrasts in the form of F-ratios with 1 and $df_{residual}$ degrees of freedom, along with a p-value corresponding to the test of the null that the contrast equals zero.

SPSS's UNIANOVA module has some options built in to do contrasts that correspond to the coding systems described in this chapter. For instance, the command below conducts a one-way ANOVA while also producing output for contrasts equivalent to those generated by Helmert, sequential (**repeated**), and indicator (**simple**) coding of groups.

```
unianova wtsc by cohort/emmeans=tables(cohort)/contrast (cohort)=
   helmert/contrast (cohort)=repeated/contrast (cohort)=simple.
```

Consult your preferred program's documentation to see if it is capable of doing comparable analyses.

10.2.4 Covariates and the Comparison of Adjusted Means

Adjusted means were introduced in sections 9.2.4 and 9.2.6 as estimates of group means if all groups were average on a covariate or covariates. We saw in those sections that the regression coefficients for indicator codes can be interpreted as differences between adjusted means whenever a covariate is included in the model along with the codes for groups, and hypothesis tests or confidence intervals used for inference.

Covariates can be included in a model when groups are represented with any coding system, including sequential, Helmert, and effect coding.

When covariates are included, the interpretation we gave to the regression coefficients in section 10.1 apply to adjusted means rather than to the unadjusted means.

To illustrate, we examine differences between the four age groups in willingness to self-censor, with shyness used as as a covariate and using Helmert coding of age cohort. Research shows that people who are relatively higher in willingness to self-censor are also relatively higher on measures of shyness (Hayes et al., 2005), so it is worth examining whether the differences in willingness to self-censor exist independent of any differences between groups in their average shyness. A measure of shyness was included in the survey and is available in the data file, so it is a simple matter to adjust for shyness by simply including it as an additional regressor in the model. The resulting regression equation is

$$\hat{Y} = 2.187 - 0.163D_1 - 0.133D_2 + 0.028D_3 + 0.281X_1 \qquad (10.6)$$

where X_1 is shyness. Corresponding regression output (from SAS, though SPSS and STATA output provide the same information) can be found in Figure 10.2. Applying the test discussed in section 9.2.2 results in $SR^2 = \Delta R^2 = .019, F(3, 456) = 4.256, p = .006$. So the groups differ on average in willingness to self-censor even after accounting for differences between them in shyness.

Setting shyness to the sample mean (in the data, $\overline{X}_1 = 2.832$) and plugging the Helmert codes into equation 10.6 generates the adjusted mean willingness to self-censor for each group:

$$\hat{Y}_1 = 2.187 - 0.163(-3/4) - 0.133(0) + 0.028(0) + 0.281(2.832) = 3.105$$
$$\hat{Y}_2 = 2.187 - 0.163(1/4) - 0.133(-2/3) + 0.028(0) + 0.281(2.832) = 3.031$$
$$\hat{Y}_3 = 2.187 - 0.163(1/4) - 0.133(1/3) + 0.028(-1/2) + 0.281(2.832) = 2.884$$
$$\hat{Y}_4 = 2.187 - 0.163(1/4) - 0.133(1/3) + 0.028(1/2) + 0.281(2.832) = 2.912$$

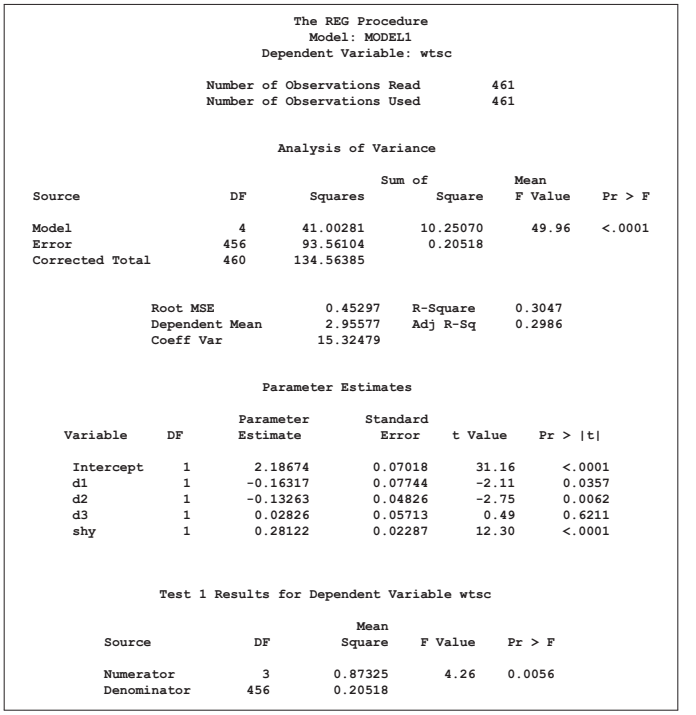
The computations described in section 10.1.2 but substituting the adjusted means for the unadjusted means reveals that the regression coefficients quantify differences between adjusted means (or means of adjusted means):

$$
\begin{aligned}
b_1 &= \frac{\hat{Y}_2 + \hat{Y}_3 + \hat{Y}_4}{3} - \hat{Y}_1 \\
&= \frac{3.031 + 2.884 + 2.912}{3} - 3.105 \\
&= -0.163 \\
b_2 &= \frac{\hat{Y}_3 + \hat{Y}_4}{2} - \hat{Y}_2 \\
&= \frac{2.884 + 2.912}{2} - 3.031 \\
&= -0.133 \\
b_3 &= \hat{Y}_4 - \hat{Y}_3 \\
&= 2.912 - 2.884 \\
&= 0.028
\end{aligned}
$$

Standard errors for these differences are available in regression output, along with t- and p-values and confidence intervals if desired. As can be seen in Figure 10.2, holding shyness constant (at the mean or any other value else), Generation Y is more willing to self-censor than those older; $t(456) = -2.107, p = 0.036$; and Generation X is more willing to self-censor than those older; $t(456) = -2.748, p = .006$; but baby boomers and pre-baby boomers do not differ significantly in willingness to self-censor; $t(456) = 0.495, p = .621$.

Complex contrasts between means were introduced in section 10.2.1. A contrast is a weighted sum of means, with the weighting determined by a group's contrast coefficient. Although equation 10.3 can be applied to adjusted means, the standard error of a contrast involving weighted means cannot be calculated using equation 10.5. The proper formula is complex, especially when more than one covariate is in the model. It is best to leave the production of the standard error for a complex contrast involving adjusted means to a computer.

In SPSS, the command below produces a complex contrast comparing the adjusted mean of Generation X to the mean of the adjusted means of all other groups, as well as a contrast comparing the mean of the adjusted means for Generations X and Y against mean of the adjusted means for baby boomers and pre-baby boomers. See a discussion of this latter contrast in section 10.2.1 for how the contrast coefficients are selected. The result of the first contrast is identical to the estimate and hypothesis test for b_1



```
                          The REG Procedure
                            Model: MODEL1
                       Dependent Variable: wtsc

                  Number of Observations Read         461
                  Number of Observations Used         461


                          Analysis of Variance

                                       Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      4       41.00281        10.25070      49.96    <.0001
Error                    456       93.56104         0.20518
Corrected Total          460      134.56385


          Root MSE              0.45297    R-Square     0.3047
          Dependent Mean        2.95577    Adj R-Sq     0.2986
          Coeff Var            15.32479


                       Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1        2.18674        0.07018      31.16      <.0001
d1            1       -0.16317        0.07744      -2.11      0.0357
d2            1       -0.13263        0.04826      -2.75      0.0062
d3            1        0.02826        0.05713       0.49      0.6211
shy           1        0.28122        0.02287      12.30      <.0001


          Test 1 Results for Dependent Variable wtsc

                                  Mean
     Source             DF       Square    F Value    Pr > F

     Numerator           3      0.87325       4.26    0.0056
     Denominator       456      0.20518
```

FIGURE 10.2. SAS output from a regression estimating willingness to self-censor from age cohort controlling for shyness.

in the example above when Helmert coding was used to code groups. The second shows that the means of these adjusted means are statistically different, contrast = −0.170, $t(456) = 3.295$, $p < .01$. SAS produces the result in the form of an F-ratio rather than a t-statistic.

```
glm wtsc by cohort with shy/emmeans=tables(cohort)/
    lmatrix cohort -1 1/3 1/3 1/3/lmatrix cohort -0.5 -0.5 0.5 0.5.
```

In SAS the comparable commands are

```
proc glm data=wtsc;
    class cohort;model wtsc=cohort shy;lsmeans cohort;
    contrast '1 vs 2 3 4' cohort 3 -1 -1 -1;
```

```
    contrast '1 2 vs 3 4' cohort 0.5 0.5 -0.5 -0.5;
run;
```

10.3 Weighted Group Coding and Contrasts

Section 10.1 described various methods for coding a multicategorical variable that produced regression coefficients that correspond to a comparison between two means. For indicator and sequential coding, the regression coefficients for each code quantified a difference between the means of two and only two groups. But for Helmert and effect coding, the regression coefficients quantified the difference between one group mean and an *unweighted* mean of the means of two or more groups.

For example, when Helmert coding was used in section 10.1.2, the regression coefficient of −0.278 for D_1 quantified the difference in mean willingness to self-censor between Generation Y (3.201) and the mean of the three means for the groups older than Generation Y (2.923). The mean for everyone older than Generation Y was constructed as

$$\frac{\overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4}{3} = \frac{3.111 + 2.857 + 2.802}{3} = 2.923$$

This is an unweighted mean, in that ignores the differences in sample sizes between the three groups that contribute to it. Notice from Table 10.2 that there are 149 in the sample from Generation X, 173 baby boomers, and 101 pre-baby boomers. So the 101 pre-baby boomers contribute as much to the construction of this mean of means as the 173 baby boomers, even though there are substantially fewer pre-baby boomers in the data. If this bothers you, then read this section, where we describe versions of Helmert, effect coding, and contrasts that acknowledge differences between the group sample sizes whenever one of the means being compared is formed as a mean of means.

10.3.1 Weighted Effect Coding

We saw in section 10.1.3 that the youngest three cohorts (Generation Y, Generation X, and baby boomers) differ from average in their willingness to self-censor, where *average* was defined as the mean of the four group means. But this was an unweighted average of the four group means, meaning it ignored the fact that the four age cohorts differ in size. If you want the average against which each mean is compared when using effect

TABLE 10.9. Weighted Effect Coding of *g* Categories

Group (j)	D_1	D_2	$\cdots$	D_j	$\cdots$	D_{g-1}
1	1	0	$\cdots$	0	$\cdots$	0
2	0	1	$\cdots$	0	$\cdots$	0
$\vdots$						
j	0	0	$\cdots$	1	$\cdots$	0
$\vdots$						
$g-1$	0	0	$\cdots$	0	$\cdots$	1
g	$-n_1/n_g$	$-n_2/n_g$	$\cdots$	$-n_j/n_g$	$\cdots$	$-n_{g-1}/n_g$

coding to incorporate group size, you can use *weighted* effect coding. It requires replacing the −1 codes used in effect coding with ratios of group sizes. More specifically, if n_g is the sample size for the group coded −1 on all D_j, replace the −1 for D_j with $-n_j/n_g$. See Table 10.9.

For example, in these data there are 38 people from Generation Y, 149 people from Generation X, 173 baby boomers, and 101 pre-baby boomers. Thus, we change the −1 values for D_j for the pre-baby boomers to $D_1 = -38/101$, $D_2 = -149/101$, and $D_3 = -173/101$ (see Table 10.10). Regressing willingness to self-censor on these weighted effect codes yields

$$\hat{Y} = 2.956 + 0.246D_1 + 0.156D_2 - 0.098D_3$$

(see Table 10.11). As with all the other coding systems used in section 10.1, the model fits the same, and it reproduces the group means. Now b_0 is the weighted mean of means (which is equivalent to just calculating the average of Y ignoring age cohort entirely):

$$b_0 = \frac{n_1\overline{Y}_1 + n_2\overline{Y}_2 + n_3\overline{Y}_3 + n_4\overline{Y}_4}{n_1 + n_2 + n_3 + n_4}$$

$$b_0 = \frac{38(3.201) + 149(3.111) + 173(2.857) + 101(2.802)}{461}$$

$$b_0 = 2.956$$

and b_j is the difference between the mean Y for the group receiving $D_j = 1$ and the weighted mean of all g group means on Y:

$$b_1 = \overline{Y}_1 - 2.956 = 3.201 - 2.956 = 0.246$$
$$b_2 = \overline{Y}_2 - 2.956 = 3.111 - 2.956 = 0.156$$
$$b_3 = \overline{Y}_3 - 2.956 = 2.857 - 2.956 = -0.098$$

As when using unweighted effect coding, we conclude that Generation Y is more willing to self-censor than average, $t(457) = 3.026, p = .003$, as is Generation X, $t(457) = 4.433, p < .001$, whereas baby boomers are less willing to self-censor than average, $t(457) = -3.139, p = .002$.

10.3.2 Weighted Helmert Coding

Weighted Helmert coding is comparable to Helmert coding, in that it generates regression coefficients that compare the mean Y of one group to the mean Y of all groups ordinally higher on the variable coding groups. However, for weighted Helmert coding the mean of Y for all groups higher than ordinal position j is a weighted mean rather than an unweighted mean.

There is no way of representing how to generate weighted Helmert codes with a simple algorithm in table form as in Table 10.7. Construction of weighted Helmert codes requires matrix algebra. But an understanding of matrix algebra is not required to implement this coding system using the syntax we provide at the end of the section. However, you do need to know how to construct the matrix that is used as input into the syntax.

The first step is the construction of a $g \times (g - 1)$ matrix that takes the form in Table 10.12, where g is the number of groups and n_{j+} is the sum of the sample sizes for groups in ordinal position j or higher on the variable defining the groups. That is,

$$n_{j+} = \sum_{i=j}^{g} n_i$$

For example, from the size of the age cohorts in the willingness to self-censor data (see Table 10.2), $n_{2+} = n_2 + n_3 + n_4 = 149 + 173 + 101 = 423$,

TABLE 10.10. Weighted Effect and Helmert Coding and the Group Means Defined in Terms of the Regression Coefficients and Regression Constant

Age cohort by increasing age	D_1	D_2	D_3
Weighted effect coding			
Generation Y	1	0	0
Generation X	0	1	0
Baby boomer	0	0	1
Pre-baby boomer	−38/101	−149/101	−173/101
Weighted Helmert coding			
Generation Y	−.7500000000	−.0141843972	−.0656934307
Generation X	.2500000000	−.6619385343	−.0656934307
Baby boomer	.2500000000	.3380614657	−.4343065693
Pre-baby boomer	.2500000000	.3380614657	.5656934307

TABLE 10.11. Estimating Willingness to Self-Censor from Age Cohort Using the Coding Systems in Table 10.10

		Coeff.	*SE*	*t*	*p*
Weighted effect coding (Pre-baby boomers uncoded) $R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.956	0.024	121.550	< .001
D_1	b_1	0.246	0.081	3.026	.003
D_2	b_2	0.156	0.035	4.443	< .001
D_3	b_3	−0.098	0.031	−3.139	.002
Weighted Helmert coding $R = 0.272, F(3, 457) = 12.207, p < .001$					
Constant	b_0	2.993	0.029	103.898	< .001
D_1	b_1	−0.268	0.088	−3.026	.003
D_2	b_2	−0.275	0.053	−5.172	< .001
D_3	b_3	−0.055	0.065	−0.846	.398

TABLE 10.12. Construction of the Input Matrix for Weighted Helmert Coding

	Column				
	1	2	3	$\cdots$	$g-1$
Row					
1	-1	0	0	$\cdots$	0
2	n_2/n_{2+}	-1	0	$\cdots$	0
3	n_3/n_{2+}	n_3/n_{3+}	-1	$\cdots$	0
$\vdots$					
$g-1$	n_{g-1}/n_{2+}	n_{g-1}/n_{3+}	$n_{g-1}/n_{(g-1)+}$	$\cdots$	-1
g	n_g/n_{2+}	n_g/n_{3+}	$n_g/n_{(g-1)+}$	$\cdots$	1

$n_{3+} = n_3 + n_4 = 173 + 101 = 274$, and $n_{4+} = 101$. So with $g = 4$ groups as in this example, the 4×3 matrix would be

$$\begin{matrix} -1 & 0 & 0 \\ n_2/n_{2+} & -1 & 0 \\ n_3/n_{2+} & n_3/n_{3+} & -1 \\ n_4/n_{2+} & n_4/n_{3+} & 1 \end{matrix}$$

or, in terms of the group sample sizes in the four age cohorts,

$$\begin{matrix} -1 & 0 & 0 \\ 149/423 & -1 & 0 \\ 173/423 & 173/274 & -1 \\ 101/423 & 101/274 & 1 \end{matrix}$$

Once this matrix is constructed, it is manipulated through matrix algebra to produce a $g \times (g-1)$ matrix that contains the $g-1$ sets of weighted Helmert codes for the g groups, where rows correspond to groups and columns are the codes D_1, D_2, and so forth. In this example, the resulting matrix is

$$\begin{matrix} -.7500000000 & -.0141843972 & -.0656934307 \\ .2500000000 & -.6619385343 & -.0656934307 \\ .2500000000 & .3380614657 & -.4343065693 \\ .2500000000 & .3380614657 & .5656934307 \end{matrix}$$

which are the codes for D_1, D_2, and D_3 found in Table 10.10. Regressing willingness to self-censor on D_1, D_2, and D_3 using these weighted Helmert codes yields the following model:

$$\hat{Y} = 2.993 - 0.268D_1 - 0.274D_2 - 0.055D_3$$

(see Table 10.4) with the same R as when any other coding system is used, as well as the same F- and p-values for testing the null that ${}_TR = 0$. And the model generates $\hat{Y}$ values that correspond to the group means.

So mathematically, this model is no different than any other model of the groups means we have constructed so far, in that it generates the same estimates of Y and fits exactly the same. But now the regression coefficient for D_j quantifies the difference between $\overline{Y}_j$ and the weighted mean of the means of Y for all groups coded higher than j on the variable quantifying the groups:

$$b_1 = \left(\frac{149\overline{Y}_2}{423} + \frac{173\overline{Y}_3}{423} + \frac{101\overline{Y}_4}{423}\right) - \overline{Y}_1 = 2.933 - 3.201 = -0.268$$

$$b_2 = \left(\frac{173\overline{Y}_3}{274} + \frac{101\overline{Y}_4}{274}\right) - \overline{Y}_2 = 2.837 - 3.111 = -0.274$$

$$b_3 = \overline{Y}_4 - \overline{Y}_3 = 2.802 - 2.857 = -0.055$$

The t-statistic and p-value for each regression coefficient tests the null hypothesis that the difference between the corresponding true means is equal to zero. As can be seen comparing the results when using unweighted to weighted Helmert coding (Tables 10.4 and 10.11), the results are very similar in this case, although this won't always be true. Generation Y self-censors less on average than those older, and Generation X self-censors on average more than those older, but baby boomers self-censor no more on average than pre-baby boomers.

The matrix computations are very tedious to do by hand. Fortunately, many good statistics programs have built-in features to do matrix computations. The SPSS code below takes the input matrix, implements the matrix algebra, and outputs the matrix of weighted Helmert codes. You can then construct D_1, D_2, and D_3 using **`if`** and **`compute`** commands. See the *Syntax Reference Manual* for guidance or consult a local expert.

```
matrix.
compute m={-1.0000, 0.0000, 0.0000;
```

```
           149/423,-1.0000, 0.0000;
           173/423,173/274,-1.0000;
           101/423,101/274, 1.0000}.
compute d=m*inv(t(m)*m).
print d.
end matrix.
```

In SAS, matrix operations can be conducted using PROC IML, which is an optional package. Check your installation. The comparable code in SAS is

```
proc iml;
m={-1.000000000 0.000000000 0.000000000,
     0.352245862 -1.000000000 0.000000000,
     0.408983451 0.631386861 -1.000000000,
     0.238770685 0.368613138 1.000000000};
d=m*inv(m`*m);
print d;
quit;
```

In STATA, try

```
mata
m=(-1.00000,0.00000,0.00000\
    149/423,-1.0000,0.00000\
    173/423,173/274,-1.0000\
    101/423,101/274, 1.0000)
d=m*luinv(m'*m)
d
end
```

10.3.3 Weighted Contrasts

Use of weighted Helmert codes generates regression coefficients and tests of significance, some of which can be interpreted as *complex contrasts*, a term and method introduced in section 10.2.1. However, in that discussion, the contrast involved a comparison of unweighted means. For example, in that section, we compared average willingness to self-censor among

Generation Y and Generation X to average willingness to self-censor among baby boomers and pre-baby boomers:

$$\begin{aligned}\text{Contrast} &= \frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2} \\ &= 3.156 - 2.830\end{aligned}$$

Those two means being compared were unweighted, because the mean of Generations Y and X was constructed merely by taking the arithmetic average of $\overline{Y}_1$ and $\overline{Y}_2$, ignoring that there are many fewer Generation Y in the sample than Generation X. Similarly, the mean of the baby boomers and pre-baby boomers was was constructed as the mean of $\overline{Y}_3$ and $\overline{Y}_4$, ignoring differences in sample size.

Weighted versions of these two means of means would give weight to Generation X relative to Generation Y, and to baby boomers relative to pre-baby boomers, in proportion to differences in their sample sizes. So rather than 3.156 for the combination of Generations X and Y, their weighted mean would be

$$\frac{38\overline{Y}_1 + 149\overline{Y}_2}{187} = \frac{38(3.201) + 149(3.111)}{187} = 3.129$$

Notice that this is closer to the mean of Generation X than Generation Y, because Generation X contributes more data to the mean. Similarly, the weighted mean for the combination of baby boomers and pre-baby boomers would be

$$\frac{173\overline{Y}_3 + 101\overline{Y}_4}{274} = \frac{173(2.857) + 101(2.802)}{274} = 2.837$$

rather than 2.830, which is closer to the mean of baby boomers, because its sample size is larger than the pre-baby boomers.

Complex contrasts can be conducted that compare weighted means to each other by using the relative sample sizes of the groups, as in the example computations above. Define a *contrast grouping* as a set of groups being combined in a contrast. A contrast always involves two, and only two, contrast groupings. In this example, contrast grouping 1 is the group defined as Generation X and Generation Y, and contrast grouping 2 is the group defined as baby boomers and pre-baby boomers. Now define n_{group_1} as the sum of the sample sizes of the groups that define contrast grouping 1, and n_{group_2} as the sum of the sample sizes of the groups that define contrast grouping 2. So in this example, $n_{group_1} = 38 + 149 = 187$ and

$n_{group_2} = 173 + 101 = 274$. Finally, define λ_j as the ratio of group j's sample size to the sample size of its corresponding contrast grouping. In this case,

$$\lambda_1 = n_1/n_{group_1} = 38/187$$
$$\lambda_2 = n_2/n_{group_1} = 149/187$$
$$\lambda_3 = n_3/n_{group_2} = 173/274$$
$$\lambda_4 = n_4/n_{group_2} = 101/274$$

With the $g = 4$ values of λ calculated, a weighted contrast is constructed as

$$\text{Contrast} = \sum_{j=1}^{g} c_j \lambda_j \overline{Y}_j \tag{10.7}$$

and its standard error estimated as

$$SE(\text{contrast}) = \sqrt{MS_{residual} \sum_{j=1}^{g} \frac{\left(c_j \lambda_j\right)^2}{n_j}} \tag{10.8}$$

where c_j is the contrast coefficients for group j as defined in section 10.2.1. The ratio of the contrast to its standard error is distributed as $t(df_{residual})$, and a p-value can be constructed using the t-distribution for testing a null hypothesis about the contrast (e.g., that the two weighted means are equal, meaning their difference is zero).

In this example, $c_1 = c_2 = 0.5$ and $c_3 = c_4 = -0.5$. Application of equation 10.7 yields

$$\begin{aligned}\text{Contrast} &= 0.5(38/187)(3.201) + 0.5(149/187)(3.111) - \\ &\quad 0.5(173/274)(2.857) - 0.5(101/274)(2.802) \\ &= 0.146\end{aligned}$$

and equation 10.8 generates

$$\begin{aligned}SE(\text{contrast}) &= \sqrt{0.273\left[\frac{\left(0.5\frac{38}{187}\right)^2}{38} + \frac{\left(0.5\frac{149}{187}\right)^2}{149} + \frac{\left(-0.5\frac{173}{274}\right)^2}{173} + \frac{\left(-0.5\frac{101}{274}\right)^2}{101}\right]} \\ &= 0.025\end{aligned}$$

Their ratio is $t = 0.146/0.025 = 5.840$, which has an exceedingly tiny two-tailed p-value derived from the $t(457)$ distribution. The null hypothesis of equality of the weighted means is rejected.

Notice that in this example, the contrast calculated above is actually one half of the difference between the weighted means rather than the difference itself:

$$\begin{aligned}\text{Contrast} &= 0.5(38/187)\overline{Y}_1 + 0.5(149/187)\overline{Y}_2 - \\ &\quad 0.5(173/274)\overline{Y}_3 - 0.5(101/274)\overline{Y}_4 \\ &= 0.5\left[\left(\frac{38}{187}\overline{Y}_1 + \frac{149}{187}\overline{Y}_2\right) - \left(\frac{173}{274}\overline{Y}_3 + \frac{101}{274}\overline{Y}_4\right)\right] \\ &= 0.5(3.129 - 2.837) \\ &= 0.146\end{aligned}$$

However, so too is the estimated standard error one-half of the standard error of the difference between the weighted means, so the result of the inference is unaffected. If this bothers you, simply multiply both by two when reporting. This correction would be important if reporting a confidence interval for the difference between weighted means, because you would want the confidence interval to be in the metric of the difference, not one-half the difference.[1]

These computations can be done by most statistical programs that allow you to specify contrast coefficients in an ANOVA procedure, and these will be done more accurately than the hand computations illustrated above. In the unweighted contrast example from section 10.2.3, we put c_j in the computer code to produce the contrast. But now, we use $c_j\lambda_j$ for the contrast coefficients instead. So in SPSS, the code to conduct this contrast would be

```
oneway wtsc by cohort/contrast 0.101604 0.398396 -0.315693 -0.184307
   /statistics descriptive.
```

or in SAS, use

```
proc glm data=wtsc;
   class cohort;model wtsc=cohort;means cohort;
   contrast '1 2 vs 3 4' cohort 0.101604 0.398396 -0.315693 -0.184307;
run;
```

[1]It is not generally true that equation 10.7 will produce one-half the difference between weighted means. Whether or not equation 10.7 produces the weighted mean difference or some multiple of it will depend on the values of c_j used.

In this example, each value of $c_j\lambda_j$ input into the code could be multiplied by 2 to rescale the contrast to the mean difference metric rather than one-half the difference.

10.3.4 Application to Adjusted Means

Weighted effect and weighted Helmert coding will produce regression coefficients that correspond to differences between weighted adjusted means when covariates are included in the model. You may be tempted to do complex weighted contrasts between adjusted means using the procedure described in section 10.3.3, substituting adjusted means for $\overline{Y}_j$. But equation 10.8 does not produce a proper estimate of the standard error of a contrast between weighted adjusted means. The computer-assisted procedures described in section 10.2.3 can be used instead, so long as the contrast coefficients fed to the computer algorithm are multiplied by the appropriate weights (i.e., use $c_j\lambda_j$ rather than c_j) to produce the contrast of interest.

10.4 Chapter Summary

In this chapter we introduced and illustrated several ways of coding a multicategorical variable so that it can be used as a regressor in a regression model. These methods, including sequential coding, Helmert coding, and effect coding, yield models that are mathematically equivalent to the model generated when indicator coding is used. Of these methods, sequential and Helmert coding are particularly useful when the multicategorical variable represents an ordinal dimension. But regardless of the method of coding used, the choice one makes about how to code groups does not affect the fit of the model or the estimates of Y is produces. Furthermore, the choice does not affect the test of the null hypothesis that the g groups don't differ on average on Y, and using regression analysis results in the same inference produced by ANOVA and ANCOVA. However, the method of coding groups will change the regression constant and the regression coefficients and how they are interpreted.

Complex contrasts between means is a staple topic in analysis of variance books, but it is still appropriate in a regression analysis book such as this because ANOVA is just a special case of linear regression analysis, and contrasts can be conducted using output from a regression analysis. Another topic commonly introduced in the context of ANOVA is the *multiple test problem*—the positive correlation between the number of tests conducted and the probability of making at least one Type I error. In the

next chapter we address the multiple test problem and its relevance not only to comparing groups but also to regression analysis more generally.