5.3.2 Categorical Independent Variables

Indicator Dummy Variables

Independent variables need not be continuous – or even quantitative. For example, suppose subjects in a drug study are randomly assigned to either an active drug or a placebo. Let Y represent response to the drug, and

$$x = \begin{cases} 1 & \text{if the subject received the active drug, or} \\ 0 & \text{if the subject received the placebo.} \end{cases}$$

The model is $E[Y|X = x] = \beta_0 + \beta_1 x$. For subjects who receive the active drug (so $x = 1$), the population mean is

$$\beta_0 + \beta_1 x = \beta_0 + \beta_1$$

For subjects who receive the placebo (so $x = 0$), the population mean is

$$\beta_0 + \beta_1 x = \beta_0.$$

Therefore, β_0 is the population mean response to the placebo, and β_1 is the difference between response to the active drug and response to the placebo. We are very interested in testing whether β_1 is different from zero, and guess what? We get exactly the same t value as from a two-sample t-test, and exactly the same F value as from a one-way ANOVA for two groups.

Exercise Suppose a study has 3 treatment conditions. For example Group 1 gets Drug 1, Group 2 gets Drug 2, and Group 3 gets a placebo, so that the Independent Variable is Group (taking values 1,2,3) and there is some Dependent Variable Y (maybe response to drug again).

Sample Question 5.3.1 *Why is $E[Y|X = x] = \beta_0 + \beta_1 x$ (with x = Group) a silly model?*

Answer to Sample Question 5.3.1 *Designation of the Groups as 1, 2 and 3 is completely arbitrary.*

Sample Question 5.3.2 *Suppose $x_1 = 1$ if the subject is in Group 1, and zero otherwise, and $x_2 = 1$ if the subject is in Group 2, and zero otherwise, and $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Fill in the table below.*

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
1			$\mu_1 =$
2			$\mu_2 =$
3			$\mu_3 =$

Answer to Sample Question 5.3.2

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
1	1	0	$\mu_1 = \beta_0 + \beta_1$
2	0	1	$\mu_2 = \beta_0 + \beta_2$
3	0	0	$\mu_3 = \beta_0$

Sample Question 5.3.3 *What does each β value mean?*

Answer to Sample Question 5.3.3 *$\beta_0 = \mu_3$, the population mean response to the placebo. β_1 is the difference between mean response to Drug 1 and mean response to the placebo. β_2 is the difference between mean response to Drug 21 and mean response to the placebo.*

Sample Question 5.3.4 *Why would it be nice to simultaneously test whether β_1 and β_2 are different from zero?*

Answer to Sample Question 5.3.4 *This is the same as testing whether all three population means are equal; this is what a one-way ANOVA does. And we get the same F and p values (not really part of the sample answer).*

Notice that x_1 and x_2 contain the same information as the three-category variable Group. If you know Group, you know x_1 and x_2, and if you know x_1 and x_2, you know Group. In models with an intercept term, a categorical independent variable with k categories is always represented by $k-1$ dummy variables. If the dummy variables are indicators, the category that does not get an indicator is actually the most important. The intercept is that category's mean, and it is called the **reference category**, because the remaining regression coefficients represent differences between the reference category and the other category. To compare several treatments to a control, make the control group the reference category by *not* giving it an indicator.

It is worth noting that all the traditional one-way and higher-way models for analysis of variance and covariance emerge as special cases of multiple regression, with dummy variables representing the categorical independent variables.

Add a quantitative independent variable

Now suppose we include patient's age in the regression model. When there are both quantitative and categorical independent variables, the quantitative variables are often called *covariates*, particularly if the categorical part is experimentally manipulated. Tests of the categorical variables controlling for the quantitative variables are called *analysis of covariance*.

The usual practice is to put the covariates first. So, we'll let X_1 represent age, and let X_2 and X_3 be the indicator dummy variables for experimental condition. The model now is that all conditional distributions are normal with the same variance σ^2, and population mean

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Sample Question 5.3.5 *Fill in the table.*

Group	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Answer to Sample Question 5.3.5

Group	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$\mu_2 = (\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\mu_3 = \quad \beta_0 \quad + \beta_1 x_1$

This is a *parallel slopes model*. That is, there is a least-squares regression line for each group, with the same slope β_1 for each line. Only the intercepts are different. This means that for any fixed value of x_1 (age), the differences among population means are the same. For any value of age (that is, holding age constant, or *controlling* for age), the difference between response to Drug A and the placebo is β_2. And controlling for age), the difference between response to Drug B and the placebo is β_3. The three group means are equal for each constant value of age if (and only if) $\beta_2 = \beta_3 = 0$. This is the null hypothesis for the analysis of covariance.

It is easy (and often very useful) to have more than one covariate. In this case we have parallel planes or hyper-planes. And at any fixed set of covariate values, the distances among hyperplanes correspond exactly to the differences among the intercepts. This means we are usually interested in testing null hypotheses about the regression coefficients corresponding to the dummy variables.

Sample Question 5.3.6 *Suppose we want to test the difference between response to Drug A and Drug B, controlling for age. What is the null hypothesis?*

Answer to Sample Question 5.3.6 $H_0 : \beta_2 = \beta_3$

Sample Question 5.3.7 *Suppose we want to test whether controlling for age, the average response to Drug A and Drug B is different from response to the placebo. What is the null hypothesis?*

Answer to Sample Question 5.3.7 $H_0 : \beta_2 + \beta_3 = 0$

Sample Question 5.3.8 *Huh? Show your work.*

Answer to Sample Question 5.3.8

$$\frac{1}{2}[\,(\beta_0+\beta_2+\beta_1x_1)+(\beta_0+\beta_3+\beta_1x_1)\,]=\beta_0+\beta_1x_1$$

$$\iff \quad \beta_0+\beta_2+\beta_1x_1+\beta_0+\beta_3+\beta_1x_1=2\beta_0+2\beta_1x_1$$

$$\iff \quad 2\beta_0+\beta_2+\beta_3+2\beta_1x_1=2\beta_0+2\beta_1x_1$$

$$\iff \quad \beta_2+\beta_3=0$$

The symbol $\iff$ means "if and only if." The arrows can logically be followed in both directions.

This last example illustrates several important points.

- Contrasts can be tested with indicator dummy variables.
- If there are covariates, the ability to test contrasts *controlling* for the covariates is very valuable.
- Sometimes, the null hypothesis for a contrast of interest might not be what you expect, and you might have to derive it algebraically. This can be inconvenient, and it is too easy to make mistakes.

Cell means coding

When students are setting up dummy variables for a categorical independent variable with p categories, the most common mistake is to define an indicator dummy variable for every category, resulting in p dummy variables rather than $p-1$ — and of course there is an intercept too, because it's a regression model and regression software almost always includes an intercept unless you explicitly suppress it. But then the p population means are represented by $p+1$ regression coefficients, and mathematically, the representation cannot be unique. In this situation the least-squares estimators are not unique either, and all sorts of technical problems arise. Your software might try to save you by throwing one of the dummy variables out, but which one would it discard? And would you notice that it was missing from your output?

Suppose, however, that you used p dummy variables but *no intercept* in the regression model. Then there are p regression coefficients corresponding to the p population means, and all the technical problems go away. The correspondence between regression coefficients and population means is unique, and the model can be handy. In particular, null hypotheses can often be written down immediately without any high school algebra. Here is how it would look for the study with two drugs and a placebo. The conditional population means is

$$E[Y|\boldsymbol{X}=\boldsymbol{x}]=\beta_1x_1+\beta_2x_2+\beta_3x_3,$$

and the table of population means has a very simple form:

Drug	x_1	x_2	x_3	$\beta_1x_1+\beta_2x_2+\beta_3x_3$
A	1	0	0	$\mu_1=\beta_1$
B	0	1	0	$\mu_2=\beta_2$
Placebo	0	0	1	$\mu_3=\beta_3$

The regression coefficients correspond directly to population (cell) means for any number of categories; this is why it's called *cell means coding*. Contrasts are equally easy to write in terms of μ or β quantities.

Cell means coding works nicely in conjunction with quantitative covariates. In the drug study example, represent age by X_4. Now the conditional population mean is

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,$$

and the cell means (for any fixed value of age equal to x_4) are

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	$\beta_1 + \beta_4 x_4$
B	0	1	0	$\beta_2 + \beta_4 x_4$
Placebo	0	0	1	$\beta_3 + \beta_4 x_4$

This is another parallel slopes model, completely equivalent to the earlier one. The regression coefficients for the dummy variables are the intercepts, and because the lines are parallel, the differences among population means at any fixed value of x_4 are exactly the differences among intercepts. Note that

- It is easy to write the null hypothesis for any contrast of collection of contrasts. Little or no algebra is required.
- This extends to categorical independent variables with any number of categories.
- With more than one covariate, we have a parallel planes model, and it is still easy to express the null hypotheses.
- The `test` statement of `proc reg` is a particularly handy tool.

Effect Coding

In *effect coding* there are $p - 1$ dummy variables for a categorical independent variable with p categories, and the intercept is included. Effect coding look just like indicator dummy variable coding with an intercept, except that the last (reference) category gets -1 instead of zero. Here's how it looks for the hypothetical drug study.

Group	x_1	x_2	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

To see what the regression coefficients mean, first define μ to be the average of the three population means. Then

$$\mu = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \beta_0,$$

so that the intercept is the mean of population means — sometimes called the *grand mean*. Now we can see right away that

- β_1 is the difference between μ_1 and the grand mean.
- β_2 is the difference between μ_2 and the grand mean.
- $-\beta_1 - \beta_2$ is the difference between μ_3 and the grand mean.
- Equal population means is equivalent to zero coefficients for all the dummy variables.
- The last category is not a reference category. It's just the category with the least convenient expression for the deviation from the grand mean.
- This pattern holds for any number of categories.

In the standard language of analysis of variance, *effects* are deviations from the grand mean. That's why this dummy variable coding scheme is called "effect coding." When there is more than one categorical independent variable, the average cell mean for a particular category (averaging across other independent variables) is called a *marginal mean*, and the so-called *main effects* are deviations of the marginal means from the grand mean; these are represented nicely by effect coding. Equality of marginal means implies that all main effects for the variable are zero, and vice versa.

Sometimes, people speak of testing for the "main effect" of a categorical independent variable. This is a loose way of talking, because there is not just one main effect for a variable. There are at least two, one for each marginal mean. Possibly, this use of "effect" blends the effect of an experimental variable with the technical statistical meaning of effect. However, it's a way of talking that does no real harm, and you may see it from time to time in this text.

We will see later that effect coding is very useful when there is more than one categorical independent variable and we are interested in *interactions* — ways in which the relationship of an independent variable with the dependent variable depends on the value of another independent variable.

Covariates work nicely with effect coding. There is no need to make a table of expected values, unless a question explicitly asks you to do so. For example, suppose you add the covariate X_1 = Age to the drug study. The treatment means (which depend on X_1 are as follows:

Group	x_2	x_3	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = \beta_0 + \beta_2 \quad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \quad + \beta_1 x_1$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_2 - \beta_3 + \beta_1 x_1$

Regression coefficients are deviations from the average conditional population mean (conditional on x_1). So, if the regression coefficients for all the dummy variables equal zero, the categorical independent variable is unrelated to the dependent variable, when one controls for the covariates.

Finally, it's natural for a student to wonder: What dummy variable coding scheme should I use? Use wWhichever is most convenient. They are all equivalent, if done correctly. They yield the same test statistics, and the same conclusions.