More about Dummy Variables

- Indicator dummy variables with intercept
- Indicator dummy variables without intercept (Cell means coding)
- Effect coding

Drug A, Drug B, Placebo

- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

Now add a quantitative variable (covariate)

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

Parallel slopes, ANCOVA

Can test contrasts *controlling* for covariates

- Valuable
- Sometimes very easy, sometimes can require a bit of algebra
- An easy example: Are responses to Drug A and B different, controlling for age?

Are responses to Drug A and B different, controlling for age?

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

$$H_0 : \beta_2 = \beta_3$$

Test whether the average response to Drug A and Drug B is different from response to the placebo, controlling for age. What is the null hypothesis?

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

$$H_0 : \beta_2 + \beta_3 = 0$$

Show your work

$$\tfrac{1}{2}[\,(\beta_0 + \beta_2 + \beta_1 x_1) + (\beta_0 + \beta_3 + \beta_1 x_1)\,] = \beta_0 + \beta_1 x_1$$

$$\iff \beta_0 + \beta_2 + \beta_1 x_1 + \beta_0 + \beta_3 + \beta_1 x_1 = 2\beta_0 + 2\beta_1 x_1$$

$$\iff 2\beta_0 + \beta_2 + \beta_3 + 2\beta_1 x_1 = 2\beta_0 + 2\beta_1 x_1$$

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

We want to avoid this kind of thing

A common error

- Categorical IV with *p* categories
- *p* dummy variables (rather than *p-1*)
- And an intercept

- There are *p* population means represented by *p+1* regression coefficients - not unique

But suppose you leave off the intercept

- Now there are *p* regression coefficients and *p* population means
- The correspondence is unique, and the model can be handy -- less algebra
- Called **cell means coding**

Cell means coding: *p* indicators and no intercept

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

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

Add a covariate: x_4

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

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	$\beta_3 + \beta_4 x_4$

Effect coding

- *p-1* dummy variables for *p* categories
- Include an intercept
- Last category gets -1 instead of zero
- What do the regression coefficients mean?

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	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Meaning of the regression coefficients

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1x_1+\beta_2x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

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

With effect coding

- Intercept is the *Grand Mean*
- Regression coefficients are deviations of group means from the grand mean
- Equal population means is equivalent to zero coefficients for all the dummy variables
- Last category is not a reference category

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1x_1+\beta_2x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

Sometimes speak of the “main effect” of a categorical variable

- More than one categorical IV (factor)
- Marginal means are average group mean, averaging across the other factors
- This is loose speech: There are actually *p* main effects for a variable, not one
- Blends the “effect” of an experimental variable with the technical statistical meaning of effect.
- It’s harmless

Add a covariate: Age = x_1

Group	x_2	x_3	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$
A	1	0	$\mu_1=\beta_0+\beta_2\quad+\beta_1x_1$
B	0	1	$\mu_2=\beta_0+\beta_3\quad+\beta_1x_1$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_2-\beta_3+\beta_1x_1$

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 IV is unrelated to the DV, controlling for the covariates.

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.

What dummy variable coding scheme should you use?

- Whichever is most convenient
- They are all equivalent, if done correctly
- Same test statistics, same conclusions