Categorical Independent Variables

STA302 Fall 2015

See last slide for copyright information

Categorical means *unordered* categories

- Like Field of Study: Humanities, Sciences, Social Sciences
- Could number them 1 2 3, but what would the regression coefficients mean?
- But you really want them in your regression model.

One Categorical Explanatory Variable

- X=1 means Drug, X=0 means Placebo
- Population mean is $E[Y|X = x] = \beta_0 + \beta_1 x$
- For patients getting the drug, mean response is
$$E[Y|X = 1] = \beta_0 + \beta_1$$
- For patients getting the placebo, mean response is
$$E[Y|X = 0] = \beta_0$$

Sample regression coefficients for a binary explanatory variable

- X=1 means Drug, X=0 means Placebo
- Predicted response is $$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$
- For patients getting the drug, predicted response is
$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 = \overline{Y}_1$$
- For patients getting the placebo, predicted response is
$$\widehat{Y} = \widehat{\beta}_0 = \overline{Y}_0$$

Regression test of $H_0 : \beta_1 = 0$

- Same as an independent t-test
- Same as a oneway ANOVA with 2 categories
- Same t, same F, same p-value.

- Now extend to more than 2 categories

Drug A, Drug B, Placebo

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

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

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	$\mu_3 = \beta_0$

Regression coefficients are *contrasts* with the category that has no indicator – the *reference* category

Indicator dummy variable coding with intercept

- Need p-1 indicators to represent a categorical explanatory variable with p categories.
- If you use p dummy variables, columns of the **X** matrix are linearly dependent.
- Regression coefficients are *contrasts* with the category that has no indicator.
- Call this the *reference category.*

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

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 regression lines

A common error

- Categorical explanatory variable 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	$\mu_3 = \beta_3$

This model is equivalent to the one with the intercepts

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	$\beta_3 + \beta_4 x_4$

Do the residuals add to zero with cell means coding?

- If so, SST = SSR + SSE
- And we have R^2

- Let **j** denote an nx1 column of ones.
- If there is a (k+1)x1 vector **a** with **Xa=j**, the residuals add up to zero.
- So the answer is Yes.

Copyright Information