Categorical Independent Variables

STA302 Fall 2014

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

Covariates

- Of course there could be more than one
- Reduce MSE, make tests more sensitive
- If values of categorical IV are not randomly assigned, including relevant covariates could change the conclusions.

Interactions

- Interaction between independent variables means "It depends."
- Relationship between one explanatory variable and the response variable *depends* on the value of the other explanatory variable.
- Can have
  - Quantitative by quantitative
  - Quantitative by categorical
  - Categorical by categorical

Quantitative by Quantitative

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\ E(Y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \end{aligned}$$

For fixed x_2

$$E(Y|\mathbf{x}) = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

Both slope and intercept depend on value of x_2

And for fixed x_1, slope and intercept relating x_2 to E(Y) depend on the value of x_1

Quantitative by Categorical

- One regression line for each category.
- Interaction means slopes are not equal
- Form a product of quantitative variable by each dummy variable for the categorical variable
- For example, three treatments and one covariate: x_1 is the covariate and x_2, x_3 are dummy variables

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &\quad + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon \end{aligned}$$

Make a table

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

What null hypothesis would you test for

- Equal slopes
- Comparing slopes for group one vs three
- Comparing slopes for group one vs two
- Equal regressions
- Interaction between group and x_1

General principle

- Interaction between A and B means
  - Relationship of A to Y depends on value of B
  - Relationship of B to Y depends on value of A
- The two statements are formally equivalent

What to do if H_0: $\beta_4=\beta_5=0$ is rejected

- How do you test Group "controlling" for x_1?
- A reasonable choice is to set x_1 to its sample mean, and compare treatments at that point.

- How about setting x_1 to sample mean of the group (3 different values)?
- With random assignment to Group, all three means just estimate $E(X_1)$, and the mean of all the x_1 values is a better estimate.

Copyright Information