Categorical Independent Variables[1]

STA 302 Fall 2017

[1] See last slide for copyright information.

Overview

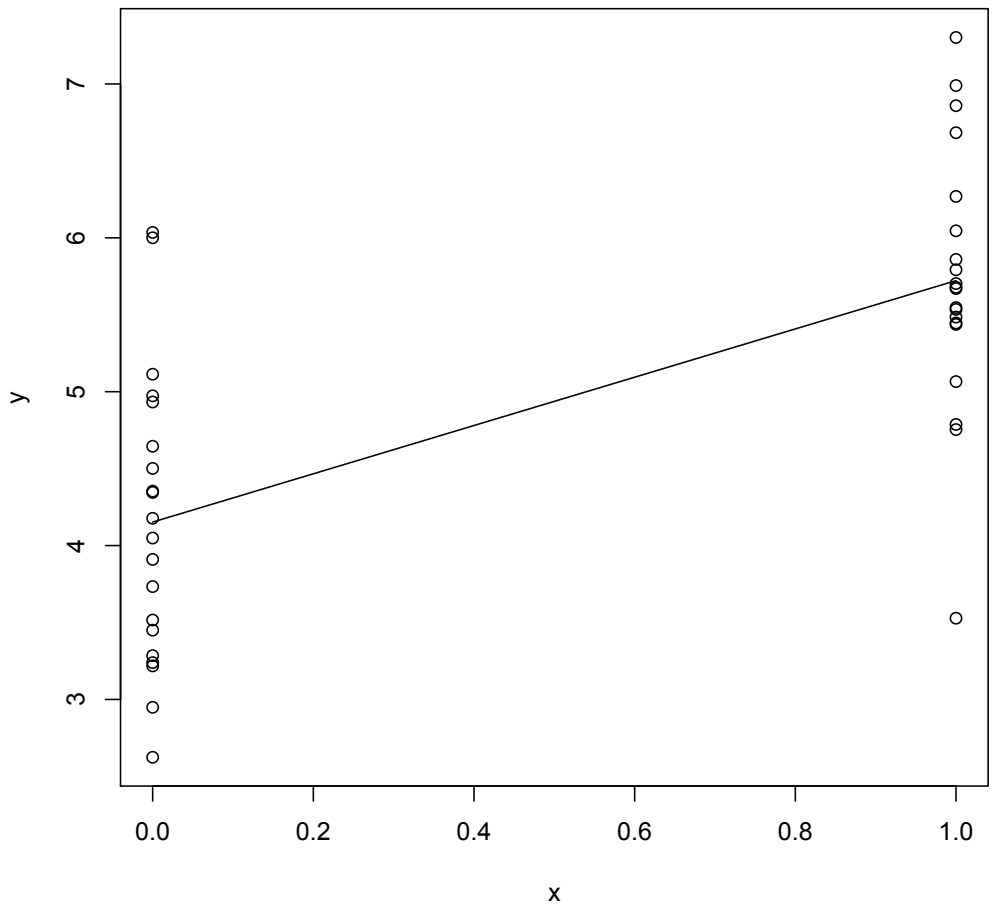
Independent variables need not be continuous

Code data so that $x = 1$ means Drug, $x = 0$ means Placebo.

- Population mean response is $E(y|x) = \beta_0 + \beta_1 x$.
- For patients getting the drug, mean response is $E(y|x = 1) = \beta_0 + \beta_1$.
- For patients getting the placebo, mean response is $E(y|x = 0) = \beta_0$.
- Difference (treatment effect) is β_1.
- Test $H_0 : \beta_1 = 0$.

Scatterplot

Showing the least-squares line



Predicted response is $\widehat{y} = b_0 + b_1 x$.

For patients getting the drug, predicted response is $\widehat{y} = b_0 + b_1 = \overline{y}_1$.

For patients getting the placebo, predicted response is $\widehat{y} = b_0 = \overline{y}_0$.

More than Two Categories

Suppose a study has 3 treatment conditions. For example

- Group 1 gets Drug 1
- Group 2 gets Drug 2
- Group 3 gets a placebo
- So that the Explanatory Variable is Group
- Taking values 1,2,3.
- The dependent variable y is response to drug.

Why is $E(y|x) = \beta_0 + \beta_1 x$ (with $x =$ Group) a silly model?

Indicator Dummy Variables

With intercept

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
- Fill in the table.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Answer

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

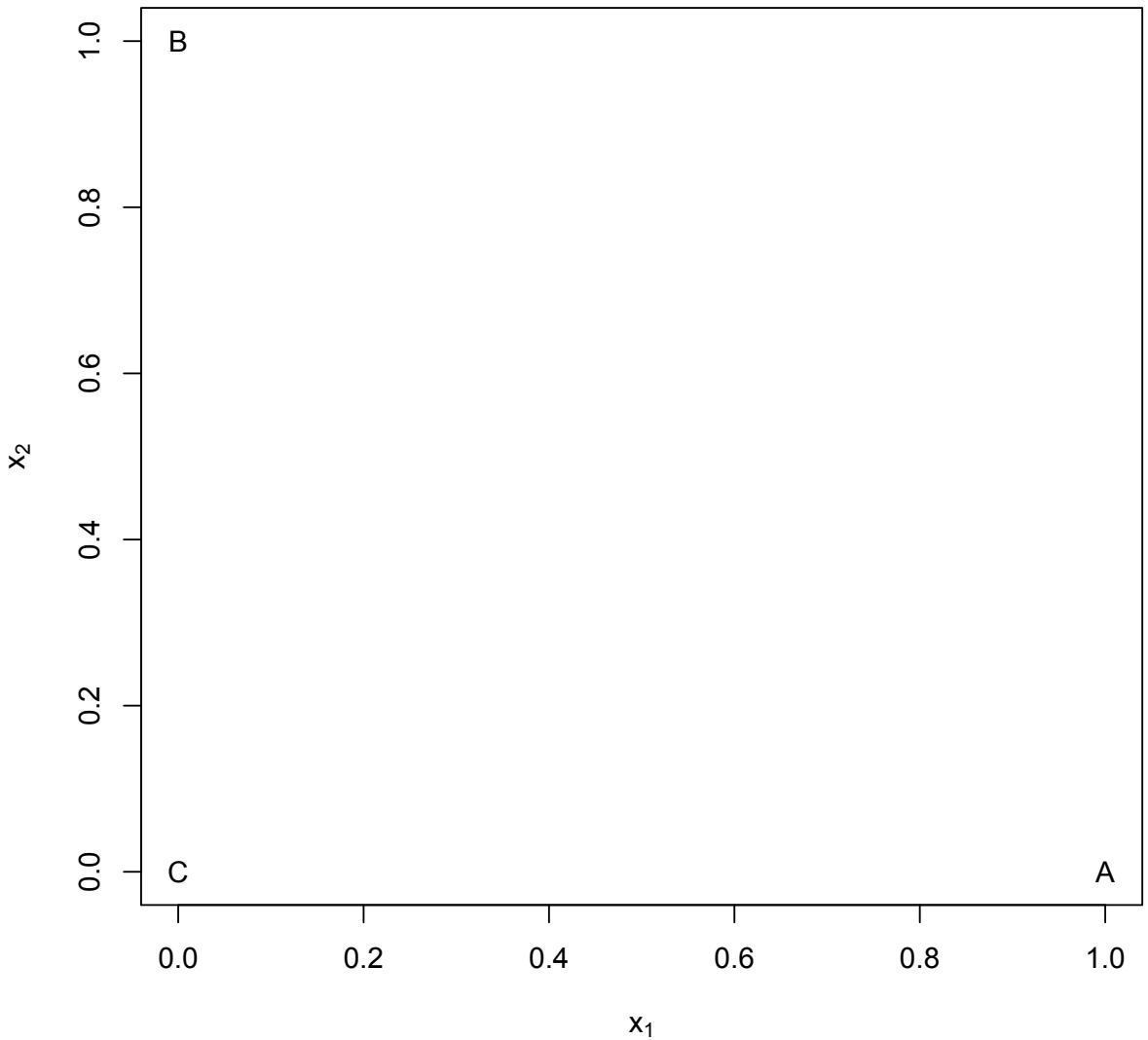
Regression coefficients are contrasts with the category that has no indicator – the *reference category*.

Indicator dummy variable coding with intercept

- With an intercept in the model, need $p-1$ indicators to represent a categorical explanatory variable with p categories.
- If you use p dummy variables and also an intercept, trouble.
- Indicators would add up to the intercept and columns of X would be linearly dependent.
- Regression coefficients are contrasts with the category that has no indicator.
- Call this the *reference category*.

$x_1 = 1$ if Drug A, zero o.w., $x_2 = 1$ if Drug B, zero o.w.

Recall $\sum_{i=1}^{n}(y_i - m)^2$ is minimized at $m = \overline{y}$



What null hypotheses would you test?

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

- Is the effect of Drug A different from the placebo? $H_0 : \beta_1 = 0$
- Is Drug A better than the placebo? $H_0 : \beta_1 = 0$
- Did Drug B work? $H_0 : \beta_2 = 0$
- Did experimental treatment have an effect? $H_0 : \beta_1 = \beta_2 = 0$
- Is there a difference between the effects of Drug A and Drug B? $H_0 : \beta_1 = \beta_2$

Now add a quantitative explanatory variable (covariate)

Covariates often come first in the regression equation

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $x_3 =$ Age
- $E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_1) + \beta_3 x_3$
B	0	1	$\mu_2 = (\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	0	0	$\mu_3 = \beta_0 + \beta_3 x_3$

Parallel regression lines.

More comments

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_1) + \beta_3 x_3$
B	0	1	$\mu_2 = (\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	0	0	$\mu_3 = \beta_0 + \beta_3 x_3$

- If more than one covariate, parallel regression planes.
- Non-parallel (interaction) is testable.
- "Controlling" interpretation holds.
- In an experimental study, quantitative covariates are usually just observed.
- Could age be related to drug?
- Good covariates reduce $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$, and make tests involving the categorical variables more sensitive.

Cell means coding: p indicators and no intercept

Example: Three treatments and no covariate.

$$E(y|\boldsymbol{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Drug	x_1	x_2	x_3	$E(y\|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
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 $p-1$ dummy variables and the intercept.
- If you have p dummy variables and the intercept, the model is over-parameterized.

Add a covariate: x_4

$$E(y|\boldsymbol{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Drug	x_1	x_2	x_3	$E(Y\|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
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 model is equivalent to the one with the intercept.

Key to the equivalence of dummy variable coding schemes

Clearly these X matrices are one-to-one.

$$\begin{pmatrix} 1 & 1 & 0 & x_1 \\ 1 & 0 & 1 & x_2 \\ 1 & 0 & 0 & x_3 \\ 1 & 1 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_n \end{pmatrix} \leftrightarrow \begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \\ 1 & 0 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n \end{pmatrix}$$

And it's a linear transformation.

Matrix multiplication

$$\begin{pmatrix} 1 & 1 & 0 & x_1 \\ 1 & 0 & 1 & x_2 \\ 1 & 0 & 0 & x_3 \\ 1 & 1 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_n \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \\ 1 & 0 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n \end{pmatrix}$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\Leftrightarrow \quad \mathbf{y} = (XA)(A^{-1}\boldsymbol{\beta}) + \boldsymbol{\epsilon}$$

Transformed X requires a transformed $\boldsymbol{\beta}$.

Copyright Information

This slide show was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website:
http://www.utstat.toronto.edu/~brunner/oldclass/302f17