Categorical Predictor Variables[1]

STA 302 Fall 2020

[1] See last slide for copyright information.

Overview

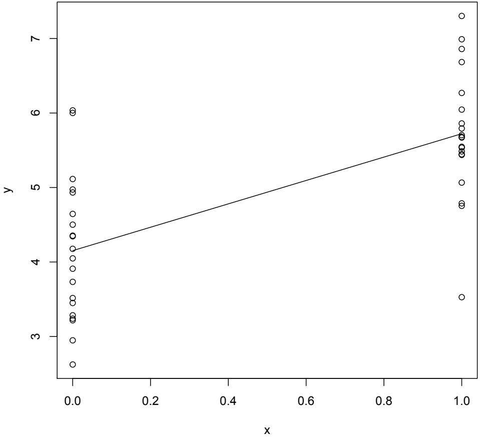
Predictor variables need not be continuous

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 $H_0 : \beta_1 = 0$.
- Same as the traditional 2-sample test.

Scatterplot

Showing the least-squares line



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 $\widehat{y} = \widehat{\beta}_0 = \overline{y}_0$.

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 Treatment
- 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 = Treatment) 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*.

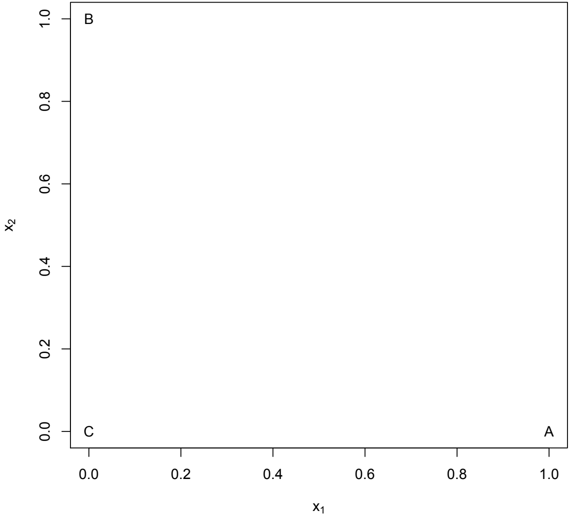
Indicator dummy variable coding with intercept

- With an intercept in the model, need $r-1$ indicators to represent a categorical explanatory variable with r categories.
- If you use r dummy variables and also an intercept, trouble.
- Indicators would add up to the intercept and columns of $\mathbf{X}$ 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.

$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$

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

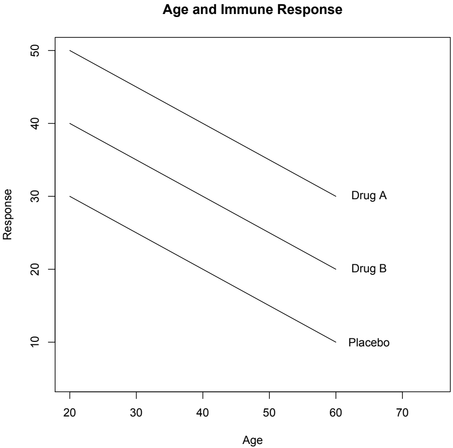
- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $x_3 =$ Age
- $E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

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	$\mu_1 =$
B	0	1	$\mu_2 =$
Placebo	0	0	$\mu_3 =$

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	$\mu_3 = \beta_0 + \beta_3 x_3$

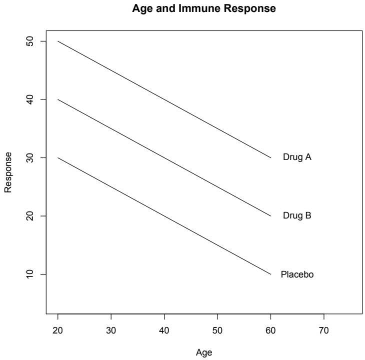
Parallel Regression Lines

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	$\mu_3 = \quad \beta_0 \quad + \beta_3 x_3$



Parallel Regression Lines

Drug	x_1	x_2	$E(y\|\mathbf{x})$
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	$\mu_3 = \beta_0 + \beta_3 x_3$



For fixed age, is there a difference in expected immune response as a function of experimental treatment? $H_0 : \beta_1 = \beta_2 = 0$.

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 $MSE = \frac{\widehat{\epsilon}^{\prime}\widehat{\epsilon}}{n-k-1}$, and make tests involving the categorical variables more sensitive.

Cell means coding: r 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 $r-1$ dummy variables and the intercept.
- If you have r dummy variables and also 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	$E(y\vert\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
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.

Which one should you use?

Choose on the basis of convenience

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	$\mu_3 = \beta_0 + \beta_3 x_3$

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	$\beta_3 + \beta_4 x_4$

- Test whether the average response to Drug A is different from the average response to Drug B, controlling for age. What is the null hypothesis? $H_0 : \beta_1 = \beta_2$.
- 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 for the model with intercept? $H_0 : \beta_2 + \beta_3 = 0$.

Huh?

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	$\mu_3 = \beta_0 + \beta_3 x_3$

Controlling for age, is the average response to Drug A and Drug B different from mean response to the placebo? What is the null hypothesis? $H_0 : \beta_2 + \beta_3 = 0$. Really? Show your work.

$$
\begin{aligned}
& \tfrac{1}{2}\left[\,(\beta_0 + \beta_2 + \beta_1 x_1) + (\beta_0 + \beta_3 + \beta_1 x_1)\,\right] = \beta_0 + \beta_1 x_1 \\
\iff \quad & \beta_0 + \beta_2 + \beta_1 x_1 + \beta_0 + \beta_3 + \beta_1 x_1 = 2\beta_0 + 2\beta_1 x_1 \\
\iff \quad & 2\beta_0 + \beta_2 + \beta_3 + 2\beta_1 x_1 = 2\beta_0 + 2\beta_1 x_1 \\
\iff \quad & \beta_2 + \beta_3 = 0.
\end{aligned}
$$

We want to avoid this kind of thing.

Easier with Cell Means Coding

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	$\beta_3 + \beta_4 x_4$

Controlling for age, is the average response to Drug A and Drug B different from mean response to the placebo? What is the null hypothesis?

$H_0 : \frac{1}{2}(\beta_1 + \beta_2) = \beta_3$, or $H_0 : \beta_1 + \beta_2 = 2\beta_3$.

Key to the equivalence of dummy variable coding schemes

Clearly these $\mathbf{X}$ matrices are one-to-one.

$$\left(\begin{array}{cccc} 1 & 1 & 0 & x_1 \\ 1 & 0 & 1 & x_2 \\ 1 & 0 & 0 & x_3 \\ 1 & 1 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_n \end{array}\right) \leftrightarrow \left(\begin{array}{cccc} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \\ 1 & 0 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n \end{array}\right)$$

And it's a linear transformation.

Matrix multiplication

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

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

Transformed $\mathbf{X}$ implies a transformed $\boldsymbol{\beta}$.

Other 1-1 linear transformations of the predictor variables can be useful

- x_1 = Verbal SAT, x_2 = Math SAT, y = First year GPA.
- $w_1 = x_1 + x_2$ is total SAT score.
- $w_2 = x_2 - x_1$ is how much better the student did in the math part.
- You might prefer $y_i = \beta_0 + \beta_1 w_{i,1} + \beta_2 w_{i,2} + \epsilon_i$.
- (w_1, w_2) is one-to-one with (x_1, x_2).
- $\mathbf{y} = (\mathbf{XA})(\mathbf{A}^{-1}\boldsymbol{\beta}) + \boldsymbol{\epsilon}$.

Interactions

- Interaction between predictor variables means "It depends."
- Relationship between one explanatory variable and the response variable *depends* on the value of another explanatory variable
- Note that an interaction is *not* a relationship between explanatory variables (in this course).

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.

Interactions between explanatory variables can be

- Quantitative by quantitative
- Quantitative by categorical
- Categorical by categorical

Quantitative by Quantitative

Represent the interaction by a *product* of explanatory variables.

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

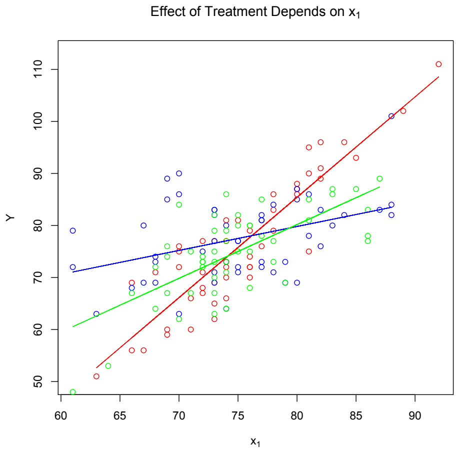
For fixed x_2,

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

- 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.
- This interpretation holds only with x_1 and x_2 (separately) in the model!

Quantitative by Categorical

- Separate regression line for each value of the categorical explanatory variable.
- Interaction means slopes of regression lines are not equal.



A Single Regression Model

- Form a product of quantitative variable times each dummy variable for the categorical variable.
- For example, three treatments and one covariate: x_1 is the covariate, and x_2 and x_3 are the dummy variables.

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

- Keep x_1, x_2 and x_3 (separately) in the model.

Fill in the table

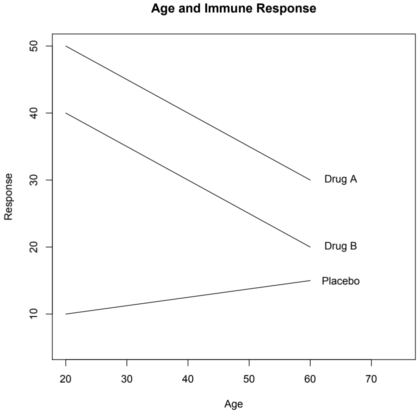
$$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Treatment	x_2	x_3	$E(y\|\mathbf{x})$
Drug A	1	0	
Drug B	0	1	
Placebo	0	0	

Treatment	x_2	x_3	$E(y\|\mathbf{x})$
Drug A	1	0	
Drug B	0	1	
Placebo	0	0	

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

Treatment	x_2	x_3	$E(y\|\mathbf{x})$
Drug A	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
Drug B	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
Placebo	0	0	$\beta_0 + \beta_1 x_1$



Treatment	x_2	x_3	$E(y\|\mathbf{x})$
Drug A	1	0	$(\beta_0+\beta_2)+(\beta_1+\beta_4)x_1$
Drug B	0	1	$(\beta_0+\beta_3)+(\beta_1+\beta_5)x_1$
Placebo	0	0	$\beta_0 \quad + \quad \beta_1 \quad x_1$

What null hypothesis would you test for

- Equal slopes. $H_0: \beta_4 = \beta_5 = 0$.
- Compare slope for Drug A versus placebo. $H_0: \beta_4 = 0$.
- Compare slope for Drug A versus Drug B. $H_0: \beta_4 = \beta_5$.
- Equal regressions. $H_0: \beta_2 = \beta 3 = \beta_4 = \beta_5 = 0$.
- Interaction between age and treatment. $H_0: \beta_4 = \beta_5 = 0$.
- Effect of experimental treatment depends on age. $H_0: \beta_4 = \beta_5 = 0$.
- For patients of average age $\overline{x}_1$, are Drugs A and B equally effective? $H_0: \beta_2 + \beta_4\overline{x}_1 = \beta_3 + \beta_5\overline{x}_1$.

Copyright Information