Gov 51: Binary and Categorical Predictors

Matthew Blackwell

Harvard University

Political effects of gov't programs

- *Progesa*: Mexican conditional cash transfer program (CCT) from ~2000
  - Welfare $$ given if kids enrolled in schools, get regular check-ups, etc.
- Do these programs have political effects?
  - Program had support from most parties.
  - Was implemented in a nonpartisan fashion.
  - Would the incumbent presidential party be rewarded?

The data

- Randomized roll-out of the CCT program:
  - treatment: receive CCT 21 months before 2000 election
  - control: receive CCT 6 months before 2000 election
- Does having CCT longer mobilize voters for incumbent PRI party?

Name	Description
`treatment`	early Progresa (1) or late Progresa (0)
`pri2000s`	PRI votes in the 2000 election as a share of adults in precinct
`t2000`	turnout in the 2000 election as share of adults in precinct

```
cct <- read.csv("data/progresa.csv")
```

Difference in means estimates

- Does CCT affect turnout?

```
mean(cct$t2000[cct$treatment == 1]) -
    mean(cct$t2000[cct$treatment == 0])
```

```
## [1] 4.27
```

- Does CCT affect PRI (incumbent) votes?

```
mean(cct$pri2000s[cct$treatment == 1]) -
    mean(cct$pri2000s[cct$treatment == 0])
```

```
## [1] 3.62
```

Binary independent variables

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- When independent variable X_i is **binary**:
  - Intercept $\hat{\alpha}$ is the average outcome in the $X = 0$ group.
  - Slope $\hat{\beta}$ is the difference-in-means of Y between $X = 1$ group and $X = 0$ group.

$$\hat{\beta} = \overline{Y}_{\text{treated}} - \overline{Y}_{\text{control}}$$

- If there are other independent variables, this becomes the difference-in-means controlling for those covariates.

Linear regression for experiments

- Under **randomization**, we can estimate the ATE with regression:

```
mean(cct$pri2000s[cct$treatment == 1]) -
  mean(cct$pri2000s[cct$treatment == 0])
```

```
## [1] 3.62
```

```
lm(pri2000s ~ treatment, data = cct)
```

```
##
## Call:
## lm(formula = pri2000s ~ treatment, data = cct)
##
## Coefficients:
## (Intercept)    treatment
##       34.49         3.62
```

Categorical variables in regression

- We often have **categorical variables**:
  - Race/ethnicity: white, Black, Latino, Asian.
  - Partisanship: Democrat, Republican, Independent
- Strategy for including in a regression: create a **series of binary variables**

Unit	Party	Democrat	Republican	Independent
1	Democrat	1	0	0
2	Democrat	1	0	0
3	Independent	0	0	1
4	Republican	0	1	0
⋮	⋮	⋮	⋮	⋮

- Then include **all but one** of these binary variables:

$$\text{turnout}_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \varepsilon_i$$

Interpreting categorical variables

$$\text{turnout}_i = \alpha + \beta_1 \text{Republican}_i + \beta_2 \text{Independent}_i + \varepsilon_i$$

- $\widehat{\alpha}$: average outcome in the **omitted group/baseline** (Democrats).
- $\widehat{\beta}$ coefficients: average difference between each group and the baseline.
  - $\widehat{\beta}_1$: average difference in turnout between Republicans and Democrats
  - $\widehat{\beta}_2$: average difference in turnout between Independents and Democrats