Causation and Regression

Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

Women as Policy Makers

- Do women promote different policies than men?
- Observational studies: compare policies adopted by female politicians with those adopted by male politicians

- Randomized natural experiment:
  - one third of village council heads reserved for women
  - assigned at the level of Gram Panchayat (GP) since mid-1990s
  - each GP has multiple villages
- What does the effects of female politicians mean?

- Hypothesis: female politicians represent the interests of female voters
- Female voters complain about drinking water while male voters complain about irrigation

The Data

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

```
women <- read.csv("data/women.csv")
```

- Does the reservation policy increase female politicians?

```r
mean(women$female[women$reserved == 1])
## [1] 1
mean(women$female[women$reserved == 0])
## [1] 0.0748
```

- Does it change the policy outcomes?

```r
## drinking-water facilities
mean(women$water[women$reserved == 1]) -
    mean(women$water[women$reserved == 0])
## [1] 9.25
## irrigation facilities
mean(women$irrigation[women$reserved == 1]) -
    mean(women$irrigation[women$reserved == 0])
## [1] -0.369
```

Linear Regression Model

- Model:

$$Y = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} X + \underbrace{\epsilon}_{\text{error term}}$$

  - Y: dependent/outcome/response variable
  - X: independent/explanatory variable, predictor
  - (α, β): coefficients (parameters of the model)
  - ϵ: unobserved error/disturbance term (mean zero)

- Interpretation:
  - $\alpha + \beta X$: mean of Y given the value of X
  - α: the value of Y when X is zero
  - β: increase in Y associated with one unit increase in X

Least Squares

- Estimate the model parameters from the data
  - $(\hat{\alpha}, \hat{\beta})$: estimated coefficients
  - $\widehat{Y} = \hat{\alpha} + \hat{\beta} x$: predicted/fitted value
  - $\hat{\epsilon} = Y - \widehat{Y}$: residuals

- We obtain these estimates via the least squares method
- Minimize the sum of squared residuals (SSR):

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

- This also minimizes the root mean squared error: $\text{RMSE} = \sqrt{\frac{1}{n}\text{SSR}}$
- In **R**, use the `lm()` function and the `coef()` to extract the estimated coefficients

Slope Coefficient = Difference-in-Means Estimator

- Randomization enables a causal interpretation of estimated regression coefficient $\rightsquigarrow$ this is not always the case

```
mean(women$water[women$reserved == 1]) -
    mean(women$water[women$reserved == 0])

## [1] 9.25

lm(water ~ reserved, data = women)

##
## Call:
## lm(formula = water ~ reserved, data = women)
##
## Coefficients:
## (Intercept)     reserved
##       14.74         9.25
```

Linear Regression with Multiple Predictors

- The model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Sum of squared residuals (SSR):

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - \cdots - \hat{\beta}_p X_{ip})^2$$

- The social pressure experiment revisited:

```
social <- read.csv("data/social.csv")
levels(social$messages) # base level is `Civic'
## NULL
fit <- lm(primary2008 ~ messages, data = social)
```

Randomization of Treatments Enables Causal Interpretation

- The `lm()` function automatically creates an indicator variable for each level of a factor variable

```
fit

##
## Call:
## lm(formula = primary2008 ~ messages, data = social)
##
## Coefficients:
##       (Intercept)    messagesControl
##           0.31454           -0.01790
## messagesHawthorne  messagesNeighbors
##           0.00784            0.06341
```

- The baseline category, the `Intercept`, is `Civic Duty`

- The predicted values give the average outcome under each condition

```r
unique(social$messages)

## [1] "Civic Duty" "Hawthorne"  "Control"    "Neighbors"

predict(fit, newdata =
            data.frame(messages =
                           unique(social$messages)))

##     1     2     3     4
## 0.315 0.322 0.297 0.378

tapply(social$primary2008, social$messages, mean)

## Civic Duty    Control  Hawthorne  Neighbors
##      0.315      0.297      0.322      0.378
```

- We can create an equivalent model by replacing the intercept with the indicator variable for the baseline treatment

```r
lm(primary2008 ~ -1 + messages, data = social)
```

Heterogenous Effects by Interaction Terms

- The model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- One unit increase in $X_2 \rightsquigarrow$ the change in average Y associated with X_1 goes up by β_3
- Back to the social pressure example:

$$\begin{aligned} Y = \; & \alpha + \beta_1 \text{primary2004} + \beta_2 \text{Neighbors} + \\ & \beta_3 \text{primary2004} \cdot \text{Neighbors} + \epsilon \end{aligned}$$

```r
## subset neighbors and control groups
social.neighbor <- subset(social, (messages == "Control") |
                                  (messages == "Neighbors"))
fit.int <-
    lm(primary2008 ~ primary2004 + messages +
           primary2004:messages, data = social.neighbor)
```

```r
coef(fit.int)

##                     (Intercept) 
##                          0.2371 
##                     primary2004 
##                          0.1487 
##               messagesNeighbors 
##                          0.0693 
## primary2004:messagesNeighbors 
##                          0.0272
```

More Heterogeneity

- The model:

$$Y = \alpha + \beta_1 \text{age} + \beta_2 \text{Neighbors} + \beta_3 \text{age} \cdot \text{Neighbors} + \epsilon$$

- Compute age:

```
social.neighbor$age <- 2008 - social.neighbor$yearofbirth
```

- Fit the model:

```
fit.age <- lm(primary2008 ~ age * messages,
              data = social.neighbor)
coef(fit.age)
##            (Intercept)                    age
##               0.089477               0.003998
##      messagesNeighbors age:messagesNeighbors
##               0.048573               0.000628
```

- Create data frames with several values of age:

```
## age = 25, 45, 65, 85 in Neighbors group
age.neighbor <-
    data.frame(age = seq(from = 25, to = 85, by = 20),
               messages = "Neighbors")
## age = 25, 45, 65, 85 in Control group
age.control <-
    data.frame(age = seq(from = 25, to = 85, by = 20),
               messages = "Control")
```

- Predict turnout for each value of age and compute average treatment effect:

```
## average treatment effect for age = 25, 45, 65, 85
ate.age <- predict(fit.age, newdata = age.neighbor) -
    predict(fit.age, newdata = age.control)
ate.age

##      1      2      3      4
## 0.0643 0.0768 0.0894 0.1020
```

Regression with a Nonlinear Term

$$
\begin{aligned}
Y &= \alpha + \beta_1 \text{age} + \beta_2 \text{age}^2 + \beta_3 \text{Neighbors} \\
&\qquad + \beta_4 \text{age} \cdot \text{Neighbors} + \beta_5 \text{age}^2 \cdot \text{Neighbors} + \epsilon
\end{aligned}
$$

```
fit.age2 <- lm(primary2008 ~ age + I(age^2) + messages +
                   age:messages + I(age^2):messages,
               data = social.neighbor)
coef(fit.age2)

##                (Intercept)                           age
##                  -9.70e-02                      1.17e-02
##                   I(age^2)              messagesNeighbors
##                  -7.39e-05                     -5.28e-02
##      age:messagesNeighbors I(age^2):messagesNeighbors
##                   4.80e-03                     -3.96e-05
```

- Make prediction:

```r
## ``Neighbors'' treatment condition
yT.hat <-
    predict(fit.age2,
            newdata = data.frame(age = 25:85,
                                 messages = "Neighbors"))
## Control condition
yC.hat <-
    predict(fit.age2,
            newdata = data.frame(age = 25:85,
                                 messages = "Control"))
```

- Plot the predicted turnout:

```r
plot(25:85, yT.hat, type = "l", xlim = c(20, 90),
     ylim = c(0, 0.5), xlab = "Age",
     ylab = "Predicted turnout rate")
lines(x = 25:85, y = yC.hat, lty = "dashed")
text(40, 0.45, "Neighbors condition")
text(45, 0.15, "Control condition")
```

Graph is Helpful