Precept 6

Jordan Klein (Demography PhD Student)
jdklein@Princeton.edu
POL 345

Agenda

- Check in/logistics
- Sharing questions
- Review
  - Quick R tips
  - confidence intervals, p-values, regression in R
- Precept Questions

Check In/Logistics

- You should have received Quiz 1 back
- Pset 2 will be finished soon
- We are working on precept7 today, Handout7 is the pre-precept assignment

Share

1. Most confusing thing from the past few lectures
2. First thing you would do if covid went away tomorrow

Quick R Tips

- To clear everything R has stored & start over, run *rm(list = ls())*
- To check code, run chunks one-by-one (not entire sections/scripts at once)
- Get in the habit of putting all variables you are using in regression in the same dataset
- Get in the habit of not replacing variables in your dataset but adding new ones

Review of confidence intervals

95% confidence interval

$$= (Estimate - 1.96 \times SE\ , Estimate + 1.96 \times SE)$$

Interpretation = *"Across repeated samples, 95% of confidence intervals will contain the true value."*

Not "There is a 95% probability that the true value lies within the interval."

Review of p-values

$$\textbf{p-value} = Pr(|Z| > Test\ statistic)$$

Interpretation = *"If the null hypothesis were true, the probability we would observe a test statistic at least as extreme is (p-value)."*

We consider a p-value to be statistically significant (sufficient to reject the null-hypothesis) when it is < .05.

Linear regression in R

Generate the model

model <- lm(y ~ x1 + x2 + …, data = your_dataset, subset = (variable you want to subset by))

Examine the model

summary(model)

Precept Questions

Context

- The effect of the Electric Company (an educational show from the 70s) on children's reading ability
- Experimental study by Cooney (1976)
- Dataset = electric-company.csv
- Treatment = showing children the program
- Outcome of interest = post.score

Name	Description
`pair`	The index of the treated and control pair (ignored here).
`city`	The city: Fresno ("F") or Youngstown ("Y")
`grade`	Grade (1 through 4)
`supp`	Whether the program replaced ("R") or supplemented ("S") a reading activity
`treatment`	"T" if the class was treated, "C" otherwise
`pre.score`	Class reading score *before* treatment, at the beginning of the school year
`post.score`	Class reading score at the end of the school year

Question 1

- Load data
- Fit a linear regression of reading score on grade
- What sort of variable has R assumed grade is?
- Under what circumstances would this be a reasonable modeling choice?

Question 2

- Create a new grade factor variable
- Refit the regression
- What do the coefficients mean?

Question 3

- Fit a regression of post.score on the treatment
- Fit a regression of post.score on the treatment & grade
- Summarize both models
- Are the estimates for the treatment coefficient different in the two models?
- Are we more or less certain about the value of the coefficient in second model (with grade) compared to the first?

Question 4

- How would you compute a 95% confidence interval for the effect the of treatment from each summary table?

Question 5

- Try the *confint* function on the models
- Can we reject the null hypothesis that the treatment effect is 0 in both models?
- What do the p-values mean?
- Why are the p-values and confidence intervals different between the two models?

Question 6

- Fit a regression model for the effect of the treatment on post.score for each grade.
- How do the treatment effects differ as grade increases?
- Are these ATEs? If so, which populations are they ATEs for?
- What do we call ATEs for specific values of pre-treatment variables?

Question 7

- How confident would you be that the treatment effects are non-zero on the basis of these models, are we less confident about some effects? If so, why do you think that is?
- How many data points are used in each model?

Question 8

- Add pre.score to the models from question 6
- Do we become more or less sure about the value of the treatment after adding pre.score? Why do you think that is?
- What are the advantages and disadvantages of these multiple models over fitting just one model?