Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 7

Download Data and Open R Script

We'll be using GSS and CA Data!

1. Download the Lecture 7 PDF and R files for this week
   - Place all files in HED612_S21 »> lectures »> lecture7
2. Open the RProject (should be in your main HED612_S21 folder)
3. Once the RStudio window opens, open the Lecture 7 R script by clicking on:
   - file »> open file... »> [navigate to lecture 7 folder] »> lecture7.R

Schedule

What we have done:

- Prediction
- Population Regression Model & OLS Prediction Line [will review today]
- Interpretation of $\hat{\beta}_1$ with continuous X [will review today]

Today:

- Confidence Intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$
- Mini R lesson on creating new variables
- Interpretation of $\hat{\beta}_1$ with categorical X
  - Homework 7 posted on D2L
  - Reading for next week: TBD [on Omitted Variable Bias]

Next week [3/3/2021]:

- Review SER vs SE of $\hat{\beta}_1$
- OLS Assumptions
- Introduction to Omitted Variable Bias
- Final project requirements and intro to possible datasets!

3/10/2021: Spring Break/Reading Day [No Class]

3/17/2021: Intro to Multivariate Regression

General Regression Purposes

Things we generally do with regression:

- Prediction
  - Here we're interested in knowing/predicting $\hat{Y}$
  - Example: Predict poverty status from owning a cell phone
  - Example: Predict academic probation for early warning system
    - We don't really care which X variables predict academic probation... (i.e., absences, going to REC center 2+ times a week, etc.)
- Hypothesis Testing about β_1
  - Here we're interested primarily the impact our X has on Y, in other words the slope and significance of $\hat{\beta}_1$
  - Example: Does smaller class sizes cause better student learning
  - Example: Does receiving federal financial aid have an effect on on-time graduation
  - Our focus is on our one independent variable of interest, and sometimes test the effect of X on multiple Y's (i.e., on-time graduation, first to second year retention, GPA, etc)

Population Regression Model and OLS Prediction Line

RQ: What is the effect of student-teacher ratio (X) on student test scores (Y)

Population Linear Regression Model: $Y_i = \beta_0 + \beta_1 X_i + u_i$

- Where Y = student test scores
- Where X = student teacher ratio

OLS Prediction Line or "OLS Regression Line" (without estimates): $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- We DROP the residual term! Why?
- Residuals = everything not included in the model that account for the difference between actual observed value of Y and Y value predicted by OLS regression
- We can only predict values of Y based on data we have!
- We use residuals to understand how good our predictions are! (SER)

Interpretation of $\hat{\beta}_1$ for Continous X

OLS Prediction Line or "OLS Regression Line" (with estimates): $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Interpretation of $\hat{\beta}_1$

- General interpretation [always true!]
  - The average effect of a one-unit increase in X is associated with a $\hat{\beta}_1$ unit change (negative = decrease or positive = increase) in Y

```
summary(mod1)
#>
#> Call:
#> lm(formula = testscr ~ str, data = caschool)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -47.727 -14.251   0.483  12.822  48.540
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
#> str          -2.2798     0.4798  -4.751 2.78e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 18.58 on 418 degrees of freedom
#> Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
#> F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Point and Interval Estimation

- Paramater
  - A summary of the population; usually unknown
- Point Estimate
  - A single number that is the best guess for the parameter (e.g., $\hat{\beta}_1$)
  - Use hypothesis testing to explore whether there is a significant relationship between X and Y
    - $H_0 : \beta_1 = 0$ (no relationship: no effect of X on Y)
    - $H_0 : \beta_1 \neq 0$ (there is a relationship: X does have an effect on Y)
- Interval Estimate
  - An interval around the point estimate, within which the parameter value is believed to fall
  - e.g., If $\hat{\beta}_1$= 2.5, we are 95% sure that $\hat{\beta}_1$ is between 1.5 and 3.5

Confidence Intervals

Confidence intervals about β_1

- General formula for confidence intervals
  - (point estimate) $\pm$ z*SE(point estimate)
  - Where z $=$ z-score associated with desired confidence interval

Confidence Interval	Z-Score
90%	1.645
95%	1.96
99%	2.576

- Formulas for 95% confidence interval (CI) of β_1
  - $\hat{\beta}_1 \pm 1.96 * \text{SE}(\hat{\beta}_1)$
  - Interpretation: We are 95% confident that the population parameter β_1 lies somewhere between [lower bound] and [upper bound]

- What happens to CI when you choose a higher “confidence interval” level? Why?
  - e.g., 99% CI instead of 95%

Confidence Interval in R

RQ: What is the effect of district average income (in $000s) (X) on student test scores (Y)?

- Write out population regression model
- Write out OLS regression without estimates

Run regression in R

- Calculate 95% CI using $\hat{\beta}_1$ and SE($\hat{\beta}_1$)
  - We are 95% confident that the population parameter β_1 lies somewhere between 1.70 and 2.06
- Calculate 99% CI using $\hat{\beta}_1$ and SE($\hat{\beta}_1$)
  - We are 99% confident that the population parameter β_1 lies somewhere between 1.64 and 2.11

Use `confint()` in R to calculate CI

Confidence Intervals and Hypothesis Testing

We always test same null hypothesis

- $H_0 : \beta_1 = 0$
- Reject H_0 if p-value is less than "alpha-level"

Relationship between confidence intervals and hypothesis tests about β_1

- Assume testing H_0 with alpha-level $= .05$
  - If p-value for H_0 is less than .05, then 95% CI will not contain zero (our value associated with the null)
  - If 95% CI does not contain zero (our value associated with the null), then p-value for H_0 is less than .05
- Assume testing H_0 with alpha-level $= .01$
  - If p-value for H_0 is less than .01, then 99% CI will not contain zero (our value associated with the null)
  - If 99% CI does not contain zero (our value associated with the null), then p-value for H_0 is less than .01

Student Exercise #1

Using CA Schools data: RQ: What is the effect of student teacher ratio (X) on test scores (Y)?

1. Write out population regression model for the effect of student teacher ratio on district test scores?
2. Run the regression in R as `stuex_mod`. Write the OLS prediction line with estimates
3. What is the point estimate for β_1? Interpret this estimate.
4. Using R's `confint()` function, what is the 95% CI for β_1? Interpret in words.
5. Calculate the 99% on your own using $\hat{\beta}_1$ and SE($\hat{\beta}_1$).

Student Exercise #1 [Solutions!]

1. Write out population regression model for the effect of student teacher ratio on district test scores?

- $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Where Y = `testscr`
  - Where X = `str`

2. Run the regression in R as `stuex_mod`. Write the OLS prediction line with estimates

- $\hat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $\hat{Y_i} = 699 + (-2.28) X_i$

3. What is the point estimate for β_1? Interpret this estimate.

- Point estimate for β_1: $\hat{\beta_1}$ = 2.28; A one unit increase in student teacher ratio is associated with a 2.28 point decrease, on average, on district student test scores

4. Using R's `confint()` function, what is the 95% CI for β_1? Interpret in words.

- `confint(stuex_mod, level = 0.95)`: We are 95% confident that the population parameter β_1 lies somewhere between -3.22 and -1.34

5. Calculate the 99% on your own using $\hat{\beta_1}$ and SE($\hat{\beta_1}$). - $\hat{\beta_1}$ $\pm$ 2.576 * SE($\hat{\beta_1}$)

- -2.28 $\pm$ 2.576 * SE(0.4798)
- -2.28 $\pm$ 1.235965
- -3.52, -1.04

Creating variables in R

Creating "analysis" variables in R

- Quantitative researchers really need two different skills
  - **Statistics**: learning how to run and interpret apppropriate statistical tests/methods according to RQs; learning how to apply findings to answer RQs; draw recommendations and implications from statistical findings
  - **Data Management**: proficiency in fundamental data management and manipulation tasks like managing data in their raw forms, creating variables, combining multiple datasets, manipulation or reshaping data, etc.
- This is not a "data management" class; or a class to teach you how to use R
- We simply use R to run the statistical tests associated with linear regression
- We, for the most part, use data that is "clean"; however you often need to create different "analysis" versions of variables
- We are gonna have a "crash course" lecture on creating categorical variables today.
  - We'll start on learning this via creating a dummy variable; then in following weeks we will create variables with 2+ categories

"Coneptual" Process for Creating "analysis" variables in R

1. Investigating values and patterns of variables from "input data"
2. Identifying and cleaning errors or values that need to be changed
3. Creating "analysis" variables
4. Checking values of analysis variables against values of input variables

- Example: Create a new dummy variable (`ba_degree`) for whether a respondent has at least a Bachelor's degree (`ba_degree` $= 1$) or less than a Bachelor's degree (`ba_degree` $= 0$)

Creating variables in R

Task: Create a new dummy version of the variable `degree` called `ba_degree`

- 1 indicates respondent has at least a Bachelor's degree
- 0 indicates respondent has less than a Bachelor's degree

Step 1: Investigate values and patterns of variable from "input data"

- use `var_label()` and `val_labels()` to check variable and value labels
- use `count()` to get a frequency count of each category
- use `count()` + `is.na()` to check if there are any missing observations

Step 2: Identifying and cleaning errors or values that need to be changed

- Most common cleaning error to fix: National surveys don't report missing as `NA`; they assign "strange" values to missing observations
  - `-99` = `item legitmate skip`
  - `-98` = `no response`
- Use `mutate()` to fix errors
  - Create version of "input" variables that code missing values (-99,-98) as true missing in R (`NA`).
  - Our `degree` variable is actually clean already! All missing categories are already coded to `NA` (they were previously IAP=-8, DK=-9, and NA=-7)
  - We'll practice this step next week on a variable that is not so clean!

Creating variables in R

Task: Create a new dummy variable (`ba_degree`)

- 1 indicates respondent has at least a Bachelor's degree
- 0 indicates respondent has less than a Bachelor's degree

Step 3: Creating "analysis" variables

- Create dummy vars via `mutate()` + `ifelse()` ; where general syntax is:

```
df <- df %>%
 mutate(NEWVAR=
          ifelse(OLDVAR+CONDITION, value if TRUE, value if FALSE))
```

Step 4: Checking values of analysis variables against values of input variables

- Use `group_by() + count()` to check new and old variables against each other

Warning: You will ONLY use the assignment operator `<-` within Steps 2 and Steps 3; using `<-` in Step 1 or Step 4 will change the original `GSS` dataset

- If you make this mistake; just reload your gss dataset!
- show in R

10 Minute Break

Interpretation of $\hat{\beta}_1$ with Categorical X

Interpretation of $\hat{\beta}_1$ with Categorical X

- Many independent variables of interest are categorical rather than continous
- How to distinguish between continuous and categorical variables when running regression?
  - Continous variables:
    - Difference between one value and another is quantitative
    - e.g., SAT score of 900 vs. 1000; income of $40k vs $45k, GPA of 2.0 vs. 2.1
  - Categorical variables:
    - Difference between one value and another cannot be measured quantitatively
    - e.g., race/ethnicity, parent education is B.A. vs M.D., political ideology

Many program evaluation questions involve a categorical independent variable of interest

- What is the effect of receiving a pell grant on on-time college completion?
- What is the effect of participating in Mexican American Studies program on high school graduation?
- What is the effect of Head Start pre-k on Kindergarten reading levels?
- What is the effect of class size (small vs large) on student learning?

General Steps for Regression with Categorical X

- Identify categories of X and choose a reference group
- Create 0/1 variables for each group
- Write out population model and OLS regression line
- Run regression in R
- Interpret estimates

Interpretation of $\hat{\beta}_1$ with Categorical X

- Y = income and X = college graduate (BA or higher)
- Choose the "reference group" or "base level"; this is who all other groups will be compared to (similar to ANOVA)
  - Non-college graduates will be our reference group
  - College graduates will be our non-reference group
  - If your categorical variable is a dummy variable, your reference category should be equal to zero and non-reference equal to 1!
  - We already did this! `ba_degree` = 0 for lower than a BA and 1 = for BA or higher

- Population regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Where Y= income
  - X= 0/1 college graduate
    - 0 = non-college graduate [reference group]
    - 1 = college graduate [non-reference group]
- OLS Prediction Line [run regression in R]
  - R will automatically assign the lowest value of X as your reference category!
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - $\hat{Y}_i = 17551.5 + 21090.2 * X_i$
- **Generic interpretation of $\hat{\beta}_1$ for categorical X**
  - **The average effect of being [specific non-reference category] as opposed to [reference category] is associated with a $\hat{\beta}_1$ change in Y**
- Specific interpretation of $\hat{\beta}_1 = 21090.2$
  - the average effect of having a **BA or higher** as opposed to having **less than a BA**, on average, is associated with a $21,090.20 increase in annual income

Same Interpretation of $\hat{\beta}_1$

- Generic interpretation of $\hat{\beta}_1$ when X=continuous
  - The average effect of a one-unit increase in X is a $\hat{\beta}_1$ unit change in the value of Y
- Generic interpretation of $\hat{\beta}_1$ when X=categorical
  - X; 0= reference group; 1= non-reference group
  - The average effect of being [non-reference group] as opposed to [reference group] is associated with a $\hat{\beta}_1$ change in Y
    - In other words: a one-unit increase in X (from X=0 to X=1) is associated with $\hat{\beta}_1$ change in Y

So interpretation of $\hat{\beta}_1$ for categorical X is the same as for continous X

- In both cases $\hat{\beta}_1$ is the effect of a one unit increase in X
- But in categorical, X can only increase one unit (from X=0 to X=1)

Interpretation of $\hat{\beta}_1$ with Categorical X

- What if we switch our reference category so that X is now a dummy variable where:
  - X=0 are BA or Higher
  - X= 1 are lower than a BA
  - make this variable `lower_ba` in R
- Population regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Where Y= income, X= 0/1 non-college graduate
- OLS Prediction Line [run regression in R]
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - $\hat{Y}_i = 38,642.5 + (-21090.2) * X_i$
- Generic interpretation of $\hat{\beta}_1$ for categorical X
  - the average effect of being [specific non-reference category] as opposed to [reference category] is associated with a $\hat{\beta}_1$ change in Y
- Specific interpretation of $\hat{\beta}_1 = -21090$
  - students answer
- How do we explain the differences in $\hat{\beta}_0$ between these two regressions?
  - Model 2 [X=0 lower than BA, X=1 BA or higher]: $\hat{\beta}_0$ = 17551.5
  - Model 3 [X=0 BA or higher, X=1 lower than BA]: $\hat{\beta}_0$ = 38642

Interpretation of $\hat{\beta}_1$ with Categorical X (more than 2 categories!)

- What is the effect of respondent political party on income?
  - Categories: democrat, republican, independent, unknown party
  - We need to create dummy variables for each of these categories first! [show in R]
- Choose category that will be our "reference group"
  - Let's choose democrats as the reference group!
  - When we have more than one category; we run the regression with all category dummies *except* the reference group!
- Population regression model
  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
  - Where Y= income;
  - X_1= 0/1 republican, X_2 = 0/1 independent, X_3 = 0/1 unknown party; Reference Category = democrats
- OLS Prediction Line [run regression in R]
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$
  - $\hat{Y}_i = 24476 + 6184 * X_{1i} + (-1736) * X_{2i} + (-3335) * X_{3i}$

Interpretation of $\hat{\beta}_1$ with Categorical X (more than 2 categories!)

- **Interpretation of $\hat{\beta}_1$: 6184**
  - Generic (for categorical): the average effect of being [specific non-reference category] as opposed to [reference category] is associated with a $\hat{\beta}_1$ change in Y
  - Specific for this example: the average effect of being republican as opposed to being a democrat is associated with a $6,184 increase in annual income
- We interpret $\hat{\beta}_2$ and $\hat{\beta}_3$ the same way!
- **Interpretation of $\hat{\beta}_2$: -1736**; the average effect of being independent as opposed to being a democrat is associated with a $1,736 decrease in annual income (but not significant!)
- **Interpretation of $\hat{\beta}_3$: -3335**; the average effect of having an unknown political party as opposed to being a democrat is associated with a $3,335 decrease in annual income (but not significant!)

Prediction with Categorical X (more than 2 categories!)

- Prediction with categorical X works the exact same way!
- Show on Whiteboard
  - Population Model
  - OLS Line with estimates
  - Calculate predicted income for republicans, independents, democrats

R Shortcut for Creating Dummies for Vars with 2+ categories

- It's a pain to create dummy versions for all categorical variables with 2+ categories
  - Some categorical variables can have many categories, which means you have to create as many dummy variables as you have categories (ugh!)
- There's an R shortcut!
  - Create one categorical variable with as many categories neeeded
    - we create it using `mutate()` + `case_when()`
  - We insert the new categorical variable into our regression
  - R "creates" the dummies for each category on the "backend"
  - R will assume lowest value category is reference; OR we can explicitly indicate what the reference category is!

Homework Data

Educational Longitudinal Study Data

- Educational Longitudinal Study of 2002
  - (https://nces.ed.gov/surveys/els2002/)[ELS Website]
  - Nationally representative, longitudinal study of 10th graders in 2002 and 12th graders in 2004
  - Students followed throughout secondary and postsecondary years
  - Surveys of students, their parents, math and English teachers, and school administrators
  - Student assessments in math (10th & 12th grades) and English (10th grade)
  - High school transcripts available for research on coursetaking

- Problem Set #7
  - You will need to download the data from D2L [all instructions are detailed in the assignment]
  - I will give you all coded needed to create new variables; but try to get a bit of intuition behind the code [all based on what we learned this lecture]