Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 3

1. Prep

2. Review of Statistics

Prep

Download Data and Open R Script

Download Data and Open R Script

1. Create a new data folder called "gss"
   - HED612_S21 »> data »> gss
2. Download the GSS 2018 dataset from D2L (under Datasets)
   - Place the "GSS2018.RData" dataset into the "gss" folder you created in the previous step
3. Download the Lecture 3 PDF and R files for this week
   - Place all files in HED612_S21 »> lectures »> lecture3
4. Open the RProject (should be in your main HED612_S21 folder)
5. Once the RStudio window opens, open the Lecture 3 R script by clicking on:
   - file »> open file... »> [navigate to lecture 3 folder] »> lecture3.R

Today and Next Week

Today

- Review of statistics (bivariate)
- Some more R Basics
- Intro to “Gold Standard” for causal inference and bivariate regression
- Class exercise

HW & Reading

- HW#3 posted on D2L
- Stock & Watson Ch. 4 [will cover Chapter 4 next 2 weeks]

Next Week

- Bivariate regression cont.

COE Student R Group

- HED students are starting an R “working group”
  - Will meet regularly
  - Troubleshoot errors
  - Work collaboratively on class assignments (highly encouraged!)
  - Learn new skills via tutorials
- I have committed to providing some guidance and will participate when I can
- So far looks like Thursdays from 5-7 may work
  - May consider meeting every two weeks...
  - If you want to join but have different availability fill out Doodle poll
  - Doodle Poll

Homework Review

- Common Questions for PS#2 [review as a class]
  - Packages vs Libraries in R
    - Packages are collections of R functions, data, and compiled code in a well-defined format, created to add specific functionality. We only need to install packages in R once via `install.packages()`
    - Libraries are the "directories" in R where the packages are stored; we need to "load" these libraries via `library()` each time we would like to use any of the functionality associated with the package
  - Rounding in R
    - It's okay if your hand calculations are slightly off from R answers due to rounding...
    - Why is it rounding to an integer?
    - R is an **object oriented programming language**; which means it saves "results" into various objects which "function" differently
    - Piping creates a tibble object...which will round to the nearest integer
    - [Show solutions in R and D2L]
- Questions or Concerns?

Review of Statistics

Hypothesis testing regarding the mean(s)

Many hypotheses we have about education research can be phrased as yes or no questions:

- Do students in smaller class sizes get better grades?
- Is there a difference in earnings between men and women college graduates?
- Do wealthier public schools get more recruiting visits by colleges and universites than poorer schools?

Intro to Stats:

- Hypothesis testing regarding the population mean
- Hypothesis testing regarding two populations

Hypothesis testing: population mean

Null hypothesis: H_0: $E(Y) = \mu_{Y,0}$

Alternative hypothesis: H_1: $E(Y) \neq \mu_{Y,0}$

- Sample mean, $\bar{Y}$ or $E(Y)$, is rarely exactly equal to the hypothesized value, $\mu_{Y,0}$, in any given sample
  - Because the true population in fact does not equal the hypothesized value (the null hypothesis is false) OR
  - Because the true population does equal the hypothesized value but $\bar{Y}$ differs from the hypothesized value because of random sampling variation

Solution: we test the null hypothesis accounting for sampling variation!

- Perform a t-test to check whether there is a significant difference between the population mean (which is equal to the mean of sample means from last week's sampling distribution) and the hypothesized value
  - Compute the standard error of $\bar{Y}$; $SE(\bar{Y}) = s_Y/\sqrt{n}$
  - Compute the t-statistic; $t = \bar{Y} - \mu_{Y,0}/SE(\bar{Y})$
  - Compute the p-value (significance probability) of null hypothesis; $p - value = 2\Phi(-|t|)$
  - P-value is the probability of drawing $\bar{Y}$ at least as far in the tails of its distribution under the assumption that the null hypothesis is correct as the sample average you actually computed

Hypothesis testing: population mean [Stock & Watson Example]

Example: In the population, college graduates earn $20 an hour in the labor market

- **Null hypothesis**: H_0: $E(Y) = \mu_{Y,20}$
- **Alternative hypothesis**: H_1: $E(Y) \neq \mu_{Y,20}$
- Random Sample of 200 college graduates
  - $\bar{Y} = \$22.64$ is the mean hourly earnings of the sample
  - $s_Y = \$18.14$ is the standard deviation

- Step 1: Compute the **standard error of** $\bar{Y}$
  - $SE(\bar{Y}) = s_Y / \sqrt{n}$
- Step 2: Compute the **T-statistic**: measures the size (in units of standard error) of the difference between the population mean and the hypothesized value relative to the variation in the sample data
  - $t = \bar{Y} - \mu_{Y,0} / SE(\bar{Y})$
- Step 3: Compute the **P-value**: the probability of observing another $\bar{Y}$ at least as different from $20 as $22.24 by pure random sampling variation assuming the null is correct
  - $p - value = 2\Phi(-|t|)$
  - If the p-value is 0.039, then there is only a 3.9% probability that a similar ($\bar{Y} = \$22.24$) would have been drawn if the null is true, we can reject the null hypothesis
  - If the p-value is 0.40 (hypothetical), then there is about a 40% probability that a similar ($\bar{Y} = \$22.24$) would have been drawn if the null is true; we cannot reject the null

Hypothesis testing: comparing means from two populations

Do men and women college graduates, on average, earn the same amount?

Null hypothesis: H_0: $\mu_m - \mu_w = 0$

Alternative hypothesis: H_1: $\mu_m - \mu_w \neq 0$

- \$21.99 is the average hourly earnings of men in the sample of college graduates
- \$18.48 is the average hourly earnings of men in the sample of college graduates
- \$21.99 - \$18.48 = \$3.52

T-statistic measures the size (in SE) of the difference between the population wage gap and the hypothesized wage gap relative to the variation in the sample data

- the greater the magnitude of t-statistic, the greater evidence against the null
- the closer the magnitude of t-statistic to zero, the more likely it is that there is no signifcant difference between the population mean and hypothesized value

The **P-value** is the probability of observing a difference of $\mu_m - \mu_w$ at least as different from zero as the observed difference of \$3.52 by pure random sampling variation

- If the p-value is 0.05, then there is about a 5% probability that a similar sample mean difference between men and women would have been drawn if the null hypothesis is true, so we can reject the null hypothesis
- If the p-value is 0.40, then there is about a 40% probability that the \$3.52 difference in sample average earnings between men and women could have arisen just by random sampling variation if the null hypothesis is true, so we would not reject the null hypothesis

Relationships between two continuous variables

Postive relationship, negative relationship, and no relationships

1. Relationship between X and Y is positive

- when X is “high”, Y tend to be “high”
- when X is “low”, Y tends to be “low”
- e.g., number of hours studying and GPA

2. Relationship between X and Y is negative

- when X is “high”, Y tend to be “low”
- when X is “low”, Y tends to be “high”
- e.g., number of school abscences and GPA

3. No relationship between X and Y

- knowing the value of X gives you does not tell you much about the value of Y
- e.g., amount of ice cream consumed and GPA

Today's Example

Using the General Social Survey 2018 [sidebar: Data Goals for this Class]:

- Nationally representative survey of adults in the United States conducted since 1972
- Collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors
  - Demographic, behavioral, and attitudinal questions
  - Covers topics like civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, etc.
- GSS Website Link

We have two variables X and Y:

- X (independent variable) = hours worked per week
- Y (dependent variable) = income

Ways to investigate this relationship between X and Y:

- Graphically: scatterplots
- Numerically: covariance (less used), correlation

Scatterplots

- Scatterplots will plot individual observations on an X and Y axis
- Draw scatterplot of X (hours worked) and Y (income) by hand
  - Add a prediction line
- Residual
  - Difference between actual observed value of Y and predicted value of Y (given X)
- Relationship between X and Y is not perfect!

Generate a scatteplot using `ggplot` in R script

Covariance

Covariance measures the extent to which two variables move together....

- If income is "high" when hours is worked is "high", then covariance is positive
- If income is "low" when hours is worked is "high", then covariance is negative

Population covariance, cov(X, Y)

- As with all population parameters, we don't know this!

Population covariance, s_{XY} or $\hat{\sigma}_{XY}$

- Estimator of population covariance
- $s_{XY} = \hat{\sigma}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

Sample Covariance

Formula $s_{XY} = \hat{\sigma}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

Example: Imagine we have 20 obs; $\bar{X} = 40; \bar{Y} = 30$

Observation 1: $X_1 = 50; Y_1 = 60$

- $(X_i - \bar{X})(Y_i - \bar{Y}) = (50 - 40)(60 - 30) = 10 * 30 = 300$
- $X_i > \bar{X}$ and $Y_i > \bar{Y}$; so $(X_i - \bar{X})(Y_i - \bar{Y})$ is positive

Observation 2: $X_1 = 45; Y_1 = 25$ - $(X_i - \bar{X})(Y_i - \bar{Y}) = (45 - 40)(25 - 30) = 5 * -5 = -25$ - $X_i > \bar{X}$ and $Y_i < \bar{Y}$; so $(X_i - \bar{X})(Y_i - \bar{Y})$ is positive

s_{XY} is the sum of these 20 calculations divided by 19 (n-1)

Sample Covariance cont.

s_{XY} is positive when X and Y move in the same direction

- $X_i > \bar{X}$ usually coupled with $Y_i > \bar{Y}$
- $X_i < \bar{X}$ usually coupled with $Y_i < \bar{Y}$

s_{XY} is negative when X and Y move in the same direction

- $X_i > \bar{X}$ usually coupled with $Y_i < \bar{Y}$
- $X_i < \bar{X}$ usually coupled with $Y_i > \bar{Y}$

Correlation

Problem with sample covariance, s_{XY}

- Covariance (like variance) dependes on the units of measurement
- We can't compare the covaraince of X and Y vs covariance of X and Z

Sample Correlation of Z and Y, r_{XY}

- Unitless measure of relationship between X and Y
- Equals sample covariance, s_{XY}, divided by the product of their individual sample standard deviations

Sample Correlation Formula $r_{XY} = \frac{s_{XY}}{s_X * s_Y} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$

Sample Correlation

$r_{XY} = \frac{s_{XY}}{s_X * s_Y} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$

Correlations result in measures between -1 and 1

“Type” of relationship

- $r_{XY} = 0$ means there is no relationship between X and Y
- $r_{XY} > 0$ means a positive correlation between X and Y
  - in other words, the variables move together
- $r_{XY} < 0$ means a negative correlation
  - in other words, the variables move in opposite directions

“Strength” of relationship

- $r_{XY} = |0.1|$ to $|0.3|$ = weak relationship
- $r_{XY} = |0.3|$ to $|0.6|$ = moderate relationship
- $r_{XY} = |0.6|$ to $|1|$ = strong relationship

Calculate correlations in R...

Linear vs Non-Linear Relationships

Problem with covariance and correlation:

- Both measure linear relationships; these measures do not detect non-linear relationships

See R script:

- Run correlation of income and age for respondents that identified as Black
- Run Scatterplots

Random Assignment

Create the variable `randomvar` where values will be randomly assigned

- Will randomly assign observations to values of 0 to 1000 for this new random var
- Run correlation between `randomvar` and income
  - What is the relationship beween `randomvar` and income (positive, negative, no correlation)? Why?

This is the intuition behind **Randomized Control Trials** as the "gold standard" of progam evaluation research

Randomization process

- Reseachers use a "coin flip" to sort participants (of a program, policy, treatment, etc.) into two groups: treatment group or control group
  - Coin flip = randomization
  - Treatment is completely unrelated to participant's background/demographic characteristics
- **On average**, the two groups should be *identical* in every way
- Treat the treatment group; control group does not get treatment
- Compare outcomes for the treatment and control group
  - $\bar{Y}_{treat} - \bar{Y}_{control}$ = treatment effect
  - examples: baby aspirin trial, Headstart, Tenessee Star Project

In-Class Exercise [in groups]

1. What is the correlation between how presitigious father's occupation is (`papres105plus`) and income (`incomev2`)?
2. What is the correlation between hours worked (`hrs1`) and income (`incomev2`)for respondents born in the U.S (`born`)? How about for those not born in the U.S.? Do these correlations differ by birth in the U.S.?

- Hint: If you run `gss %>% select(born) %>% val_labels()` you'll see that the value `1` is "yes" and value `2` is "no."
- Use tidyverse approach to run correlation; use the `filter()` function to run a correlation for each group

1. What is the correlation between hours worked (`hrs1`) and income (`incomev2`) for respondents with at least a bachelors degree? How about for those with less than a bachelors degree?

- Hint: If you run `gss %>% select(degree) %>% val_labels()` you'll see that the value `1` is "high school", the value `2` is "junior college", value `3` is "bachelor", and value `4` is "graduate".
- Use tidyverse approach to run correlation; use the `filter()` function to run a correlation for those those with (at least) a bachelors and those without.

Code solutions on next slide...

In-Class Exercise [Solutions]

```r
#1
gss %>% summarise(cor(papres105plus, incomev2, use = "complete.obs"))

#2
gss %>% select(born) %>% val_labels()

gss %>% filter(born==1) %>%
   summarise(cor(hrs1, incomev2, use = "complete.obs"))

gss %>% filter(born==2) %>%
   summarise(cor(hrs1, incomev2, use = "complete.obs"))

#3
gss %>% select(degree) %>% val_labels()

gss %>% filter(degree>=3) %>%
      summarise(cor(hrs1, incomev2, use = "complete.obs"))

gss %>% filter(degree<=2) %>%
      summarise(cor(hrs1, incomev2, use = "complete.obs"))
```