Introduction to Multivariate Regression & Econometrics

HED 612

Lecture 4

Prep

Download Data and Open R Script

Download Data and Open R Script

1. Download the Lecture 4 PDF and R files for this week
   - Place all files in HED612_S21 »> lectures »> lecture4
2. Open the RProject (should be in your main HED612_S21 folder)
3. Once the RStudio window opens, open the Lecture 4 R script by clicking on:
   - file »> open file... »> [navigate to lecture 4 folder] »> lecture4.R

We will be using the GSS and CA School datasets today so no need to re-downlaod

Homework Review

- Common Issues and Concerns
  - Technical Issues with R
    - Errors/Issues arise whether you are new to R or have been using R for 10+ years
    - Don't be afraid to get errors; don't let errors discourage you; do your best to solve the problem but if you can't figure it out ask for help!
    - You are not the only one getting errors! Nearly half the class emailed me in the last 6 hours.
    - Precisely why I use D2L discussion boards when teaching methods classes; please don't feel embarrassed to post!
    - As a class lets all subscribe to these discussion boards! [Show on D2L]
  - GSS Data
    - Income example is great conceptually as we try to "understand" regression; variable is created a little wonky!
    - I could pick another example; but the wonkyness is part of life as a quantitative research!
    - We need good researchers that understand and can manage data well in positions that prevent these sorts of issues [You!]

HED R Coding Group

THURSDAYS (for Spring 2021)

5PM – 7PM

http://**bit.ly/hedrcoding**

Building a **supportive R coding community**.

All levels encouraged to attend!

Special guest appearances by Dr. Karina Salazar.

Today and Next Week

Today

- Intro to Bivariate Regression
  - linear regression model
  - estimating parameters

HW & Reading

- HW#4 posted on D2L
- Stock & Watson Ch. 4 [finish if you haven't]

Next Week

- Bivariate regression
  - Prediction
  - Model fit

Introduction to Bivariate Regression

Purpose of Regression

- Regression analysis is a statistical method that helps us analyze and understand the relationship between 2+ variables
- What is the purpose of regression in **descriptive research** (sometimes called "observational studies" or "predictive" studies)?
  - To understand **relationship(s)** between one dependent variable (Y) to one or more indepedent variable (X, Z, etc.)
  - Not concerned with "direction" or "cause": Does X cause Y? Does Y cause X?
  - Interested in "prediction"
  - Example: predict poverty status based on having a cell phone
- What is the purpose of regression in **econometrics research** (sometimes called "causal studies")?
  - To estimate the **causal** effect of an independent variable (X) on a dependent variable (Y)
  - Very concerned with "direction" or "cause": Does X cause Y?
  - Interested in recreating experimental conditions or what would have happened under a randomized control trial
  - Example: What is the effect of class size on student learning?
- Most of my research is descriptive; but I teach this class in a "causal" way... why?
  - One type of research is not better than the other; it's just really important to understand the difference. *Ex: Lack of a cell phone doesn't cause poverty!*
  - Causal research forces you to be very purposeful about your models!
  - Policy makers/decision makers don't just care if there is a relationship between class size and student learning; they want to know if we decrease class size by two students what is the expected change in student test scores

Regression: Models, Variables, Relationships

- **Linear Regression Model vs Non-Linear Regression Models**
  - Linear regression model (general linear model)
    - the dependent variable is continuous
    - e.g., GPA, test scores, income
    - **the focus of this class!**
  - Non-linear regression models (logit, ordinal, probit, poisson, negative binomial)
    - the dependent variable is non-continuous (i.e., categorical, binary, counts)
    - e.g., persistence, likert scales, type of major
    - **focus of HED 613** (Spring 2022/Maybe Fall 2021)
- **Bivariate vs Multivariate Regression**
  - Bivariate regression (sometimes also called univariate, simple regression)
    - One dependent variable (Y) and one independent variable of interest (X_1)
  - Multivariate regression (for econometrics/causal inference)
    - One dependent variable (Y) and one independent variable of interest (X_1); **and** multiple control variables (X_2,X_3,X_4,etc.)
- **Linear Relationship vs Non-Linear Relationship between X and Y**
  - Draw linear vs non-linear relationship scatterplots; show in R
  - We will focus on modeling linear relationship between X and Y for first half of the course
  - Then we will cover non-linear relationships

Rest of Lecture: Same Example as Last Week

- We will be using the same example as last week:
  - What is the effect of hours worked (X) on income (Y)?
- We'll be using the continuous version of income from PS3
  - `realrinc`

Slope Measures relationship between X and Y

- Research Question
  - What is the effect of hours worked (X) on annual income (Y)
- What do we want to measure?
  - The relationship between hours worked (X) and income (Y)
  - If we increase the number of hours worked per week by one additional hour, how much do we expect annual income to change?
- We want to measure β
  - $\beta = \frac{(Y_2 - Y_1)}{(X_2 - X_1)} = \frac{\Delta Y}{\Delta X}$
  - In other words, the slope of the relationship between X and Y
  - Under the assumption of a linear relationship
- Draw line showing linear relationship between X and Y
  - $x_1 = 31, x_2 = 32, y_1 = \$30,000, Y_2 = \$35,000$
  - Calculate slope at different points and for different ΔX

Population Linear Regression Model

Population Linear Regression Model

Population Linear Regression Model

- $Y_i = \beta_0 + \beta_1 X_i + u_i$

Where:

- Y_i = income for person i
- X_i = hours worked per week for person i
- β_0 ("population intercept") = average income for someone with X=0
- β_1 ("population regression coefficient") = average effect of a one-unit increase in X on the value of Y
- u_1 ("error term" or "residual") = all other variables not included in your model that affect the value of Y

Draw Picture

- Scatterplot of the population
- Population regression model line
- Label the following:
  - β_0
  - β_1
  - residual (predicted - value of Y_i)

Population Regression Model

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

Contains two population parameters

- β_0 ("population intercept") = average income for someone with X=0
- β_1 ("population regression coefficient") = average effect of a one-unit increase in X on the value of Y

- How do we know these are population parameters? (hint: notation)
- Do we usually know the value of β_0 or β_1?

Population regression coefficient, β_1

What is the effect of hours worked per week (X) on income (Y)?

- Answer: population regression coefficient, β_1
- Estimating β_1 is the fundamental goal of causal inference/this course

What is the population regression coefficient, β_1?

- β_1 measures the average change in Y for a one-unit increase in X
- Think of β_1 as measuring the slope of our prediction line!
- $\beta_1 = \frac{\Delta Y}{\Delta X} = \frac{\Delta Income}{\Delta HoursWorked}$
- Example: $\beta_1 = \frac{\$5000 \Delta Income}{1hour \Delta HoursWorked}$ = \$5,000

Interpretation (we will use this all semester!)

- General interpretation:
  - On average, a one unit-increase in X is associated with a β_1 increase (or decrease) in the value of Y
- Interpretation from example above:
  - On average, a one-hour increase in hours worked per week (X) is associated with a \$5,000 (β_1) increase in annual income (Y)
- Interpret if $\beta_1 = \$2,000$; or $\beta_1 = \$4,000$

Population regression coefficient, β_1

Some important things to remember:

- If β_1 (i.e., the relationship between X and Y) is linear, then the average change in Y for a one-unit increase in X is the same no matter the starting value of X
    - Like plot example from earlier
- β_1 measures the **average** effect on Y for a one-unit increase in X; this effect on an individual observation may be different than this average effect!
- β_1 is a population parameter. We hardly ever know population parameters. So we **estimate** β_1 using sample data!

Population Intercept, β_0

What is the effect of hours worked per week (X) on annual income (Y)?

- $Y_i = \beta_0 + \beta_1 X_i + u_i$

β_0 is the "population intercept"

- $\beta_0 =$ the average value of Y when X=0
- Here, β_0, is the average annual income for someone that works zero hours per week (X=0)
- Usually, we are not substantively interested in β_0
- Sometimes β_0 is non-sensical or there's too few observations at X=0 to calculate a precise estimate (e.g., effect of age on income)

Population Linear Regression Line

Population Linear Regression *LINE*

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

We sometimes deconstruct the Population Linear Regression Model into two parts:

(1) **Population** Linear Regression *LINE*/ Regression Function: $Y_i = \beta_0 + \beta_1 X_i$
(2) **Population** "Error" or "Residual" Term: u_i

- Population regression line: just a linear prediction line, like the one in the scatterplot *if* the scatterplot contained all observation in the population
- Population regression line measures the "average" or "expected" relationship between X and Y, ignoring variables that we excluded from the model (i.e., u_i)

Population Linear Regression *LINE*

Population regression line and Expected Value, E(Y)

- Expected value of Y (for a sample mean/ one variable)
  - $E(Y) = \mu_Y$

- Expected value of Y, given the value of X (relationship between two variables)
  - $E(Y|X) = \beta_0 + \beta_1 X_i$
  - the population regression line is expected value of Y for a given value of X

- Population regression line and prediction
  - If we know value of parameters, β_0 and β_1, we can predict value of Y
  - Example: $\beta_0 = \$5,000$ and $\beta_1 = \$2,000$
  - (1) Predict the value of Y (annual income) for someone that works 20 hours per week
  - (2) Predict the value of Y (annual income) for someone that works 45 hours per week

u_i as "Error Term"

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= income; X_i= hours worked
- In causal inference research:
  - Error term u_i represents (consists of) *all other variables besides X that are not included in your model* that affect the dependent variable
  - In other words, the error term consists of all other factors (i.e., variables) responsible for the difference between the i^{th} district's average test score and the value predicted by the regression line
  - This interpretation will become *super* important down the road!
- Example of Y= income; X_i= hours worked; the error term u_i would consist of other factors besides hours worked that have an effect on yearly income!
  - Occupation: a hedge fund manager can 20 hours a week and make millions (maybe not last week tho!); an essential worker can 80+ hours and still only $40k
  - Race: BIPOC face discrimination in labor wages
  - Gender pay gap!
- In other social science based statistics classes
  - Interpret the u_i as the overall error in the prediction of Y due to *random variation*

u_i as "Residual"

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= income; X_i= hours worked
- u_i as the residual
  - Population regression line represents the predicted value of Y (income) for each value of X (hours worked)
  - Residual = the predicted value of Y - observed value of Y for any given value of X
- Easier to conceptually think about u_i in terms of each observation, i
  - Y_i = actual value of income for person i
  - $Y_i = \beta_0 + \beta_1 X_i$ = Population Regression line
    - The predicted value of income for person i with hours worked = X_i
  - Residual, u_i
    - The difference between actual value, Y_i, and predicted value from the population regression line for observation i
    - $u_i = Y_i - (\beta_0 + \beta_1 X_i)$

BREAK [5-10 min]

Estimating Regression Parameters

General things we do in regression analysis

1. **Estimation** [Today]
   - How do we choose estimates of β_0 and β_1 using sample data?
2. **Prediction** [Next Week]
   - What is the predicted value of Y for someone with a particular value of X?
3. **Hypothesis testing** [focus of the rest of the semester]
   - Hypothesis testing and confidence intervals about β_1

Step 1 of regression: Estimate Parameters

Population linear regression model

- $Y_i = \beta_0 + \beta_1 X_i + u_i$

Goal of estimation is to:

- Use sample data to estimate the population intercept, β_0, and the population regression coefficient, β_1
- $\hat{\beta}_0$ is an estimate of β_0
- $\hat{\beta}_1$ is an estimate of β_1
  - How dow we know these estimates are based on sample data and not population parameters? (hint: notation!)

Estimation problem:

Need to develop a method for choosing values of $\hat{\beta}_0$ and $\hat{\beta}_1$

Estimation (population mean)

We faced a similar estimation problem in intro to stats!

- Use sample to calculate the "best" estimate of the population mean, μ_Y
- We decided sample mean, $\bar{Y}$, was the "best" estimate!

Criteria we used to determine $\bar{Y}$ was "best" estimate of μ_Y

- m is all potential estimates for μ_Y
- Goal: choose the value, m , that minimizes the "sum of squares"
  - Sum of squares $= \sum_{i=1}^{n} (Y_i - m)^2$
  - $\bar{Y}$ is the value of m that minimizes sum of squares
  - So $\bar{Y}$ is the "least squares" estimator

Draw scatterplot:

- (1) Horizontal line representing sample mean
  - Show formula for sum of square errors

Estimation (regression)

Problem in regression:

- Need to develop method for selecting the "best" estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$
- Solution: similar to what we do for population mean!

First some terminology:

- Y_i is the actual observed value of Y for individual i
- $\hat{Y}_i$ is the predicted value of Y_i, based on sample data!
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Estimated residual, $\hat{u}_i$ is the difference between actual Y_i and predicted $\hat{Y}_i$
  - $Y_i - \hat{Y}_i = \hat{u}_i$
  - $Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{u}_i$
  - Residuals are sometimes called "errors"

Estimation (regression)

Criteria for choosing "best" estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$

- Select values that minimize "sum of squared residuals"

Sum of squared residuals (or sometimes called "sum of squared errors"):

- $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- $\sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$
- $\sum_{i=1}^{n} (u_i)^2$

Ordinary Least Squares is a linear method for estimating parameters in a linear regression model

- Method draws a line through the sample data points that minimizes the sum of squared residuals, or in other words, the differences between the observed values and the corresponding fitted values
- Minimization is achieved via calculus (derivatives). R will calculate this for you (phew!)
- Best estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are those that any other alternatives would result in a higher sum of squared residuals

OLS Prediction Line

Population Linear Regression Model

- $Y_i = \beta_0 + \beta_1 X_i + u_i$

OLS Prediction Line or "OLS Regression Line" (based on sample data)

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- Our OLS prediction line chose the best estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ as those that any other alternatives would result in a higher sum of squared residuals
- Draw this out...

Let's write run our first regression and write out our models!

RQ: What is the effect of hours worked per week on annual income?

- Run regression in R
  - `mod1 <- lm(realrinc ~ hrs1, data=gss)`
  - `summary(mod1)`

```
Call:
lm(formula = realrinc ~ hrs1, data = gss)

Residuals:
   Min     1Q Median     3Q    Max
-42321 -14391  -6999   5486 143635

Coefficients:
            Estimate Std. Error t value          Pr(>|t|)
(Intercept)  6960.84    2536.32   2.744           0.00615 **
hrs1          454.69      57.21   7.947 0.00000000000000445 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28140 on 1174 degrees of freedom
  (1172 observations deleted due to missingness)
Multiple R-squared:  0.05105,   Adjusted R-squared:  0.05024
F-statistic: 63.16 on 1 and 1174 DF,  p-value: 0.000000000000004447
```

$\hat{\beta}_0$

$\hat{\beta}_1$

Let's write run our first regression and write out our models!

RQ: What is the effect of hours worked per week on annual income?

Write out and label everything within the following: [we will be doing this all semester!]

1. Population regression model
   - Label Y; Label X
2. OLS Prediction Line (without estimates)
   - Define $\hat{\beta}_0$?
   - Define $\hat{\beta}_1$?
3. OLS Prediction Line (with estimates)
   - Interpret $\hat{\beta}_0$ given the estimate
   - Interpret $\hat{\beta}_1$ given the estimate
4. Predict the expected value of $\hat{Y}_i$ for someone that works 60 hours a week.

In-Class Group Exercise

RQ: What is the effect of age on annual income??

- hint! X = `age` and Y = `realrinc`

Write out and label everything within the following

- Recommendation: Practice how to write out equations in Word; touch-screen devices share your screen via whiteboard

1. Population regression model
   - Label Y; Label X
2. OLS Prediction Line (without estimates)
   - Define $\hat{\beta}_0$?
   - Define $\hat{\beta}_1$?
3. OLS Prediction Line (with estimates)
   - Run regression in R and print to get estimates:
     - `mod2 <-  lm(realrinc ~ age, data=gss)`
     - `summary(mod2)`
   - Interpret $\hat{\beta}_0$ given the estimate
   - Interpret $\hat{\beta}_1$ given the estimate
4. Predict the expected value of $\hat{Y}_i$ for someone that is 18 years old.

In-Class Group Exercise [Solutions]

RQ: What is the effect of age on annual income??

1. Population regression model
   - $Y_i = \beta_0 + \beta_1 X_i + u_i$
   - Y = annual income; X = age
2. OLS Prediction Line (without estimates)
   - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
   - $\hat{\beta}_0$? = Sample population intercept
     - i.e., the average value of Y when X=0
   - $\hat{\beta}_1$ = Sample regression coefficient
     - i.e., the average change in Y for one-unit increase in X
3. OLS Prediction Line (with estimates)
   - $\hat{Y}_i = \$8,620 + \$368 X_i$
   - $\hat{\beta}_0$ = \$8,620
     - On average, someone who is age zero has an annual income of \$8,620
     - Example of non-sensical $\hat{\beta}_0$
   - $\hat{\beta}_1$? = \$368
     - On average, a one-year increase in age is associated with a \$368 increase in annual income
4. Predict the expected value of $\hat{Y}_i$ for someone that is 18 years old.
   - $E(\hat{Y}_i | X = 35) = \hat{Y}_i = \$8,620 + \$368 * 18$
   - $E(\hat{Y}_i | X = 35) = \hat{Y}_i = \$8,620 + \$6,624$
   - $E(\hat{Y}_i | X = 35)$ = \$15,244