Social Sciences Intro to Statistics

Week 7.1 Introduction to Bivariate Regression

Week 7: Learning goal - Demonstrate estimation and prediction of bivariate regression analysis in R.

Introduction

Lecture overview:

- Introduction to Bivariate Regression
- Scatterplot, Covariance, Correlation
- Population linear regression model

Introduction to Bivariate Regression

"Bivariate regression" refers to regression models with two variables, a Y variable ("dependent variable" or "outcome") and a single X variable ("independent variable")

"Multivariate regression" refers to regression models with a Y variable and two or more X variables, this we will cover later in the course in week 10.

Today we will start with the fundamental concepts of bivariate regression. All of these concepts will be the same when we move on to multivariate regression in later weeks.

Libraries, data, functions

Load packages:

```
library(tidyverse)
library(ggplot2)
library(haven)

load(url('https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/main/data/els/output_

# ELS data frames
els <- df_els_stu_allobs_fac
```

Bivariate Regression

“Bivariate regression” refers to regression models with two variables, a Y variable (“dependent variable” or “outcome”) and a single X variable (“independent variable”).

“Multivariate regression” refers to regression models with a Y variable and two or more X variables

This lecture – which we will teach over several weeks – teaches the fundamental concepts of bivariate regression. All of these concepts will be similar when we move on to multivariate regression.

There are many ways to investigate this relationship between X and Y:

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

Relationships between two continuous variables

Postive relationship, negative relationship, and no relationships

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 (X) studying and GPA (Y)
- e.g., cost of attendance (X) and student debt (Y)

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 absences and GPA

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 (defined as "research" or "master's" universities by the Carnegie Classification that are

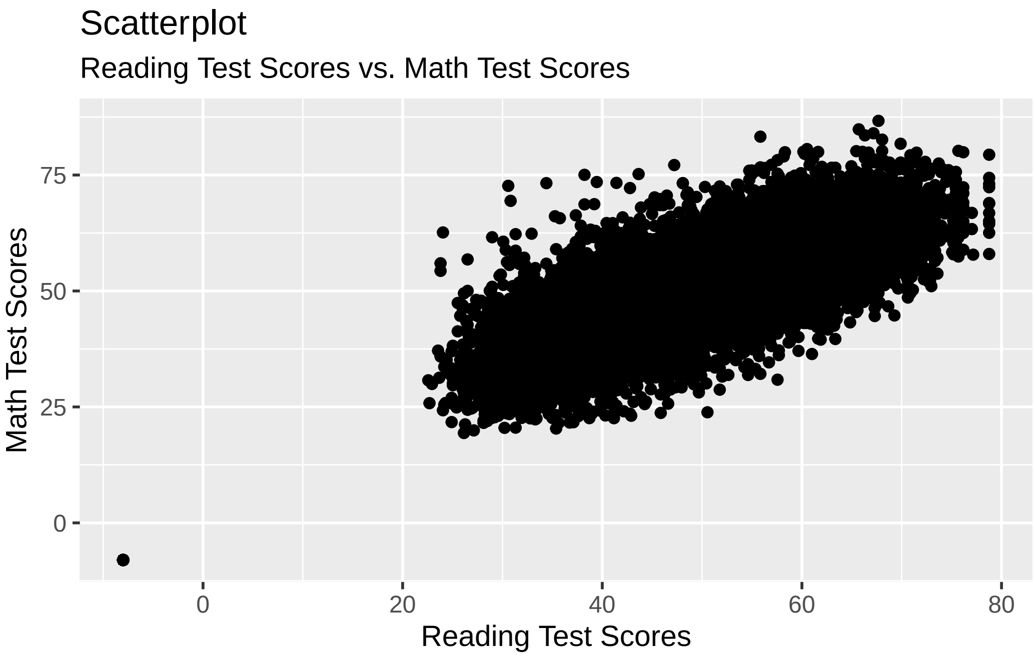
Scatterplot

Scatterplots will plot individual observations on an X and Y axis

We will use the data frame `els` to run regression models of the relationship between reading test standardized scores `bytxrsd` and math test standardized scores `bytxmstd`.

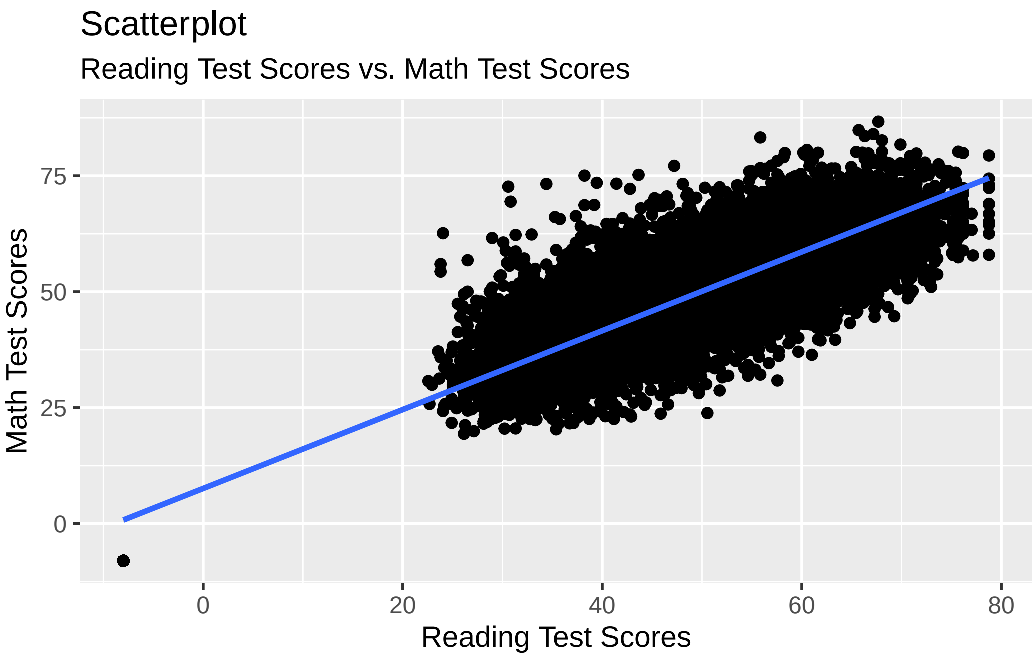
Draw scatterplot of X (`bytxrsd`) and Y (`bytxmstd`)

```
els_read_math <- els %>% ggplot(aes(x = bytxrstd, y = bytxmstd)) + geom_point() +
  labs(subtitle="Reading Test Scores vs. Math Test Scores",
       y="Math Test Scores",
       x="Reading Test Scores",
       title="Scatterplot")

els_read_math
```



Create scatterplot with "prediction" line

```r
els_read_math <- els %>% ggplot(aes(x = bytxrstd, y = bytxmstd)) + geom_point() +
  labs(subtitle="Reading Test Scores vs. Math Test Scores",
       y="Math Test Scores",
       x="Reading Test Scores",
       title="Scatterplot") +
  stat_smooth(method = 'lm')

els_read_math
```



Residual

Residual is the difference between actual observed value of Y and predicted value of Y (given X). The predicted value of Y for a given value of X is represented by the "prediction line" above

Covariance

Covariance measures the extent to which two variables move together

- If math test score is "high" when reading test score is "high", then covariance is positive
  - ("high" means a value that is higher than the mean value for the variable)

- If math test score is “low” when reading test score is “high”, then covariance is negative
  - (low” means a value that is lower than the mean value for the variable)

Population covariance, denoted $cov(X,Y)$ or σ_{XY}

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

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

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

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

$\hat{\sigma}_{XY} = s_{XY}$ is the sum of these 20 calculations divided by 19 (n-1)

$\hat{\sigma}_{XY} = s_{XY}$ 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 $Y_i < \bar{Y}$

$\hat{\sigma}_{XY} = s_{XY}$ 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 $Y_i > \bar{Y}$

```r
cov(els$bytxrstd, els$bytxmstd, use = 'complete.obs') # positive covariance
#> [1] 135.301
```

Correlation

Problem with sample covariance, s_{XY}

- The value of covariance dependes on the units of measurement of the underlying variable
- We can't compare the covaraince of X and Y vs covariance of X and Z

which covariance is larger?

- covariance between reading test score and math test score? or covariance between reading test score and hours spend each week on homework in school?

```
cov(els$bytxrstd, els$bytxmstd, use = 'complete.obs')
#> [1] 135.301
cov(els$bytxrstd, els$bys34a, use = 'complete.obs')
#> [1] 17.66619
```

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...

- correlation between reading test score and math test score is stronger than the correlation between reading test score and hours spend each week on homework in school.

```
cor(els$bytxrstd, els$bytxmstd, use = 'complete.obs')
#> [1] 0.847582
cor(els$bytxrstd, els$bys34a, use = 'complete.obs')
#> [1] 0.2387056
```

Linear vs Non-Linear Relationships

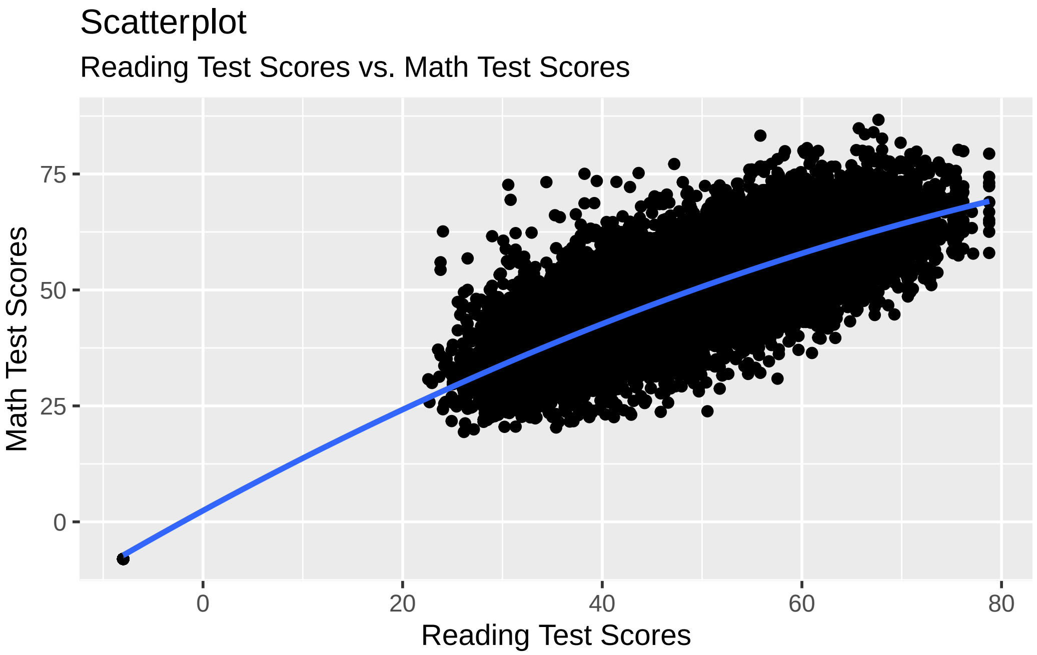
Problem with covariance and correlation:

- Both measure linear relationships, defined as relationships between two variables that are captured by a straight line;
- These measures do not detect non-linear relationships, defined as relationships between two variables that are not captured by a straight line

Below, scatterplot of relationship between cost of attendance (X) and debt (Y) with a linear prediction line fitted

- this is the relationship captured by covariance and correlation

```
els %>% ggplot(aes(x = bytxrstd, y = bytxmstd)) +
  geom_point() +
    labs(subtitle="Reading Test Scores vs. Math Test Scores",
       y="Math Test Scores",
       x="Reading Test Scores",
       title="Scatterplot") +
  stat_smooth(method = 'lm', formula = y ~ x + I(x^2), size = 1)
```



Population linear regression model

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?

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 causal effect on student achievement

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
- **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**
  - We will focus on modeling linear relationship between X and Y
  - we will cover non-linear relationships at end of quarter if we have time

Population Linear Regression Model

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

Sometimes, we deconstruct the Population Linear Regression Model into two parts:

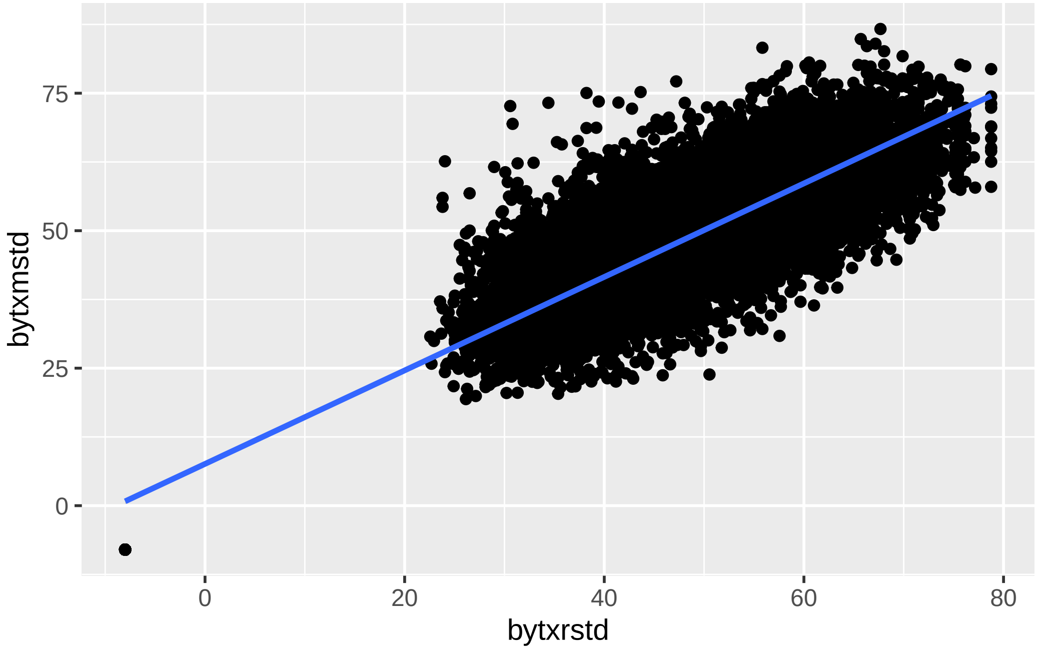
(1) **Population** Linear Regression *LINE*: $Y_i = \beta_0 + \beta_1 X_i$

- sometimes called population linear regression function

(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)

```
els %>% ggplot(aes(x = bytxrstd, y = bytxmstd)) +
  geom_point() + stat_smooth(method = 'lm')
```



Population regression line and Expected Value, E(Y)

- Expected value of Y (one variable)
  - $E(Y) = \mu_Y$
  - expected value of variable Y equals the population mean value of variable Y
- Expected value of Y, given the value of X (conditional expectation; 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

```r
# Fit the linear model
model <- lm(bytxmstd ~ bytxrstd, data = els)

# Extract the coefficients
coefficients <- coef(model)
beta0 <- coefficients[1]  # Intercept
beta1 <- coefficients[2]  # Slope

# Print the coefficients
beta0
#> (Intercept)
#>    7.592331
beta1
#>  bytxrstd
#> 0.8500363
```

- Population regression line and prediction
  - If we know value of parameters, β_0 and β_1, we can predict value of Y
  - Example: imagine $\beta_0 = 7.59$ and $\beta_1 = 0.85$
  - (1) Predict the value of Y (math reading score) for student whose reading test score is 40
    - $E(Y|X) = \beta_0 + \beta_1 X_i = 7.59 + 0.85 \times 40 = 41.59$
  - (2) Predict the value of Y (math reading score) for student whose reading test score is 72
    - $E(Y|X) = \beta_0 + \beta_1 X_i = 7.59 + 0.85 \times 72 = 68.79$

u_i

u_i as "Error Term" or "unobserved variables"

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= institution-level student debt; X_i= cost of attendance
- 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), aside from X= cost of attendance, responsible for the difference between the actual institution-level student debt at university i and the value for university i that is predicted by the regression line
  - This interpretation will become *super* important down the road!

- Example of Y = math test score; X_i= reading test score; the error term u_i would consist of other factors besides reading test score that have an effect on the math test score.
  - how many math classes have students taken, how many hours they spend on homework for their math classes, past math test scores

- In other social science based statistics classes (outside of the discipline of economics)
  - Interpret the u_i as the overall error in the prediction of Y due to *random variation*

Thinking about u_i as "Residual"

- Population linear regression model
  - $Y_i = \beta_0 + \beta_1 X_i + u_i$
  - Y= math test score; X_i= reading test score

- u_i as the residual
  - Population regression line represents the predicted value $\hat{Y}$ of the outcome (math test score) for each value of X (reading test score)
  - Residual = the predicted value of Y (for a given value of X) minus the observed value of Y

- Easier to conceptually think about u_i in terms of each observation, i
  - Y_i = actual value of institution-level student debt for person i
  - $\hat{Y}_i = \beta_0 + \beta_1 X_i$ = Population Regression line
    - The predicted value of math test score for student i with reading test score = X_i
  - Residual, u_i
    - The difference between actual value, Y_i, and predicted value from the population regression line for observation i
    - $u_i = Y_i - (\beta_0 + \beta_1 X_i) = Y_i - \hat{Y}_i$
    - $u_i = Y_i - \hat{Y}_i$