

Bivariate Regression I: Conceptual Overview and Estimation

Dr. Michael Fix
mfix@gsu.edu

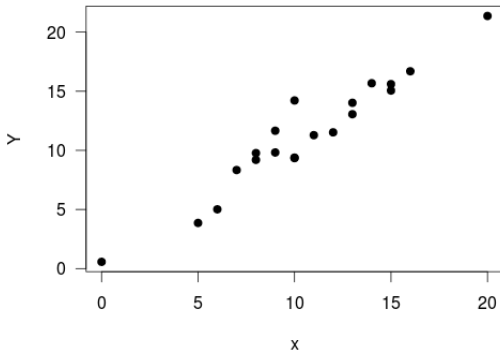
Georgia State University

1 February 2024

Note: The slides are distributed for use by students in POLS 8810. Please do not reproduce or redistribute these slides to others without express permission from Dr. Fix.

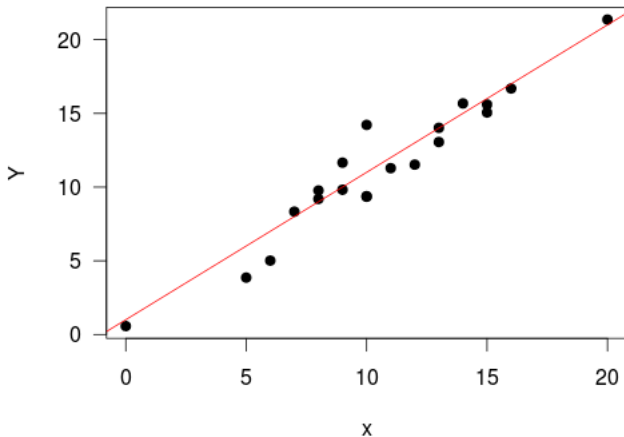
Fundamentals of Regression

- Regression involves the relationship between two (or more) variables:
 - The dependent variable (regressand/response): Y
 - The independent variable (regressor/factor): X
- Graphically, we can represent this with a scatter plot:



Fundamentals of Regression

- Intuitively, we see a line that can be drawn
- How do we get the best line?



Fundamentals of Regression

Least Squares

- The goal is to find a predicted value for Y represented by \hat{Y}
- We want to find a line with the basic formula: $\hat{Y} = a + bX$
- Our goal is a line that is the closest to all of the points
- To do this we want to minimize deviation: $d = Y - \hat{Y}$
- Sum this to get the whole and use the square to remove the problem of negatives:

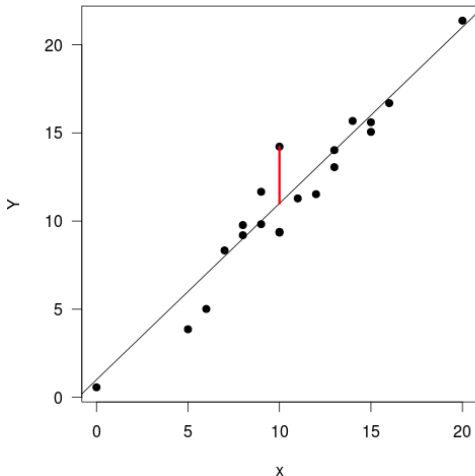
$$\sum d^2 = \sum (Y - \bar{Y})^2 \quad (1)$$

- This method is known as *Ordinary Least Squares (OLS)*

Fundamentals of Regression

Least Squares

- Conceptually we can represent this in graphical form.



Formula for Regression Line

- We need to find the formula for the line that minimizes the sum of squared errors

$$\hat{Y} = a + bX \quad (2)$$

- b indicates the slope of the line
 - This value provides substantive information
 - The change in Y for each unit increase in X
- a indicates the y -intercept of the line
 - This is the value of Y when $X = 0$

Computing OLS Estimates

- b can be calculated from the deviations of X and Y from their respective means:

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} \quad (3)$$

- a is found by solving equation (2) to get:

$$a = \bar{Y} - b\bar{X} \quad (4)$$

Computing OLS Estimates in R

- OLS is computationally simple enough that in the bivariate case, with a small N , we can hand calculate our estimates
- However, we do not generally do this as it is inefficient and doesn't scale up well


```
### Load necessary packages ----
# Use install.packages() if you do not have this package
library(tidyverse) # Data manipulation
library(stargazer) # Creates nice regression output tables

### Load your data ----
# We are using V-Dem version 12
my_data <- readRDS("data/vdem12.rds")

# Let's change names of some of these variables for the sake of simplicity
my_data <- my_data |>
  rename(democracy = v2x_polyarchy, gdp_per_capita = e_gdppc)

### Run a bivariate OLS ----
# We are going to use lm() function (which means linear model).
# Always check function help page!
?lm
help(lm)

# Here is how you specify your variables:
# lm(dependent_variable ~ independent_variable(s), data = your_data)
# ~ => this is tilda

# For example:
lm(democracy ~ gdp_per_capita, data = my_data)
```

Regression Output

```
lm(democracy ~ gdp_per_capita, data = my_data)

# This produces very little info, so we save this output as a list object and then examine it:

my_lm <- lm(democracy ~ gdp_per_capita, data = my_data) # creates a list object called my_lm

summary(my_lm) # gives more detailed output
```

```
> # For example:
> lm(democracy ~ gdp_per_capita, data = my_data)

Call:
lm(formula = democracy ~ gdp_per_capita, data = my_data)

Coefficients:
(Intercept)  gdp_per_capita
    0.2158      0.0117

> summary(my_lm) # gives more detailed output

Call:
lm(formula = democracy ~ gdp_per_capita, data = my_data)

Residuals:
    Min       1Q   Median       3Q      Max
-2.03380 -0.16797 -0.05647  0.14826  0.58390

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2158381  0.0018741  115.17  <2e-16 ***
gdp_per_capita 0.0117026  0.0001469   79.68  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2348 on 21377 degrees of freedom
(6001 observations deleted due to missingness)
Multiple R-squared:  0.229,    Adjusted R-squared:  0.229
F-statistic: 6349 on 1 and 21377 DF, p-value: < 2.2e-16
```

Better Regression Output using stargazer()

```
### Stargazer package ----  
# Let's create better looking output using stargazer function  
stargazer(my_lm, type = "text") # Change type to latex if you're importing to LaTeX  
  
# Let's make it much better and export it to latex!  
stargazer(my_lm,  
  type = "latex",  
  title = "The relationship between democracy and GDP per capita",  
  covariate.labels = c("GDP per capita"),  
  dep.var.labels = c("Electoral Democracy Index"),  
  ci.level = 0.95,  
  star.cutoffs = c(0.05),  
  notes.align = "l",  
  notes.append = FALSE,  
  notes.label = "Notes",  
  notes = "*p < 0.05. Standard errors are in parentheses.")
```

Better Regression Output using stargazer()

Table 1: The relationship between democracy and GDP per capita

	<i>Dependent variable:</i>
	Electoral Democracy Index
GDP per capita	0.012* (0.0001)
Constant	0.216* (0.002)
Observations	21,379
R ²	0.229
Adjusted R ²	0.229
Residual Std. Error	0.235 (df = 21377)
F Statistic	6,349.082* (df = 1; 21377)

Notes

* $p < 0.05$. Standard errors are in parentheses.

Why regression?

	Description	Explanation	Prediction
Task	Summarize data	Correlation/causation	Forecast OOS / future data
Emphasis	Data	Theory / Hypotheses	Outcomes
Focus	Univariate	Multivariate	Multivariate
Typical Application	Summarize / "reduce" data	Discuss marginal associations between predictors and an outcome of interest	Optimize out-of-sample predictive power / minimize prediction error

Where Do We Go From Here?

- How to use OLS for hypothesis testing
- Assumptions of the OLS Estimator
- Model fit
- Beyond the bivariate case

What Won't We Do?

- Multiple Regression
- Measurement models
- Time series
- Machine Learning