Note: Annotations start at slide 12

Introduction to Econometrics: Linear Regression

Hammad Shaikh

January 18, 2018

Regression Overview

- Empirical analysis in economics is to provide precise quantitative answers to questions of economic interest
  - What is the effect of reducing class size on test scores?

- Economic model relates economic variables of interest to one another using a equation
  - Achievement $=$ f(effort, class size, parental investment)

- Econometric model completes an economic model by specifying any additional uncertainty
  - Achievement $=$ f(effort, class size, parental investment, ϵ)

Linear regression model

- Y = dependant / outcome / response variable
  - What are plausible Y's in class size reduction policy?

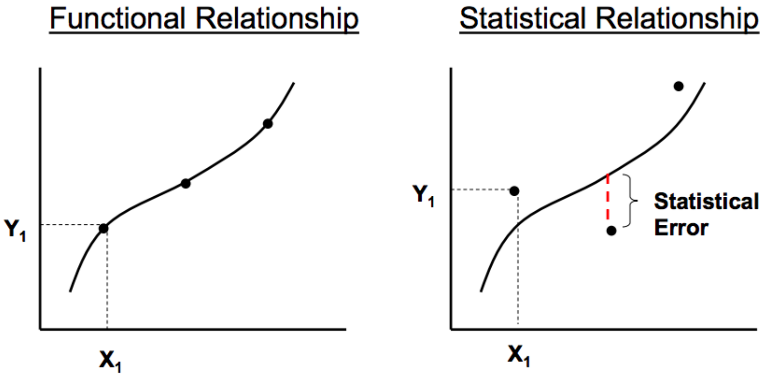
- X = independent / explanatory / predictor variable
  - Contains treatment of interest and other factors that effect Y
  - What are the X's in class size reduction policy?

- Simple regression: $Y = \beta_0 + \beta_1 X + \epsilon$

- Multiple regression: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$

Functional vs. Statistical Relationship

- Regression model describes the statistical relationship between outcome Y and response variable(s) X



Relationship Between X and Y

- The covariance is a measure of the linear association between X (class size) and Y (test score)
    - $S_{xy} = \widehat{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
    - Units are Units of X $\times$ Units of Y (No. of students $\times$ Score)

- Cov(X,Y) > 0 means a positive relation between X and Y

- Correlation is a unit less measure of the strength of linear relationship between X and Y
    - $\rho_{xy} = \frac{S_{xy}}{S_x S_y}$ is a number between -1 and 1
    - $\rho_{xy} = 1$ means perfect positive linear relationship

Simple Regression Example

- Question: What is the relationship between class size and test scores in California?

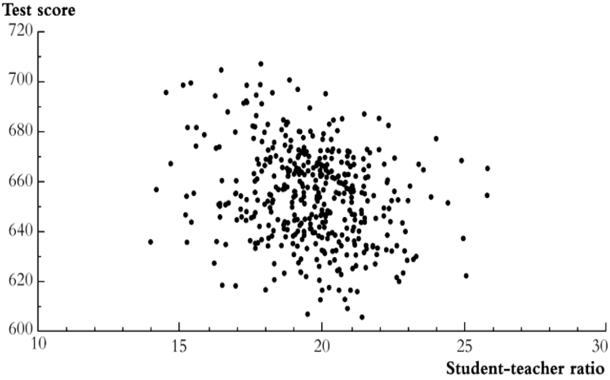
- Data available from 420 California school districts
  - 5th grade district average math and reading score
  - Student to Teacher Ratio (STR): number of students divided by number of teachers (within district)

- What is the regression model of interest?

Test Score and Student to Teacher Ratio

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.

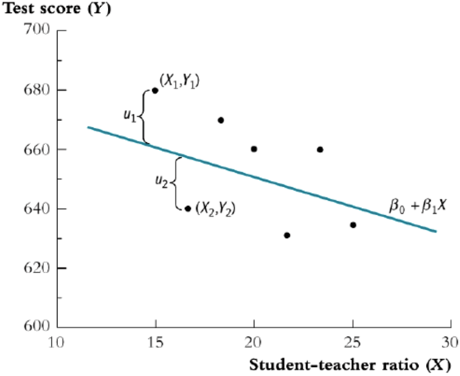


- We want to model above relationship with a simple linear regression

Estimating Simple Regression

FIGURE 4.1 Scatter Plot of Test Score vs. Student-Teacher Ratio (Hypothetical Data)

The scatterplot shows hypothetical observations for seven school districts. The population regression line is $\beta_0 + \beta_1 X$. The vertical distance from the i^{th} point to the population regression line is $Y_i - (\beta_0 + \beta_1 X_i)$, which is the population error term u_i for the i^{th} observation.



- Simple regression estimates: $\widehat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}$, $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$
    - Known as Ordinary Least Squares (OLS) estimator

Effect of STR on Achievement

- $TestScore_d = \beta_0 + \beta_1 STR_d + \epsilon_d$
  - We want to estimate $\beta_1 = \frac{\triangle TestScore}{\triangle STR}$. Interpret β_1?

- Line of best fit: $\widehat{TestScore}_d = \widehat{b}_0 + \widehat{b}_1 STR_d$
  - $(\widehat{b}_0, \widehat{b}_1)$ found by minimizing $\sum_{i=1}^{n}(TestScore_d - \widehat{TestScore}_d)^2$

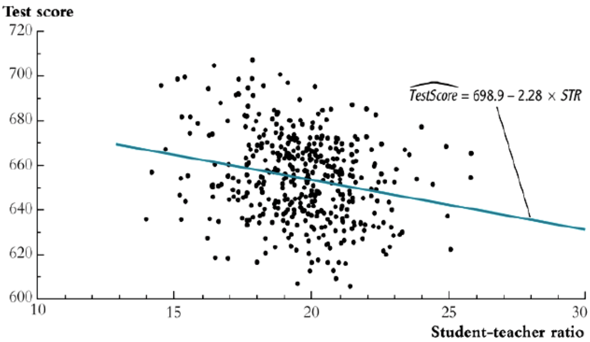
- $\widehat{b}_1 = \frac{\widehat{Cov}(TestScore_d, STR_d)}{\widehat{Var}(STR_d)}$ and $\widehat{b}_0 = \overline{TestScore} - \widehat{b}_1 \overline{STR}$

Effect of STR on Achievement Cont.

- Districts with larger class sizes (higher STR) are associated with lower test scores

FIGURE 4.3 **The Estimated Regression Line for the California Data**

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



Effect of STR on Achievement Cont.

- Estimated model: $\widehat{TestScore}_d = 698.9 - 2.28 STR_d$

- Primary estimate of interest is $\widehat{b}_1 = -2.28$
  - Districts with one more student per teacher on average are associated with 2.28 points lower test scores

- How to interpret intercept of $\widehat{b}_0 = 698.9$?

Properties of Slope Estimator

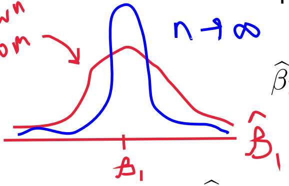
Recall $\begin{cases} \text{Rand. var } \hat{\beta}_1 \\ \text{Estimate } \hat{b}_1 \end{cases}$

- We generally want estimators to be unbiased and consistent
  - Slope estimator $\widehat{\beta}_1$ unbiased if $E(\widehat{\beta}_1) = \beta_1$
  - Slope estimator $\widehat{\beta}_1$ consistent if $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ as n grows large

$\epsilon \perp X$ is assumption discussed later

- It can shown (using CLT) that if $\underbrace{\epsilon \text{ independent of } X}$ then the distribution of slope estimator $\widehat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}$ is:

$\hat{b}_1$ drawn from



$$\widehat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$$

Notice: $n \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}_{n\sigma_X^2}$

- Show that $\widehat{\beta}_1$ is unbiased and consistent

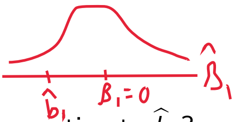
Unbiased: $E(\hat{\beta}_1) = \beta_1$
↳ mean of normal

Consistent: $Var(\hat{\beta}_1) \to 0$ as $n \to \infty$
so $\hat{\beta}_1 \to \beta_1$

Simple Linear Regression and Hypothesis Testing

- Simple linear regression: $TestScore_d = \beta_0 + \beta_1 STR_d + \epsilon_d$
  - β_0 (intercept) and β_1 (slope) are unknown parameters
  - Use sample $(STR_d, TestScore_d)_{d=1}^{n}$ to make inference about the simple linear regression parameters

Note: It is possible $\underbrace{\beta_1 = 0}_{\text{no effect}}$ but $\hat{b}_1 < 0$



- Question: How much can we trust the primary estimate $\widehat{b}_1$?

→ $\hat{b}_1 < 0$ is more trustable if $Var(\hat{\beta}_1)$ is small

→ Want to check $\hat{b}_1 < 0$ not just due to sampling error

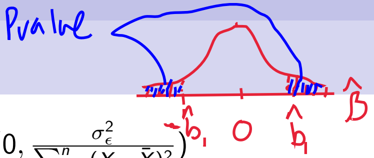
- Null Hypothesis: Class size has no effects on achievement

$H_0: \beta_1 = 0$

- Alternative Hypothesis: Class size effects achievement

$H_1: \beta_1 \neq 0$ ($\beta_1 < 0$ or $\beta_1 > 0$)

SLR and Hypothesis Testing Cont.



- Under $H_0 : \beta_1 = 0$ we have $\widehat{\beta}_1 \sim N(0, \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$
  - Since ϵ unknown, σ_ϵ^2 is also unknown. The solution is to replace it with s_e^2, the sample variance of the residuals

$s_e^2 = \frac{1}{n}\sum_i e_i^2$, e_i is residual, $e_i = y_i - \hat{y}_i$

- If $H_1 : \beta_1 \neq 0$ we compute p-value $= 2 * Pr(\widehat{\beta}_1 > \widehat{b}_1)$
  - $\widehat{b}_1$ is very significant if p-value < 0.01, significant if p-value < 0.05, and marginally significant if p-value < 0.1

- Computing p-value involves $SE(\widehat{b_1}) = \sqrt{Var(\widehat{\beta_1})}$
  - Typically (not always) if $|\frac{\widehat{b_1}}{SE(\widehat{b_1})}| > 2$ then $\widehat{b}_1$ is significant

$t_{stat} = \frac{\hat{b}_1 - 0}{SE(\hat{b}_1)} > 2 \Longrightarrow$ Pvalue < 0.05

- P-value ≈ 0 for class size application

$\rightarrow$ class size related to test scores

Fitness of Regression Model

- R^2 measures the proportion of variation in the outcome (Y) explained by the independent variable(s) (X)
  - R^2 is a number between 0 and 1 (R^2 is unitless)
  - $R^2 = 1$ means regression model perfectly fits the data

- $R^2 = \frac{SSR}{SST}$; SST = Sum of Square Total, SSR = Sum of Square Regression
  - $\underbrace{\text{SST}}_{\text{variation in } y} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $\underbrace{\text{SSR}}_{\text{Var. in } y \text{ explained by reg. } \hat{y}} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$

- R^2 applies to both simple and multiple linear regression

Ex. $R^2 = 0.3$ for $\widehat{Mark}_i = 40 + 5\,LectureAttend_i$

→ 30% of variation in grades is explained by lecture attendance

Simple Linear Regression Summary

Population → $\underbrace{B_0, B_1}_{\text{parameters}}$

Sample → $\underbrace{\hat{b}_0, \hat{b}_1}_{\text{Estimates}}$

- The population linear regression model
  - $Y = \beta_0 + \beta_1 X + \epsilon$

  B1 is param. of interest

- Line of best fit and OLS estimator
  - $\underbrace{\widehat{\beta}_1}_{\text{Slope Estimator}} = \frac{Cov(X,Y)}{Var(X)}$ and $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$

- Hypothesis testing
  - $\underbrace{H_0 : \beta_1 = 0}_{\text{X and Y not related}}$ and $H_1 : \beta_1 \neq 0$

- Measures of fit for simple regression: $\underbrace{\widehat{y} = \widehat{b}_0 + \widehat{b}_1 x}_{\text{Line of best fit}}$
  - Correlation and R^2

Extending to Multiple Regression

$Y = \beta_0 + \beta_1 X + \varepsilon$

- Results from simple linear regression are usually not causal
  - Many other factors that affect both X and Y are not accounted for in the model
    - Can bias slope estimates (omitted variable bias)

ε related to X is a problem

- Returns to education: $AdultIncome_i = \beta_0 + \beta_1 YrsEduc_i + \epsilon_i$
  - What are some variables in ϵ_i that may bias $\widehat{b}_1$?

↑ Parent Inc $\left\{\begin{array}{l}\uparrow \text{Yrs Educ} \\ \uparrow \text{Income}\end{array}\right.$, $\underbrace{\uparrow \text{ability}}_{\text{Hard to measure}}$ $\left\{\begin{array}{l}\uparrow \text{Yrs Educ} \\ \uparrow \text{Income}\end{array}\right.$

- Two solutions to help obtain causal result:
  - $\underbrace{\text{1) Randomized control trial}}_{\text{Best method to get causality}}$, or 2) Multiple regression

Randomized Control Trial (RCT)

↑ X_i not related to ε_i

RCT is gold standard

- Simple regression model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

We want $\varepsilon \perp X$

Sample

Control

Treatment

- In a RCT the Xs are randomly assigned to individuals
  - No omitted variable bias since X_i independent to ϵ_i
    - Now $\widehat{b}_1$ has a causal interpretation

→ Only diff. b/w control & Treatment is X

- Correlation does not imply causation?
  - Generally true for observational data, but false for experimental data where treatment variable is randomly assigned

Problem ε_i related to Yrs Educ

- Returns of education: $Y_i = \beta_0 + \beta_1 YrsEduc_i + \epsilon_i$
  - Can we randomly assign years of education to individuals?

→ Not ethical, so no

Multiple Regression

- Slope estimate in simple regression can be biased from omitted variables related to X and Y
  - Solution is to include the omitted variables into the model

→ Compare people with diff. X_1, but same $X_2, X_3, \ldots, X_k$

- Multiple regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$
  - β_1 = effect of changing X_1 on Y holding $X_2, \ldots, X_k$ constant
  - $\widehat{b}_1$ can be causal if all relevant variables are included
    - Conditional independence: ϵ indep. to X_1 given $X_2, \ldots, X_k$

→ Cond. indep: Given $X_2, \ldots, X_k$, ε unrelated to X_1

- Returns to education:
  $Y_i = \beta_0 + \beta_1 YrsEduc_i + \beta_2 Exp_i + \beta_3 ParentIncome_i + \epsilon_i$

→ Cond. indep: Yrs Educ indep. of ε given Exp & ParentIncome

↳ Most likely false as ability & motivation are still ommited

Regression Table Example

→ 4 regressions below
↳ Two Outcomes

→ ① $\hat{wage}_i = 11 + 2\,Yrs\,Educ_i$

Table: Income and Health Returns to Education (Fake Data)

	①	②	③	④
Y =	Hourly Wage	Hourly Wage	Years Lived	Years Lived
Constant ($\hat{b}_0$)	11***	10***	65***	66***
	(2.5)	(0.1)	(10)	(10)
Years of Educ ($\hat{b}_1$)	2***	1***	2***	3***
	(0.5) $se(\hat{b}_1)$	(0.1)	(0.25)	(0.3)
Experience		3***		0.5**
		(0.8)		(0.245)
Parent income ($1000)		0.1**		0.15*
		(0.048)		(0.075)
R-square	0.15	0.30	0.10	0.20
No. of indivisuals	15000	15000	15000	15000

X {Years of Educ, Experience, Parent income ($1000)}

Stars denote level of significance *10%,** 5%, and ***1%.

N=15'000 indiv. in data

- Regression table generally contain coefficient estimates, standard errors, no. of observations, and R^2

$\hat{b}_1$ in ④: People with an extra year of educ. are associated with living 3 yrs longer on avg. after controlling for work exp. and their parent income

Summary of Linear Regression

- Goal: examine causal relationship between outcome Y and explanatory variable X
- Simple linear regression is a good starting point
  - Slope estimate is likely biased due to omitted variables that effect both X and Y
- Experiments (RCTs) are ideal for determining causal relationship between X and Y
  - Costly and sometimes unfeasable
- Multiple regression can control for several relevant variables
  - Obtain causal relationship under conditional independance