Introduction to Simple Linear Regression

Hammad Shaikh

September 30, 2019

Regression Overview

X → Y

Policy variable Outcome

- 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?

X Y

- 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)

Stochastic error

Linear regression model

$X \longrightarrow Y$

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

↳ $Y \in$ {test score, Parent satisfaction}

- 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?

↳ $X \in$ {class size, Student-teacher ratio}

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

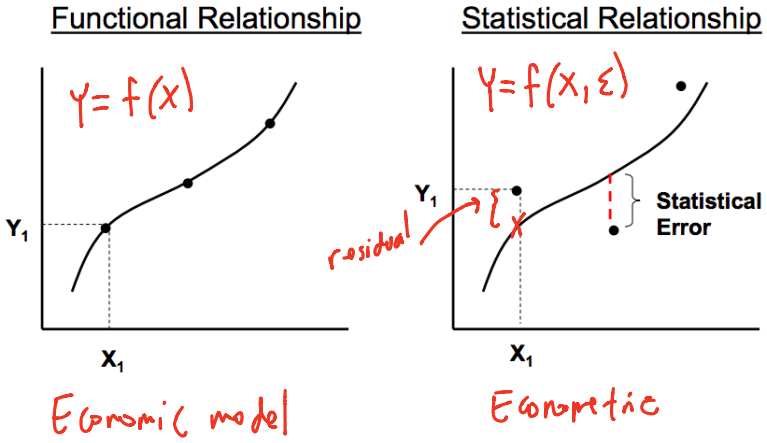
↳ i) Data (Y, X), ii) Parameters β_0, β_1, iii) error

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

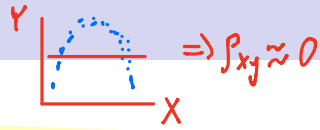
X_1: class size X_k: Hour study

Functional vs. Statistical Relationship

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



Relationship Between X and Y



$\Rightarrow \rho_{xy} \approx 0$

- 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)
    - ↳ Hard to interpret magnitude
- 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

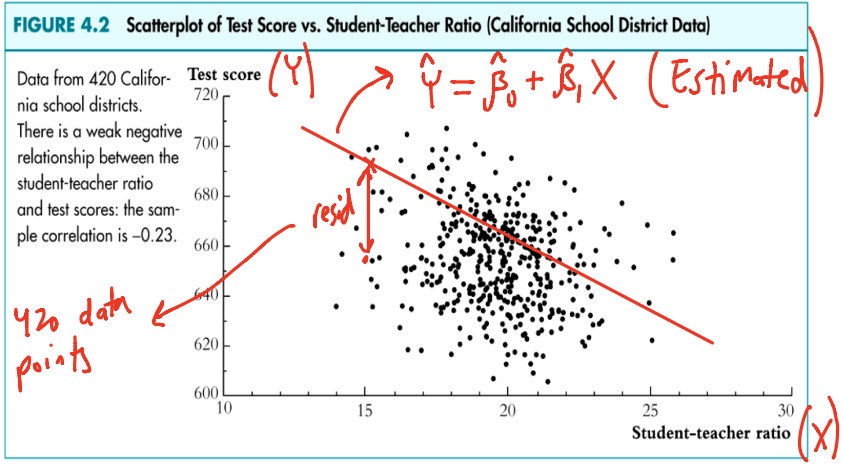
$$\text{units } Corr(X,Y) = \frac{\text{Units } X \cdot \text{Units } Y}{\text{Units } x \cdot \text{Units } Y}$$

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?

$$MathGS_d = \beta_0 + \beta_1 STR_d + \varepsilon_d \quad d \in \{1, 2, \ldots, 420\}$$

Test Score and Student to Teacher Ratio

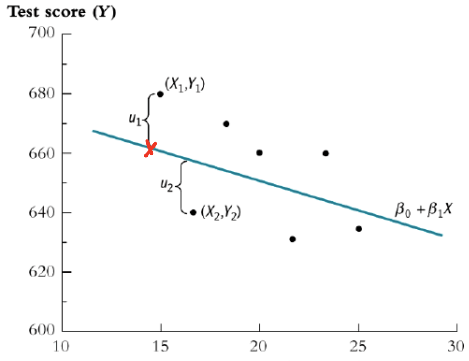


- We want to model above relationship with a simple linear regression

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.



$$\min_{\beta_0, \beta_1} \sum_{d=1}^{D} u_d(\beta_0, \beta_1)^2$$

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

Effect of STR on Achievement

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

↳ when STR ↑ by 1 the impact on test scores in β_1 points on avg.

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

↳ OLS

$e(\hat{b}_0, \hat{b}_1)$

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

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 $\widehat{b}_0 = 698.9$?

↳ STR = 0 is not in data and hence no interpretation.