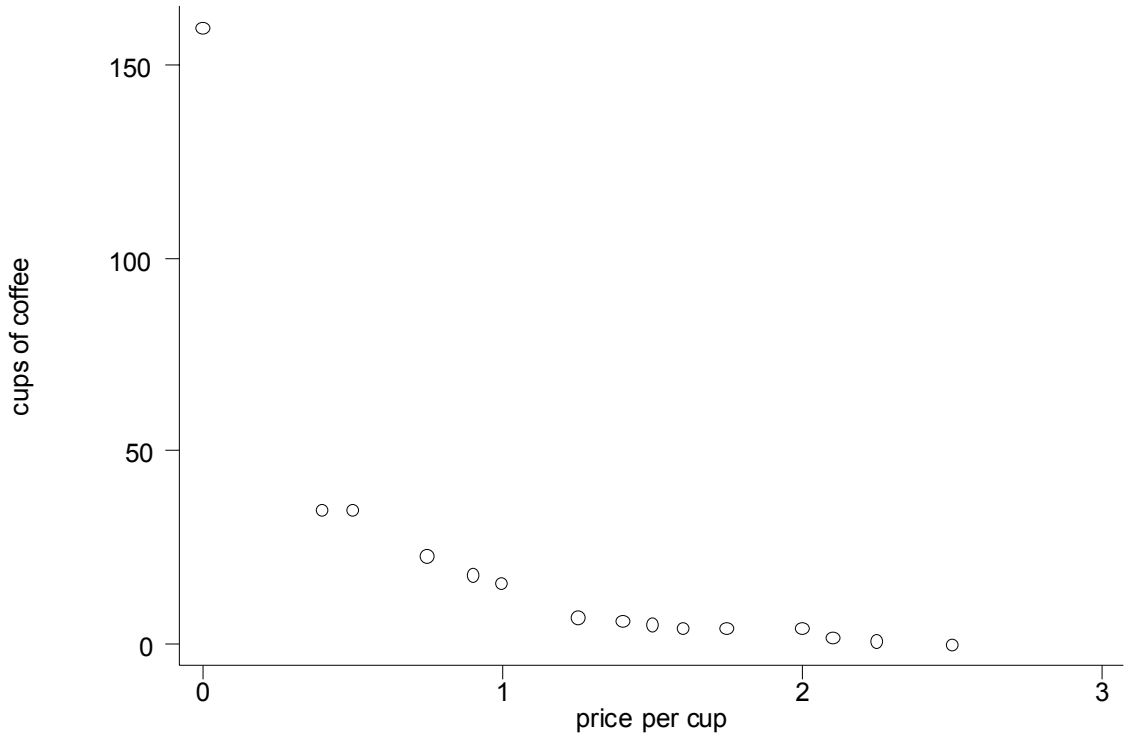
Section 3. Simple Regression (One Regressor)

1. **Introduction: Fitting a Line through a cloud**
2. **Coffee example**
3. **Global Warming example**
4. **CA test score example**
5. **Stata in action**
6. **Which line to choose?**
7. **What's to come?**
8. **Population Regression Line**

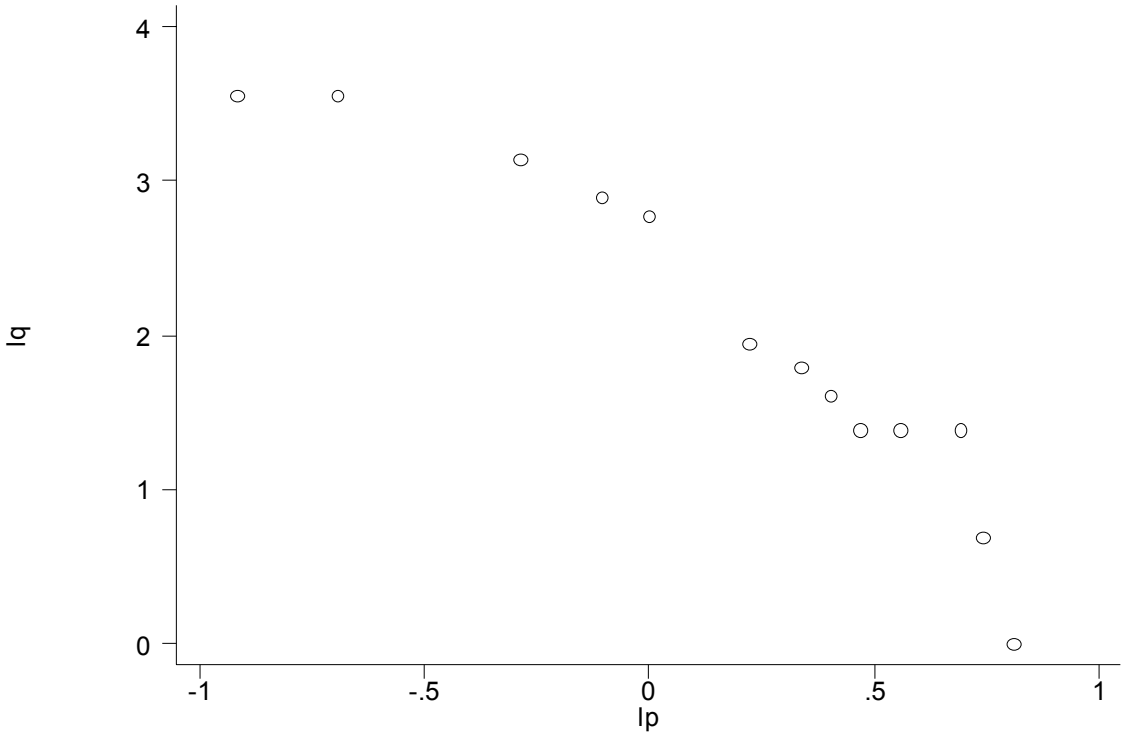
1. Introduction: Fitting a Line through a Cloud

- e.g. Demand for Coffee, Global warming,
  CA test scores and student-teacher ratios
- Why is drawing lines useful?
  - Describing data
  - Testing hypotheses
  - Prediction
- Which line to chose?
  - Minimizing "residual" or "error" terms u_i
- Note: slope and intercept are random variables
- Note: It's usually the population we care about

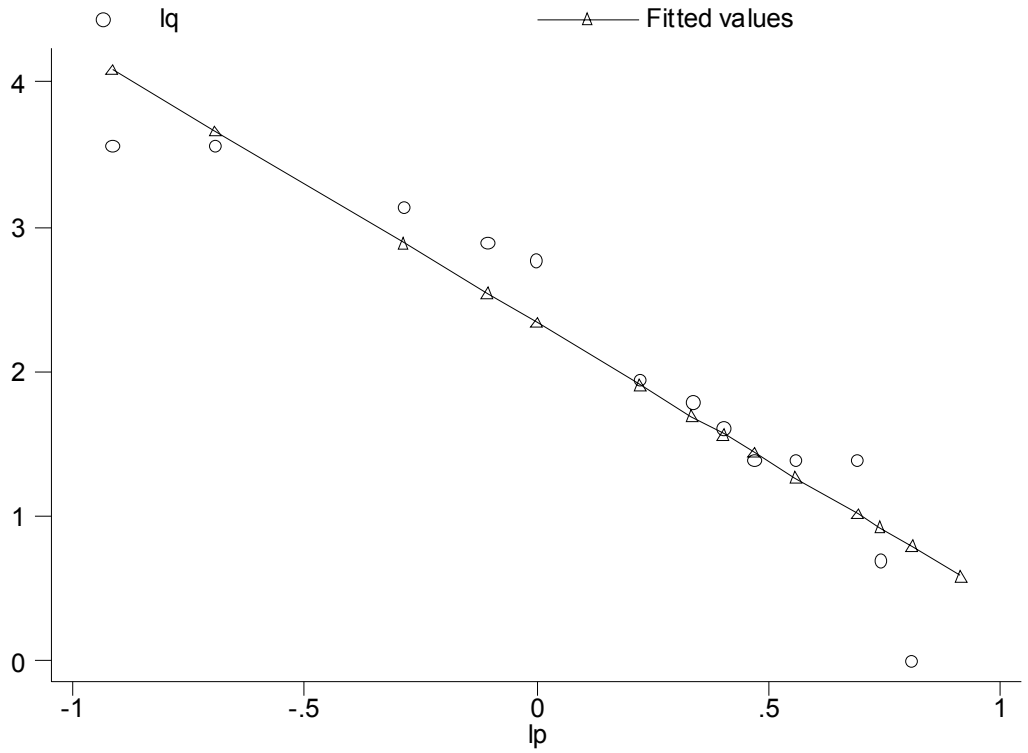
2. Coffee Demand again



2. Coffee Example (in logarithms)

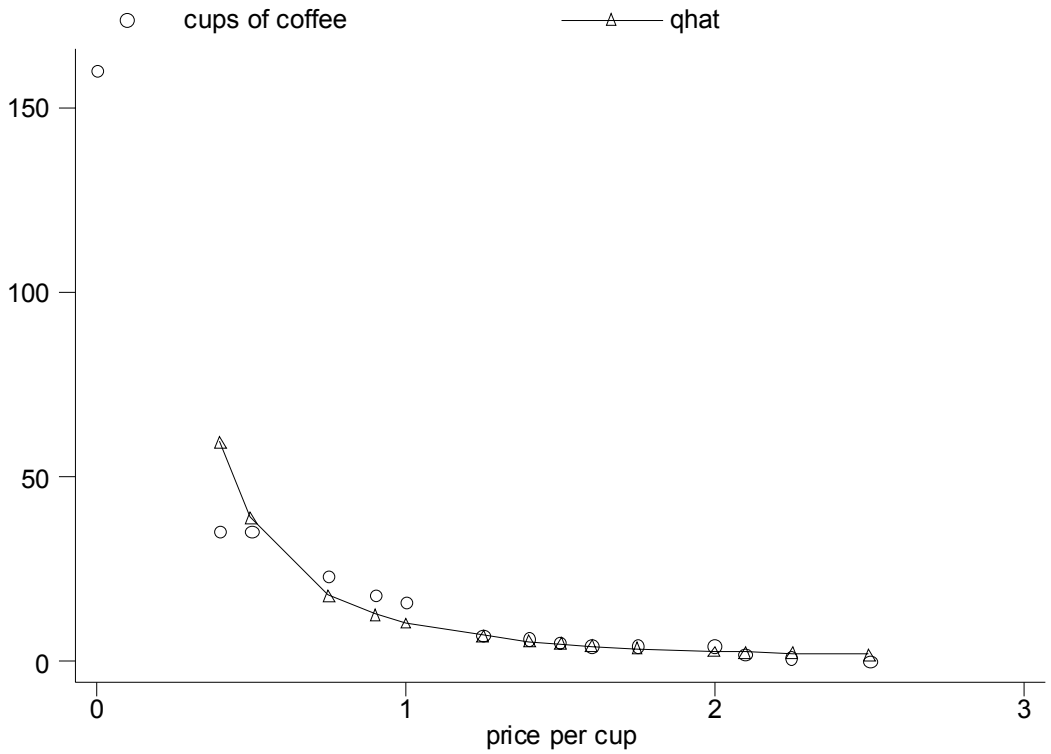


2. Coffee example (with a line)



What's the slope of the line?
About how many cups would you sell at $1?

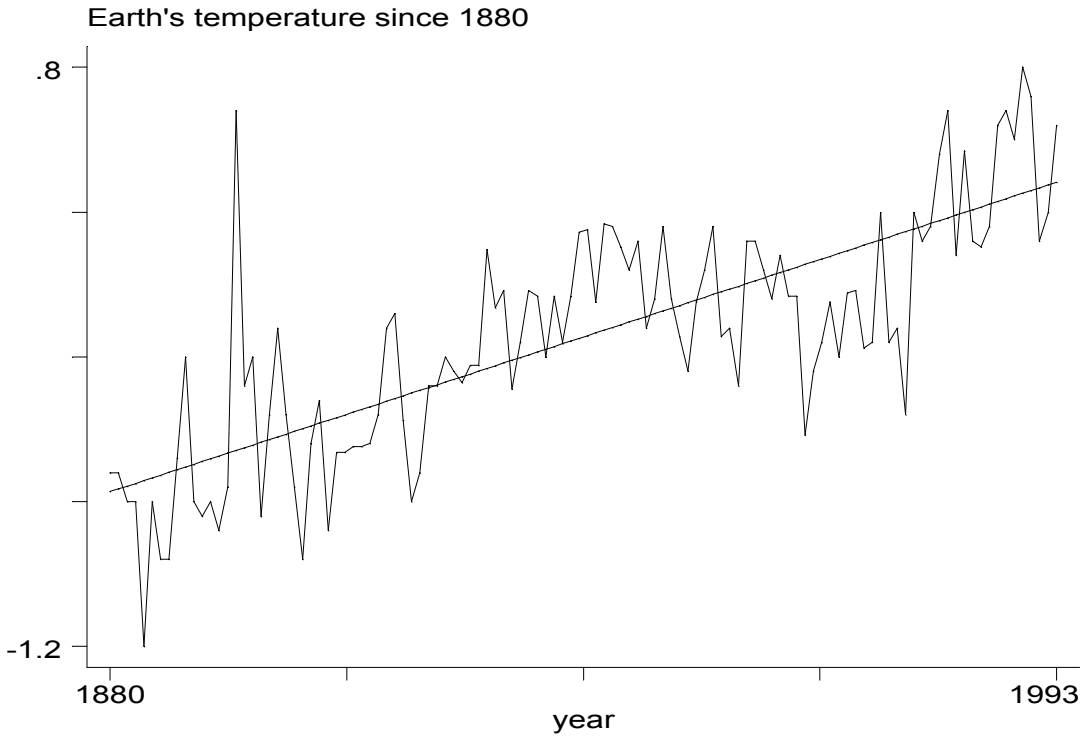
2. Coffee example (with a curve)



Predicted values qhat = exp(lqhat)

About how many cups would you sell at $1?

3. Global Warming Example



Is the slope statistically different from zero?

4. CA Test Score Example

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.

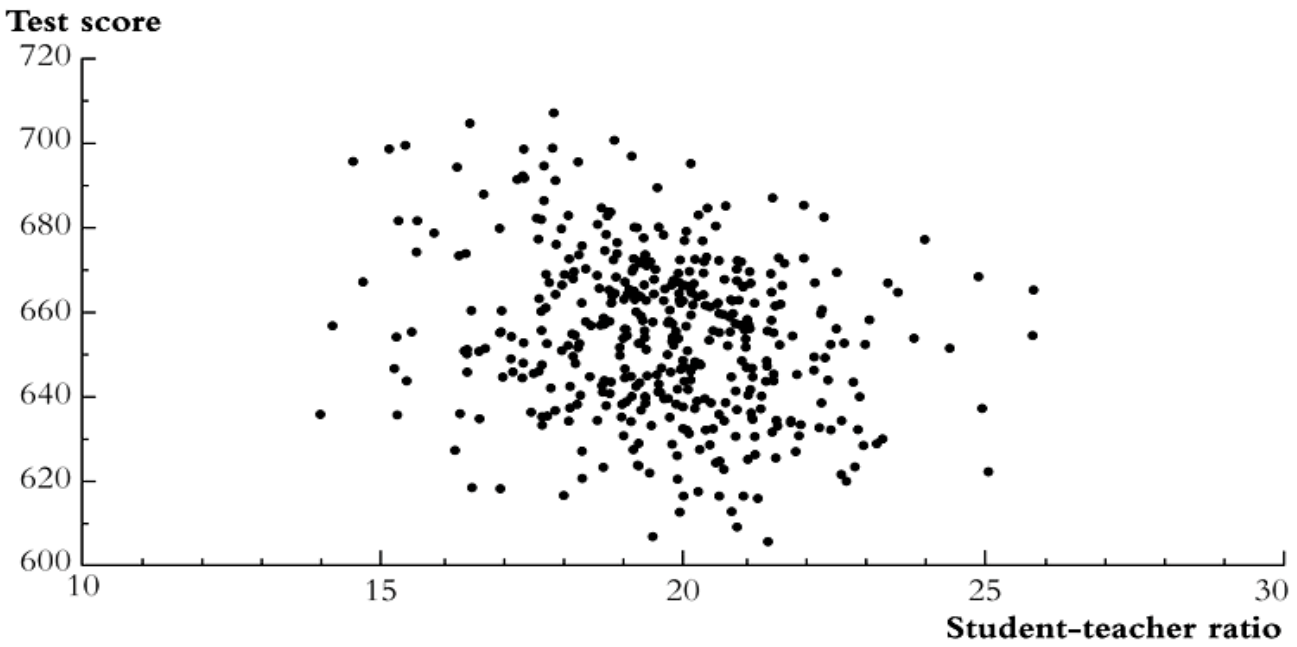
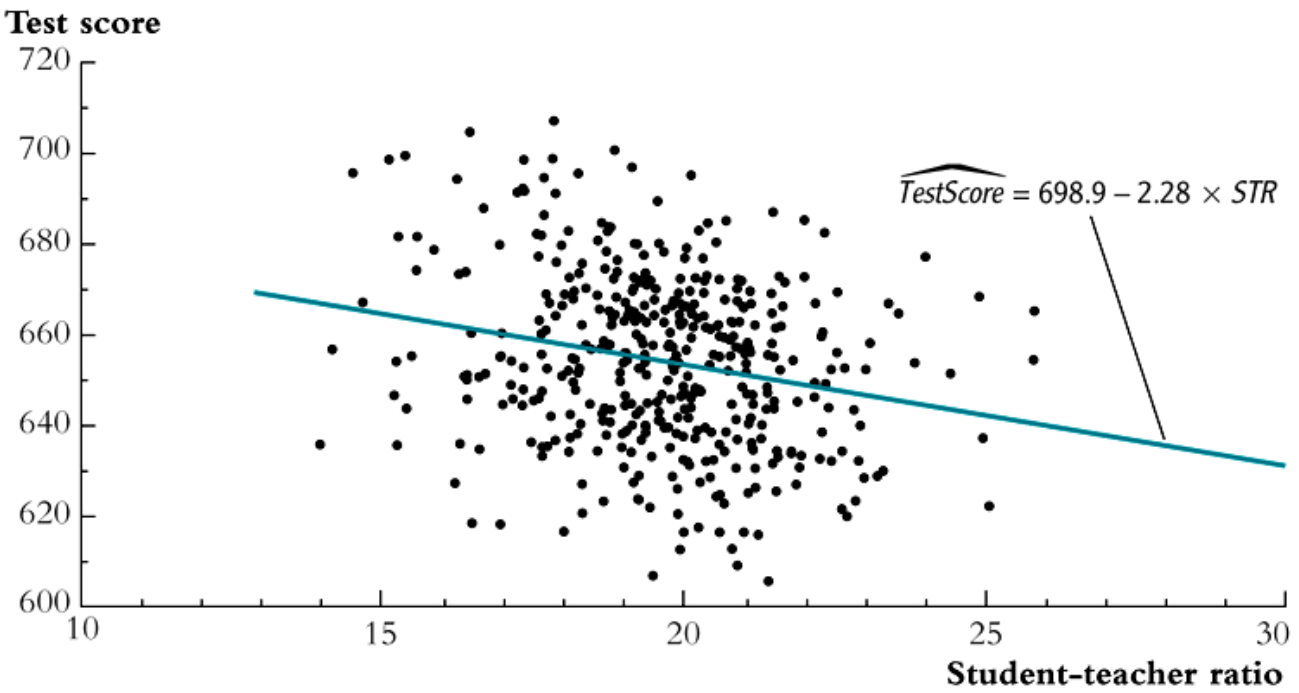


FIGURE 4.3 The Estimated Regression Line for the California Data

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.



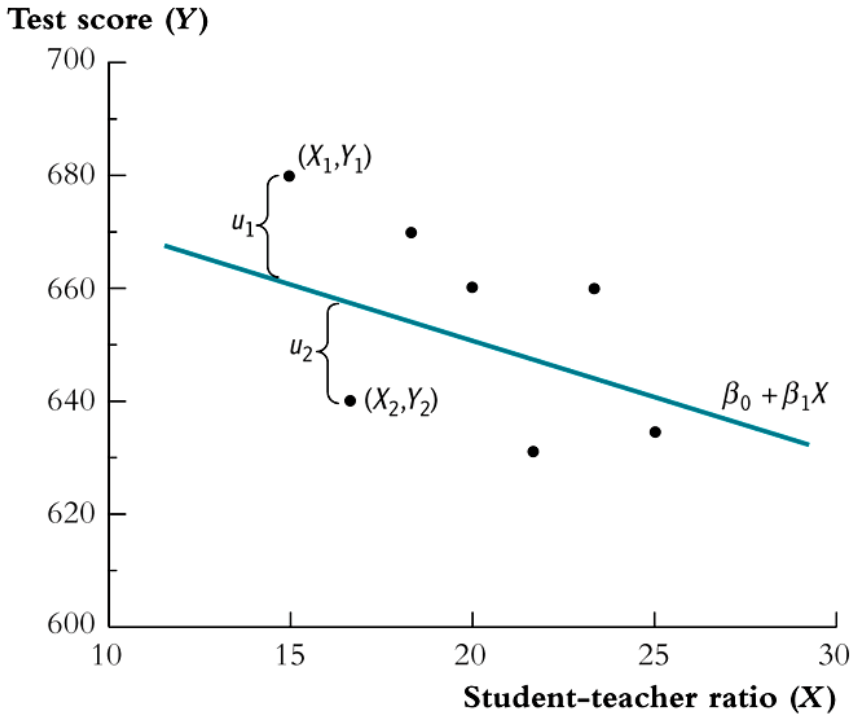
5. Stata in Action

- Stata example

6. Which line to choose? "Error terms"

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.



7. What's to come?

- **Decide which parameters in population we care about (β_0,β_1)**
  - **- just like we did with μ**
- **Draw a sample and estimate parameters**
  - **- just like we did with μ**
- **Construct CI for parameters, test hypotheses, make predictions.**
  - **- just like..**

8. Population regression line: terms

Terminology for the Linear Regression Model with a Single Regressor

The linear regression model is:

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

where:

the subscript i runs over observations, $i = 1, \ldots, n$;

Y_i is the *dependent variable*, the *regressand*, or simply the *left-hand variable*;

X_i is the *independent variable*, the *regressor*, or simply the *right-hand variable*;

$\beta_0 + \beta_1 X$ is the *population regression line* or *population regression function*;

β_0 is the *intercept* of the population regression line;

β_1 is the *slope* of the population regression line; and

u_i is the *error term*.

Next time..

- Estimators for intercept β_0 and slope β_1
- Confidence intervals for β_0 , β_1

Lesson #5: Simple Regression (One Regressor)

1. **Introduction: Fitting a Line through a cloud**
2. **Cereal bars example**
3. **Global Warming example**
4. **CA test score example**
5. **Stata in action**
6. **Which line to choose?**
7. **What's to come?**
8. **Population Regression Line**

Appendix 4.1

The California Test Score Data Set

The California Standardized Testing and Reporting data set contains data on test performance, school characteristics, and student demographic backgrounds. The data used here are from all 420 K–6 and K–8 districts in California with data available for 1998 and 1999.

Test scores are the average of the reading and math scores on the Stanford 9 Achievement Test, a standardized test administered to fifth-grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as "full-time-equivalents"), number of computers per classroom, and expenditures per student. The student-teacher ratio used here is the number of full-time equivalent teachers in the district, divided by the number of students. Demographic variables for the students also are averaged across the district. The demographic variables include the percentage of students in the public assistance program CalWorks (formerly AFDC), the percentage of students that qualify for a reduced price lunch, and the percentage of students that are English learners (that is, students for whom English is a second language). All of these data were obtained from the California Department of Education (www.cde.ca.gov).