

Econ 312: Introduction to Econometrics

Simple Linear Regression

Sang-Yeob Lee

February 16, 2010

Linear Regression with one regressor

- Linear regression allows us to estimate, and make inference about, **population** slope coefficients. Ultimately our aim is to estimate the causal effect on Y of a unit change in X - but for now, just think of the problem of fitting a straight line to data on two variables, Y and X .

The problems of statistical inference for linear regression are, at a general level, the same as for estimation of the mean.

Statistical or econometric, inference about the slope entails:

- Estimation:
 - How should we draw a line through the data to estimate the population slope? (Answer: ordinary least square)
 - What are advantages and disadvantages of OLS?
- Hypothesis testing:
 - How to test if the slope is zero?
- Confidence interval:
 - How to construct a confidence interval for the slope?

Linear Regression: Some Notation and Terminology

The *population regression line*:

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

β_1 = slope of population regression line or slope coefficient

$$= \frac{\Delta Y}{\Delta X}$$

= change in Y for a unit change in X

- Why are β_0 and β_1 “population” parameters?
- We would like to know the population value of β_1 .
- We don't know β_1 , so must estimate it using data.

General notation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \dots, n$$

- X is the independent variable or regressor
- Y is the dependent variables
- β_0 = intercept or constant
- β_1 = slope (or slope coefficient)
- ϵ_i = regression error or stochastic error

Why must the error term be included in the regression model?

- The regression error consists of omitted factors, or possibly measurement error in the measurement of Y . In general, these omitted factors are other factors that influence Y , other than the variable X .

Empirical Problem: Class size and educational output

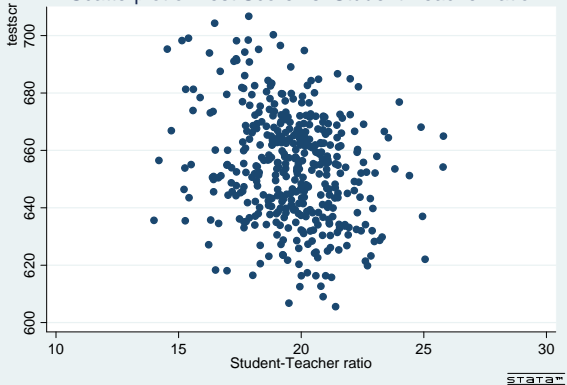
- Policy question: What is the effect on test scores of reducing class size by one students per class?
- We must use data to find out (is there anyway to answer this without data?)
- Population regression line: $TestScore = \beta_0 + \beta_1 STR$

All K-5 and K-8 California School district (n=420)

Variables:

- 5th grade test score, district average
- Student-teacher ration(STR)=no. of students in the district divided by no. of full time equivalent teachers.

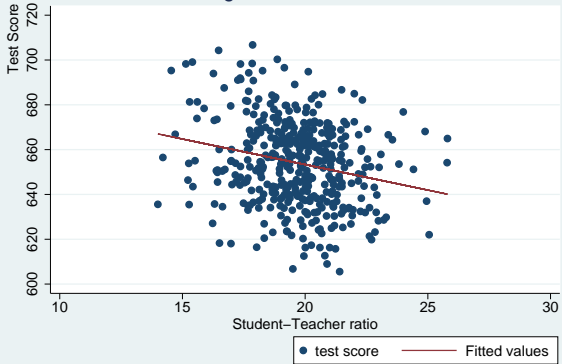
Scatterplot of Test Score vs. Student-Teacher ratio



We need to get some numerical evidence on whether districts with low STRs have higher test scores-But how?

- Estimation: Estimate β_0 and β_1
- Hypothesis Testing: Test “null” hypothesis that $\beta_1 = 0$, against the “alternative” hypothesis that $\beta_1 > 0$
- Confidence Interval: Estimate an interval for β_1

The Estimated Regression Line for the California Data



STATA

The Ordinary Least Square Estimator

- How can we estimate β_0 and β_1 from data?
- Recall that \bar{Y} was the least squares estimator of μ_Y : \bar{Y} solves,

$$\min_m \sum_{i=1}^n (Y_i - m)^2$$

By analogy, we will focus on the least squares (“ordinary least squares” or “OLS”) estimator of the unknown parameters β_0 and β_1 , which solves,

$$\min_{b_0, b_1} \sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2$$

The OLS estimator
solves: $\min_{b_0, b_1} \sum_{i=1}^n (Y_i - (b_0 + b_1 X_i))^2$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (“predicted value”) based on the estimated line.
- This minimization problem can be solved using calculus.
- The result is the OLS estimator of β_0 and β_1 .

The OLS Estimator, Predicted Values, and Residuals

- The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- The OLS predicted values \hat{Y}_i and residuals $\hat{\epsilon}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \dots, n$$

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, \dots, n$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{\epsilon}_i$) are computed from a sample of n observations of X_i and Y_i , $i = 1, \dots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (ϵ_i).

Application to the California Test Score-Class Size data

- Estimated regression line: $\widehat{TestScore} = 689.9 - 2.28 \times STR$
 - Estimated slope = $\hat{\beta}_1 = -2.28$
 - Estimated intercept = $\hat{\beta}_0 = 698.9$
- Interpretation of the estimated slope and intercept
 - District with one more student per teacher on average have test scores that are 2.28 points lower. That, is $\frac{\Delta TestScore}{\Delta STR} = -2.28$
 - The intercept means (taken literally) that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9.
 - This interpretation of the intercepts makes no sense-it extrapolates the line outside the range of the data-here, the intercept is not economically meaningful.

Predicted values & Residuals:

- One of the districts in the data set is Antelope, CA, for which $STR = 19.33$ and $TestScore = 657.8$
 - $\widehat{TestScore} = 689.9 - 2.28 \times STR$
 - Predicted value:

$$\widehat{TestScore}_{Antelope} = 689.9 - 2.28 \times 19.33 = 654.8$$

- residual:

$$\hat{\epsilon}_{Antelope} = 657.8 - 654.8 = 3.0$$

OLS regression: STATA output

```
. regress testscr str
```

Source	SS	df	MS			
Model	7794.11004	1	7794.11004	Number of obs =	420	
Residual	144315.484	418	345.252353	F(1, 418) =	22.58	
				Prob > F =	0.0000	
				R-squared =	0.0512	
				Adj R-squared =	0.0490	
				Root MSE =	18.581	
Total	152109.594	419	363.030056			

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.4798256	-4.75	0.000	-3.22298	-1.336637
_cons	689.933	9.467491	73.82	0.000	680.3231	717.5428

$\widehat{TestScore} = 689.9 - 2.28 \times STR$ (we'll discuss the rest of this output later.)