

Introduction to Linear Regression (SW Chapter 4)

EC 471
Spring 2004

4-1

We need to think about a regression model.

- We want to examine the effect of **class size**, measured in terms of **STR** (Student Teacher Ratio), on **Test Score**.
 - Class size (STR) → Independent variable
 - Test score → Dependent variable (what we want to examine)
- A regression model to consider is

$$\text{Test Score} = \beta_0 + \beta_1 \text{STR} + u$$

4-3

Think about an empirical question about educational output

- Policy question: What is the effect of **class size** by one student per class?
- What is the right output (performance)?
 - parent satisfaction
 - student personal development
 - future adult welfare
 - future adult earnings
 - **performance on standardized tests**

The California Test Scores

All K-6 and K-8 California schools

Variables:

- 5th grade test scores (Standardized combined math and reading)
- Student-teacher ratio (STR) = total no. full-time teachers in district divided by no. full-time students

An initial look at the California test score data:

TABLE 4.1 Summary of the Distribution of Student-Teacher Ratios and Fifth-Grade Test Scores for 420 K-8 Districts in California in 1998

	Average	Standard Deviation	Percentile						
			10%	25%	40%	50% (median)	60%	75%	90%
Student-teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

4-5

The class size/test score policy question:

- What is the effect on test scores of reducing STR by one student/class?
- Object of policy interest: $\frac{\Delta \text{Test score}}{\Delta \text{STR}}$
- This is the slope of the line relating test score and STR. Why?

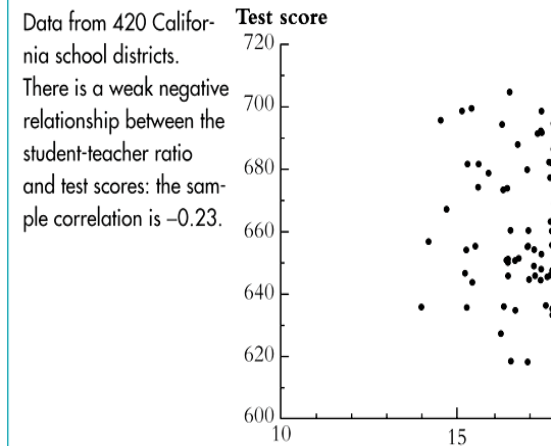
→ We are interested in:

- Sign of change (positive or negative?)
- Magnitude of change..

4-7

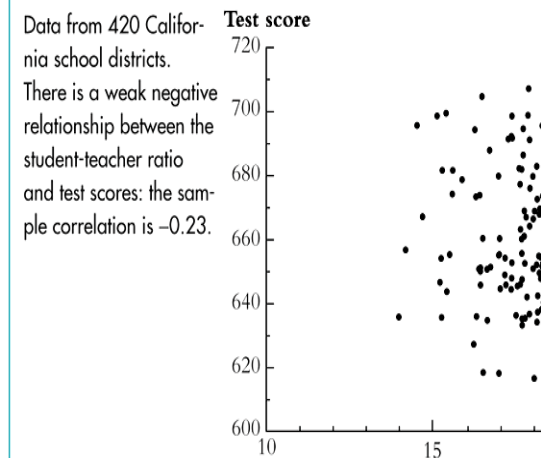
Do districts with smaller classes (lower student-teacher ratios) have higher test scores?

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio



This suggests that we want to draw a regression line from the Test Score v. STR scatterplot – b

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio



Some Notation and Terminology

(Sections 4.1 and 4.2)

The *population regression model*:

$$\text{Test Score} = \beta_0 + \beta_1 \text{STR} + u$$

β_1 = slope of population regression line

$$= \frac{\Delta \text{Test score}}{\Delta \text{STR}}$$

= change in test score for a unit change in STR

- 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.

4-9

How can we estimate β_0 and β_1 from data?

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:

- 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 (pp. 4.2). The result is the OLS estimators of β_0 and β_1 .

4-11

Three questions

- Positive or negative change?
→ β_1 = magnitude of change
- Really, does class size matter?
significant?
→ **Testing hypothesis**
 $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$
- Can we use the regression model to predict?
→ (*Predicted*) Test Score

The OLS Estimator, Predicted Values

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

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

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

The OLS predicted values \hat{Y}_i and residuals \hat{u}_i are

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

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

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and predicted values (\hat{Y}_i) are based on a sample of n observations of X_i and Y_i , $i = 1, \dots, n$. The unknown true population intercept (β_0), slope (β_1), and predicted values (Y_i) are

Example (n = 3)

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

<u>X</u>	<u>Y</u>
1	3
3	5
2	4

Find $\hat{\beta}_0$ and $\hat{\beta}_1$.

4-13

OLS regression: STATA output

```
regress testscr str, robust
Regression with robust standard errors
Number of obs = 420
F( 1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581
```

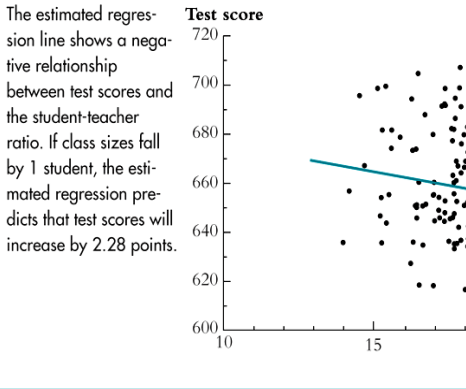
	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
str	-2.279808	.5194892	-4.39	0.000	-3.300945 -1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602 719.3057

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

(We'll discuss the rest of this output later)

4-15

FIGURE 4.3 The Estimated Regression Line for the California Test Scores



Estimated slope = $\hat{\beta}_1 = -2.28$
 Estimated intercept = $\hat{\beta}_0 = 698.9$
 Estimated regression line: $\widehat{TestScore}$

Interpretations of

1. Marginal Effects = estimated

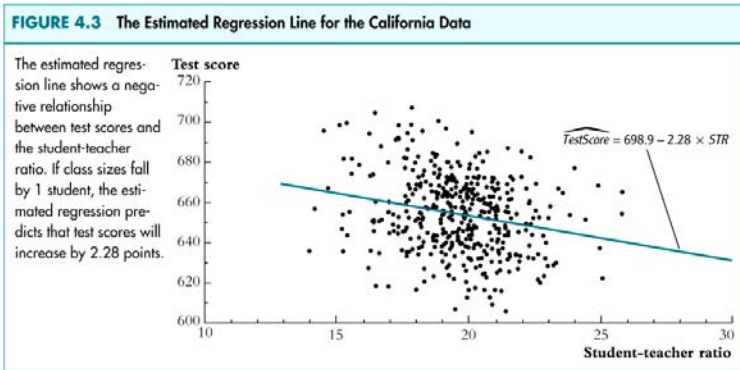
$$\widehat{TestScore} = 698.9 -$$

- Districts with one more student have test scores that are 2.28 points lower

$$\frac{\Delta Test\ score}{\Delta STR} = -2.28$$

- The intercept (taken literally) means that, if the student-teacher ratio were zero, the estimated line, districts with zero students would have a (predicted) test score of 698.9
- This interpretation of the intercept is not meaningful because it extrapolates the line outside the range of the data. In this application, the intercept is not meaningful.

2. Predicted values & residuals:



For instance, one of the districts in the data set is Antelope, CA, for which $STR = 19.33$ and actual Test Score = 657.8.

Predicted value: $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$

Residual: $\hat{u}_{Antelope} = \text{Actual} - \text{Predicted} = 657.8 - 654.8 = 3.0$

4-17

3. Testing Hypothesis

The OLS regression line is an estimate, computed using our sample of data; a different sample would have given a different value of $\hat{\beta}_1$.

How can we:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$?
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$?
- construct a confidence interval for β_1 ?

We proceed in four steps:

1. The probability framework for linear regression
2. Estimation
3. Hypothesis Testing
4. Confidence intervals

4-19

Notations

Population regression

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

$u_i = \text{error term}$

$\beta_0, \beta_1 = \text{parameters}$

Sample (estimated) regression

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$Y_i = \hat{Y}_i + \hat{u}_i$$

$\hat{u}_i = \text{residuals}$

$\hat{\beta}_0, \hat{\beta}_1 = \text{estimates}$

or

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

A. Probability Framework for

Population

population of interest (ex: all people)

Random variables (random sample)

Ex: (Test Score, STR)

Joint distribution of (Y,X)

The key feature is that we suppose a linear relation in the population that regression relation is the “population linear regression”

The Population Linear Regression Model (Section 4.3)

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

- X is the *independent variable* or *regressor*
- Y is the *dependent variable*
- β_0 = intercept
- β_1 = slope
- u_i = “error term”
- The error term 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

4-21

Data and sampling

The population objects (“parameters”) β_0 and β_1 are unknown; so to draw inferences about these unknown parameters we must collect relevant data.

Simple random sampling:

Choose n entities at random from the population of interest, and observe (record) X and Y for each entity

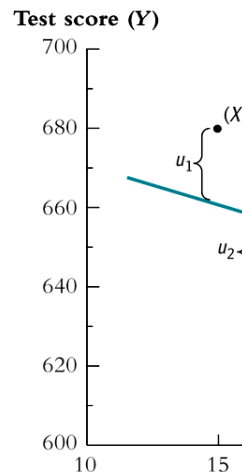
Simple random sampling implies that $\{(X_i, Y_i)\}, i = 1, \dots, n$, are *independently and identically distributed* (i.i.d.). (Note: (X_i, Y_i) are distributed independently of (X_j, Y_j) for different observations i and j .)

4-23

Ex.: The population regression line

FIGURE 4.1 Scatter Plot of Test Score vs. Student

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.



What are some of the omitted factors

Task at hand: to characterize the properties of the OLS estimator. To do so, we need the following assumptions:

The Least Squares

1. The conditional distribution of u given $X = x$ is centered at zero, that is, $E(u|X = x) = 0$.
2. $(X_i, Y_i), i = 1, \dots, n$, are i.i.d.
3. X and u have four moments,

$$E(X^4) < \infty \text{ and } E(u^4) < \infty$$

We'll discuss these assumptions

Least squares assumption #1: $E(u|X = x) = 0$.

For any given value of X , the mean of u is zero

Example: Assumption #1 and the

$$\text{Test Score}_i = \beta_0 + \beta_1 \text{STR}_i$$

$u_i = \text{other factors}$

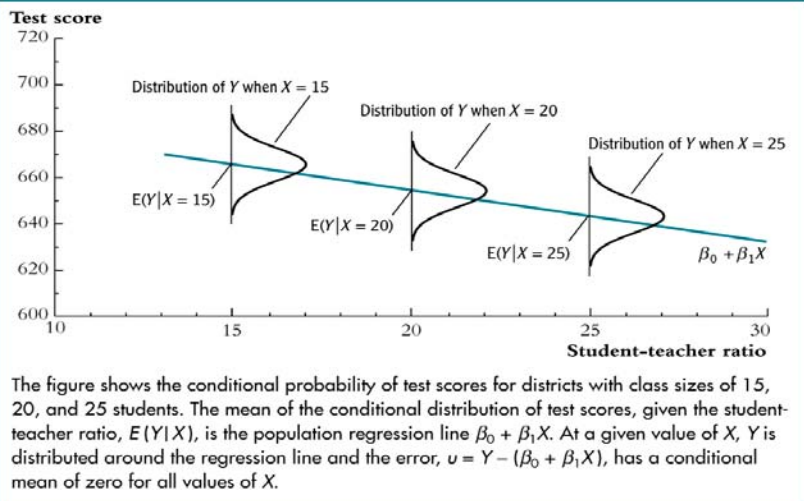
“Other factors:”

- home environment conducive
- **family income** is a useful predictor

So $E(u|X=x) = 0$ means $E(\text{family income} | \text{STR}) = 0$ (which implies that family income is uncorrelated). *This assumption is often violated.*

We have omitted variables when this assumption is not satisfied.

FIGURE 4.4 The Conditional Probability Distributions and the Population Regression Line



4-25

Least squares assumption #2:

$(X_i, Y_i), i = 1, \dots, n$ are i.i.d. (identical and independently distributed)

If some observations have a large variance as in most “cross-sectional” data, the **identical** assumption is violated.

The main place we will encounter non-**independent** sampling is when data are recorded over time (“time series data”) – this will introduce some extra complications.

Least squares assumption #3:

$$E(X^4) < \infty \text{ and } E(u^4) < \infty$$

Because $Y_i = \beta_0 + \beta_1 X_i + u_i$, assumption #3 can equivalently be stated as, $E(X^4) < \infty$ and $E(u^4) < \infty$.

Assumption #3 is generally plausible. The data implies finite fourth moments. Test scores automatically satisfy this too.

4-27

B. Estimation: the Sampling Distribution of $\hat{\beta}_1$

(Section 4.4)

$\hat{\beta}_1$ has a **sampling distribution**.

- What is $E(\hat{\beta}_1)$? (where is it centered)
 - $E(\hat{\beta}_1) = \beta_1$ (implying: $\hat{\beta}_1$ is an unbiased estimator)

- What is $\text{var}(\hat{\beta}_1)$? (measure of sampling uncertainty)

$$\text{var}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]} \quad \text{where } k = \# \text{ of indep.}$$

4-29

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$, where the variance of this distribution, $\sigma_{\hat{\beta}_1}^2$, is

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \quad (4.14)$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma_{\hat{\beta}_0}^2)$, where

$$\sigma_{\hat{\beta}_0}^2 = \frac{1}{n} \frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \quad \text{where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)} \right) X_i. \quad (4.15)$$

4-31

Or, precisely,

$$\text{var}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{1}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]}$$

- Alternatively, robust error given in the next page.

When n is large, the sampling distribution is approximated by a normal distribution.

Theorem)

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \text{ is approximately } d$$

Back to Previous example ($n = 3$)

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

<u>X</u>	<u>Y</u>
1	3
3	5
2	4

$$\text{Find } \text{var}(\hat{\beta}_1) = \frac{\frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]}$$

C. Hypothesis Testing (Section 4.5)

Suppose a skeptic suggests that reducing the number of students in a class has no effect on learning or, specifically, test scores. The skeptic thus asserts the hypothesis,

$$H_0: \beta_1 = 0$$

We wish to test this hypothesis using data – reach a tentative conclusion whether it is correct or incorrect.

4-33

General Form of the t -Statistic

In general, the t -statistic has the form

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}} \quad (4.18)$$

where the SE of the estimator is the square root of an estimator of the variance of the estimator.

Applied to a hypothesis about β_1 :

$$t = \frac{\text{estimator} - \text{hypothesized value}}{\text{standard error of the estimator}}$$
$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

4-35

Null hypothesis and **two-sided** a

$$H_0: \beta_1 = 0 \text{ vs. } H_a$$

or, more generally,

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_a$$

where $\beta_{1,0}$ is the hypothesized value

Null hypothesis and **one-sided** a

$$H_0: \beta_1 = \beta_{1,0} \text{ vs. } H_a$$

In economics, it is almost always with stories in which an effect is it is standard to focus on two-sided

where β_1 is the value of $\beta_{1,0}$ hypothesis (for example, if the null value is

What is $SE(\hat{\beta}_1)$?

$SE(\hat{\beta}_1)$ = the square root of variance of the sampling distribution

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2} = \sqrt{\text{var}(\hat{\beta}_1)}$$

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

OK, this is a bit nasty, but:

- There is no reason to memorize this
- It is computed automatically by regression software
- $SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}^2}$ is reported by regression software

<u>X</u>	<u>Y</u>
1	3
3	5
2	4

Find t-statistic for $\beta_1 = 0$. (that is, $\beta_{1,0} = 0$)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}_{\hat{\beta}_1}^2}}$$

4-37

Decision Rule:

- Reject at 5% significance level if $|t| > t_c$
where t_c is the critical value with $df = n - k - 1$
- Reject at 5% significance level if $p\text{-value} < 5\%$.
where $p\text{-value}$ is $p = \Pr[|t| > |t^{act}|] =$ probability in tails of normal outside $|t^{act}|$

Example: Test Scores and SAT

Estimated regression line: $\widehat{TestScore} = \beta_0 + \beta_1 SAT$

Regression software reports the

$$SE(\hat{\beta}_0) = 10.4$$

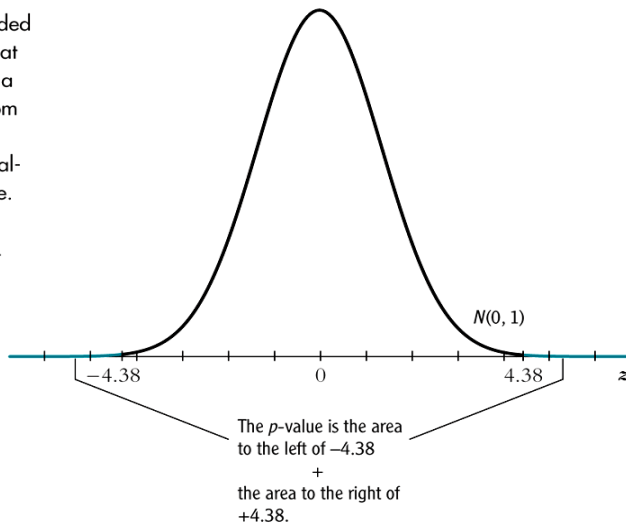
$$t\text{-statistic testing } \beta_{1,0} = 0 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- The 1% 2-sided significance level is used to reject the null at the 1% significance level
- Alternatively, we can compute the p-value

4-39

FIGURE 4.6 Calculating the p -Value of a Two-Sided Test When $t^{act} = -4.38$

The p -value of a two-sided test is the probability that $|Z| \geq |t^{act}|$, where Z is a standard normal random variable and t^{act} is the value of the t -statistic calculated from the sample. When $t^{act} = -4.38$, the p -value is only .00001.



The p -value based on the large- n standard normal approximation to the t -statistic is 0.00001 (10^{-4})

4-41

Example: Test Scores and STR, California data

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

$$SE(\hat{\beta}_0) = 10.4 \quad SE(\hat{\beta}_1) = 0.52$$

95% confidence interval for $\hat{\beta}_1$:

$$\begin{aligned} \{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\} &= \{-2.28 \pm 1.96 \times 0.52\} \\ &= (-3.30, -1.26) \end{aligned}$$

“Reject the null if the C.I. does not include the value under the null.”

Equivalent statements:

- The 95% confidence interval does not include zero;
- The hypothesis $\beta_1 = 0$ is rejected at the 5% level

4-43

D. Confidence intervals (Section 4.4)

In general, if the sampling distribution is approximately normal for large n , then a 95% confidence interval can be constructed as estimator $\pm t_c \times SE$.

So: a $(1-\alpha)100\%$ confidence interval is

$$\{\hat{\beta}_1 \pm t_c \times SE\}$$

A convention for reporting estimates

Put standard errors in parentheses

$$\begin{aligned} \widehat{TestScore} &= 698.9 - 2.28 \\ &\quad (10.4) \quad (0.52) \end{aligned}$$

This expression means that:

- The estimated regression line is $\widehat{TestScore} = 698.9 - 2.28 \times STR$
- The standard error of $\hat{\beta}_0$ is 10.4
- The standard error of $\hat{\beta}_1$ is 0.52

OLS regression: STATA output

```
regress testscr str, robust
```

Regression with robust standard errors

```
Number of obs = 420
F( 1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581
```

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
str	-2.279808	.5194892	-4.38	0.000	-3.300945 -1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602 719.3057

so:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

(10.4) (0.52)

$$t(\beta_1 = 0) = -4.38, p\text{-value} = 0.000$$

$$95\% \text{ conf. interval for } \beta_1 \text{ is } (-3.30, -1.26)$$

4-45

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ where } X \text{ is binary } (X_i = 0 \text{ or } 1):$$

- When $X_i = 0$: $Y_i = \beta_0 + u_i$
- When $X_i = 1$: $Y_i = \beta_0 + \beta_1 + u_i$

thus:

- When $X_i = 0$, the mean of Y_i is β_0
- When $X_i = 1$, the mean of Y_i is $\beta_0 + \beta_1$

that is:

- $E(Y_i | X_i=0) = \beta_0$
- $E(Y_i | X_i=1) = \beta_0 + \beta_1$

so:

$$\begin{aligned} \beta_1 &= E(Y_i | X_i=1) - E(Y_i | X_i=0) \\ &= \text{population difference in group means} \end{aligned}$$

4-47

Regression when X is Binary

Sometimes a regressor is binary

- $X = 1$ if female, = 0 if male
- $X = 1$ if treated (experimental), = 0 if control
- $X = 1$ if small class size, = 0 if large

So far, β_1 has been called a “slope” but it doesn't make much sense if X is binary.

How do we interpret regression coefficients when X is binary?

Example: *TestScore* and *STR*, California

Let

$$D_i = \begin{cases} 1 & \text{if } STR_i \leq 20 \\ 0 & \text{if } STR_i > 20 \end{cases}$$

The OLS estimate of the regression of *TestScore* to D (with standard errors)

$$\widehat{TestScore} = 650.0 + 7.4 D_i$$

(1.3) (1.8)

Difference in means between groups

$SE = 1.8$

Compare the regression results with the group means, computed directly:

Class Size	Average score (\bar{Y})	Std. dev. (s_Y)	N
Small ($STR > 20$)	657.4	19.4	238
Large ($STR \geq 20$)	650.0	17.9	182

Estimation: $\bar{Y}_{small} - \bar{Y}_{large} = 657.4 - 650.0 = 7.4$

Test $\Delta=0$: $t = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)} = \frac{7.4}{1.83} = 4.05$

95% confidence interval = $\{7.4 \pm 1.96 \times 1.83\} = (3.8, 11.0)$

This is the same as in the regression!

$$\widehat{TestScore} = 650.0 + 7.4 \times D$$

(1.3) (1.8)

4-49

Other Regression Statistics (Section 4.8)

A natural question is how well the regression line “fits” or explains the data. There are two regression statistics that provide complementary measures of the quality of fit:

- The *regression R^2* measures the fraction of the variance of Y that is explained by X ; it is unitless and ranges between zero (no fit) and one (perfect fit)
- The *standard error of the regression* measures the fit – the typical size of a regression residual – in the units of Y .

4-51

Summary: regression when X

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

- β_0 = mean of Y given that $X = 0$
- $\beta_0 + \beta_1$ = mean of Y given that $X = 1$
- β_1 = difference in group means
- $SE(\hat{\beta}_1)$ has the usual interpretation
- t -statistics, confidence intervals
- This is another way to do difference-in-means analysis
- The regression formulation is useful because we have additional regressors

The R^2

Write Y_i as the sum of the OLS predicted value and residual:

$$Y_i = \hat{Y}_i + e_i$$

The R^2 is the fraction of the sample variance “explained” by the regression, that is

$$R^2 = \frac{ESS}{TSS}$$

where $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ and $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$

$$R^2 = \frac{ESS}{TSS}, \text{ where } ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

The R^2 :

- $R^2 = 0$ means $ESS = 0$, so X explains none of the variation of Y
- $R^2 = 1$ means $ESS = TSS$, so $Y = \hat{Y}$ so X explains all of the variation of Y
- $0 \leq R^2 \leq 1$
- For regression with a single regressor (the case here), R^2 is the *square* of the correlation coefficient between X and Y

Note: R^2 does not mean much, though.

By adding more variables, R^2 always increase.

4-53

$$SER = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2}$$

The SER :

- has the units of u , which are the units of Y
- measures the spread of the distribution of u
- measures the average “size” of the OLS residual (the average “mistake” made by the OLS regression line)
- The **root mean squared error (RMSE)** is closely related to the SER :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2}$$

This measures the same thing as the SER – the minor difference is division by $1/n$ instead of $1/(n-2)$.

4-55

The Standard Error of the Regressor

The standard error of the regression is the sample standard deviation of the residuals:

$$\begin{aligned} SER &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2} \\ &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2} \end{aligned}$$

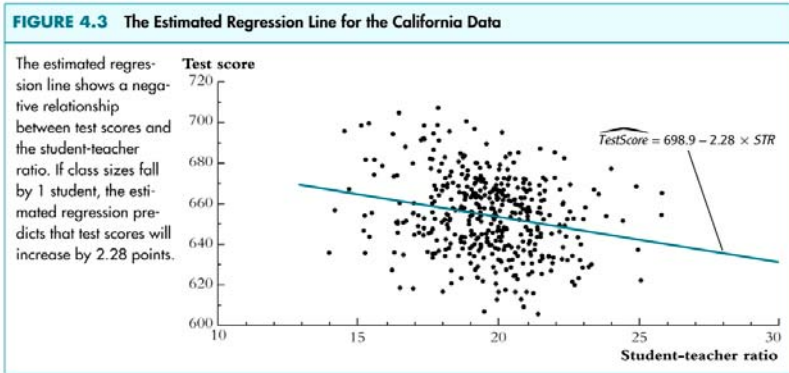
(the second equality holds because...)

Technical note: why divide by $n-2$?

$$SER = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2}$$

- Division by $n-k-1$ is a “degrees of freedom” adjustment
- When n is large, it makes negligible difference – n , $n-1$, or $n-2$ are used – although the formula uses $n-2$ when there is one regressor
- For details, see Section 15.4

Example of R^2 and SER



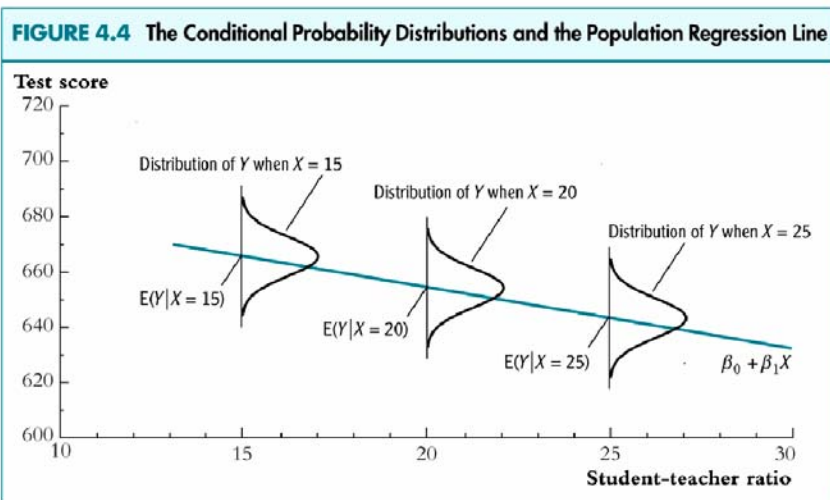
$$\widehat{TestScore} = 698.9 - 2.28 \times STR, R^2 = .05, SER = 18.6$$

(10.4) (0.52)

The slope coefficient is statistically significant and large in a policy sense, even though STR explains only a small fraction of the variation in test scores.

4-57

Homoskedasticity in a picture:



- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u does **not** change with (depend on) x

4-59

A Practical Note: Heteroskedasticity, and the Formulas for the Standard Errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

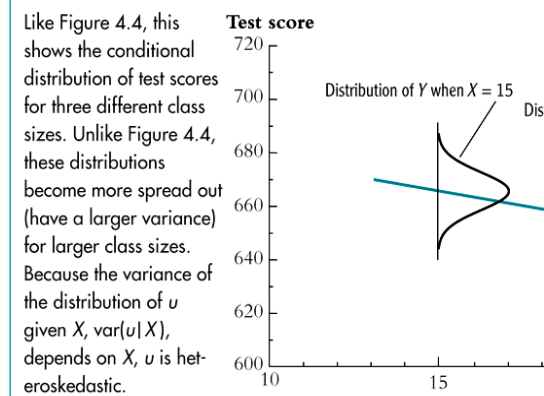
- What do these two terms mean?
- Consequences of homoskedasticity
- Implication for computing standard errors

What do these two terms mean?

If $\text{var}(u|X=x)$ is constant – that is, the conditional distribution of u given $X=x$ has the same variance for all x , then u is said to be **homoskedastic**. If $\text{var}(u|X=x)$ varies with x , then u is said to be **heteroskedastic**.

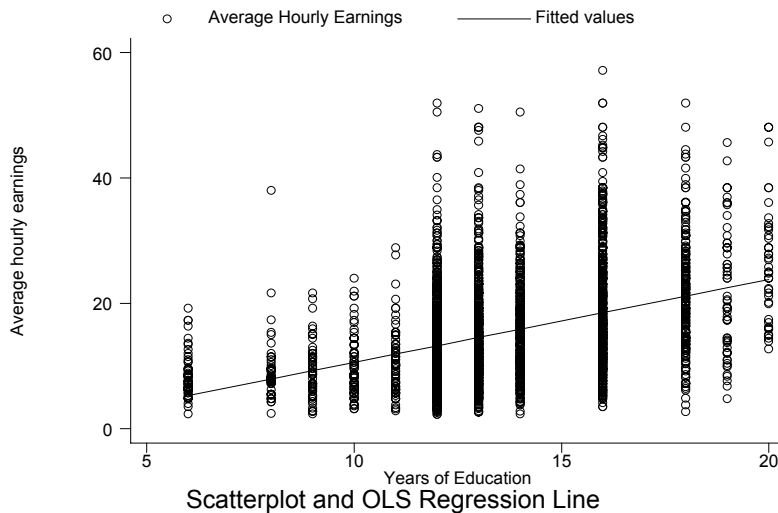
Heteroskedasticity in a picture:

FIGURE 4.7 An Example of Heteroskedasticity



- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u depends on x (heteroskedastic).

An real-world example of *heteroskedasticity* from labor economics: average hourly earnings vs. years of education (data source: 1999 Current Population Survey)



4-61

So far we have (without saying so) assumed that u is homoskedastic:

Recall the three least squares assumptions:

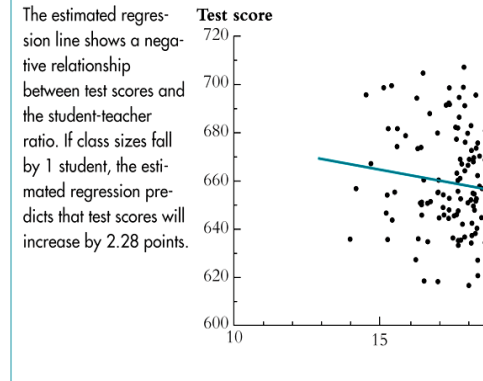
1. The conditional distribution of u given X has mean zero, that is, $E(u|X = x) = 0$.
2. $(X_i, Y_i), i = 1, \dots, n$, are i.i.d.
3. X and u have four finite moments.

Heteroskedasticity and homoskedasticity concern $\text{var}(u|X=x)$. Because we have not explicitly assumed homoskedastic errors, we have implicitly allowed for heteroskedasticity.

4-63

Is heteroskedasticity present in t

FIGURE 4.3 The Estimated Regression Line for the California



Hard to say... looks nearly homoskedastic. Variance might be tighter for large values.

What if the errors are in fact homoskedastic?

- You can prove some theorems. In particular, the Gauss-Markov theorem states that OLS is the estimator with minimum variance among all estimators that are linear and unbiased (in (Y_1, \dots, Y_n) ; see Section 15.5).
- The formula for the variance of the standard error simplifies (Appendix A). If σ_u^2 is constant, then

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_x^2)^2}$$

Note: $\text{var}(\hat{\beta}_1)$ is inversely proportional to the spread in X . More spread in X means more

General formula for the standard error of $\hat{\beta}_1$ is the $\sqrt{\quad}$ of:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \times \frac{\frac{1}{n-2} \sum_{i=1}^n (X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2}$$

Special case under homoskedasticity:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Sometimes it is said that the lower formula is simpler.

4-65

The critical points:

- If the errors are homoskedastic and you use the heteroskedastic formula for standard errors (the one we derived), you are OK
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, the standard errors are wrong.
- The two formulas coincide (when n is large) in the special case of homoskedasticity
- The bottom line: you should always use the heteroskedasticity-based formulas – these are conventionally called the **heteroskedasticity-robust standard errors**.

4-67

The homoskedasticity-only formula for $\hat{\beta}_1$ and the “heteroskedasticity formula that is valid under heteroskedasticity” are different. In general, you get different standard errors for different formulas.

Homoskedasticity-only standard errors are the default setting in regression software. Sometimes the only setting that gives you the general “heteroskedasticity-robust” standard errors you must use.

If you don’t override the default setting for heteroskedasticity, you will get the wrong standard errors (and wrong t -statistics and confidence intervals).

Heteroskedasticity-robust standard errors

```
regress testscr str, robust
```

Regression with robust standard errors

	Coef.	Robust Std. Err.	t
testscr			
str	-2.279808	.5194892	-4.39
_cons	698.933	10.36436	67.44

Use the “, robust”

Summary and Assessment (Section 4.10)

- The initial policy question:
Suppose new teachers are hired so the student-teacher ratio falls by one student per class. What is the effect of this policy intervention (this “treatment”) on test scores?
- Does our regression analysis give a convincing answer?
Not really – districts with low *STR* tend to be ones with lots of other resources and higher income families, which provide kids with more learning opportunities outside school...this suggests that $\text{corr}(u_i, STR_i) > 0$, so $E(u_i|X_i) \neq 0$.

4-69

Ideal Randomized Controlled Experiment

- *Ideal*: subjects all follow the treatment protocol – perfect compliance, no errors in reporting, etc.!
- *Randomized*: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- *Controlled*: having a control group permits measuring the differential effect of the treatment
- *Experiment*: the treatment is assigned as part of the experiment: the subjects have no choice, which means that there is no “reverse causality” in which subjects choose the treatment they think will work best.

4-71

Digression on Causality

The original question (what is the effect of an intervention that reduces class size) is about a causal effect: the effect of the treatment is β_1 .

- But what is, precisely, a causal effect?
- The common-sense definition of a causal effect is not precise enough for our purposes.
- In this course, we define a causal effect as the effect that is measured in an *ideal randomized controlled experiment*.

Back to class size:

- What is an ideal randomized controlled experiment for measuring the effect on *Test Scores* of a change in *Class Size*?
- How does our regression analysis differ from this ideal?
 - The treatment is not randomized
 - In the US – in our observational data – higher family incomes are linked to smaller classes **and** higher test scores
 - As a result it is plausible that the effect of class size is confounded
 - If so, Least Squares Assumptions 1 and 2 are violated
 - If so, $\hat{\beta}_1$ is biased: does an increase in class size seem more important than it really is?