Lecture Notes #2
Economics 120B
Econometrics

Prof. Dahl

UC San Diego

Linear Regression with One Regressor
(SW Chapter 4)

- Linear regression allows us to estimate, and make inferences 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 or of the differences between two means. Statistical, or econometric, inference about the slope entails:

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

Linear Regression: Some Notation and Terminology

(SW Section 4.1)

The ***population regression line*:**

$$\textit{Test Score} = \beta_0 + \beta_1 STR$$

β_1 = slope of population regression line

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

= change in test score for a unit change in STR

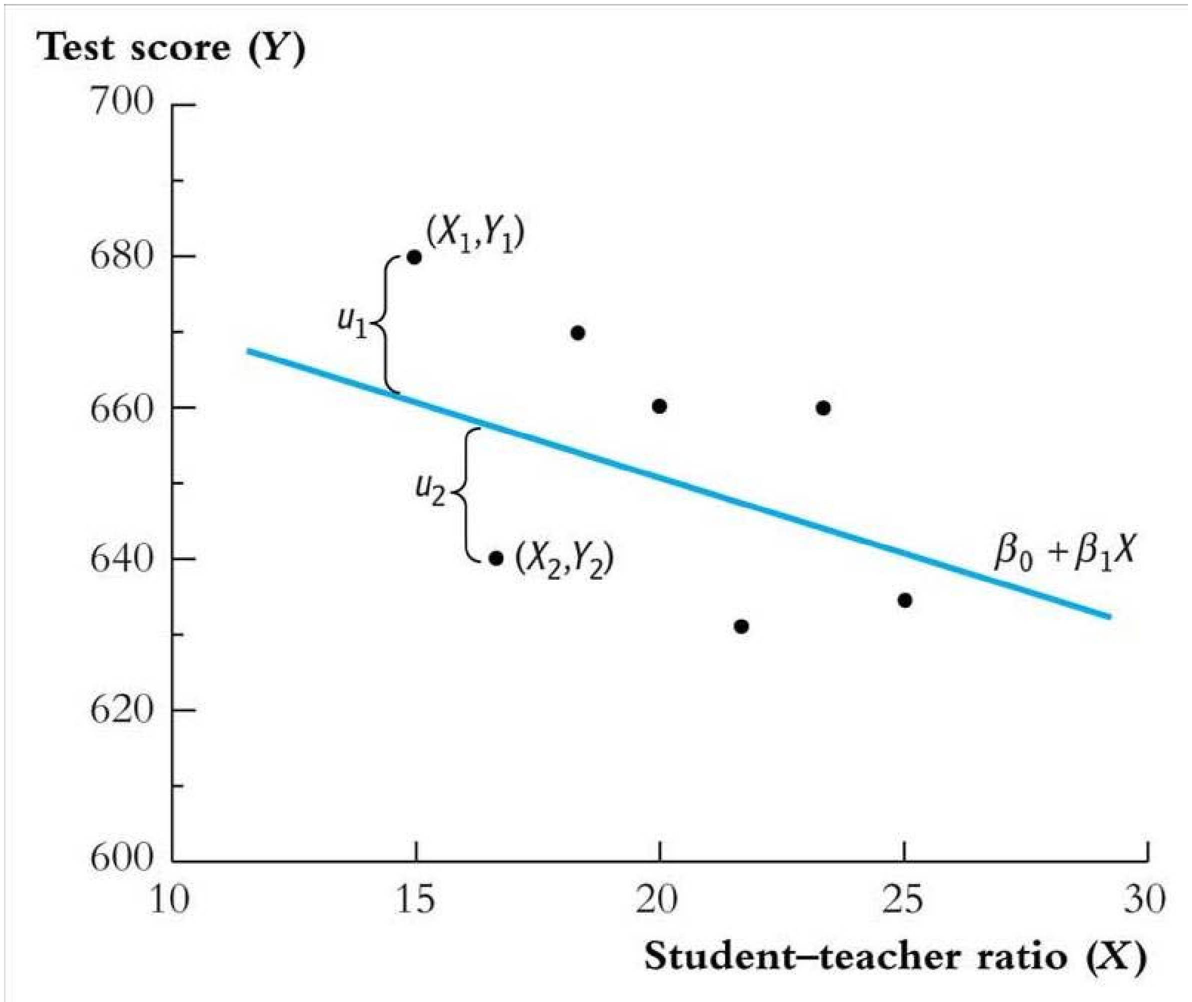
- *Why are β_0 and β_1 "population" parameters*?
- We would like to know the population value of β_1.
- We don't know β_1, so must estimate it using data.

The Population Linear Regression Model – general notation

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

- X is the ***independent variable*** or ***regressor***
- Y is the ***dependent variable***
- β_0 = ***intercept***
- β_1 = ***slope***
- u_i = the regression ***error***
- 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

This terminology in a picture: Observations on *Y* and *X*; the population regression line; and the regression error (the "error term"):



The Ordinary Least Squares Estimator

(SW Section 4.2)

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,

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

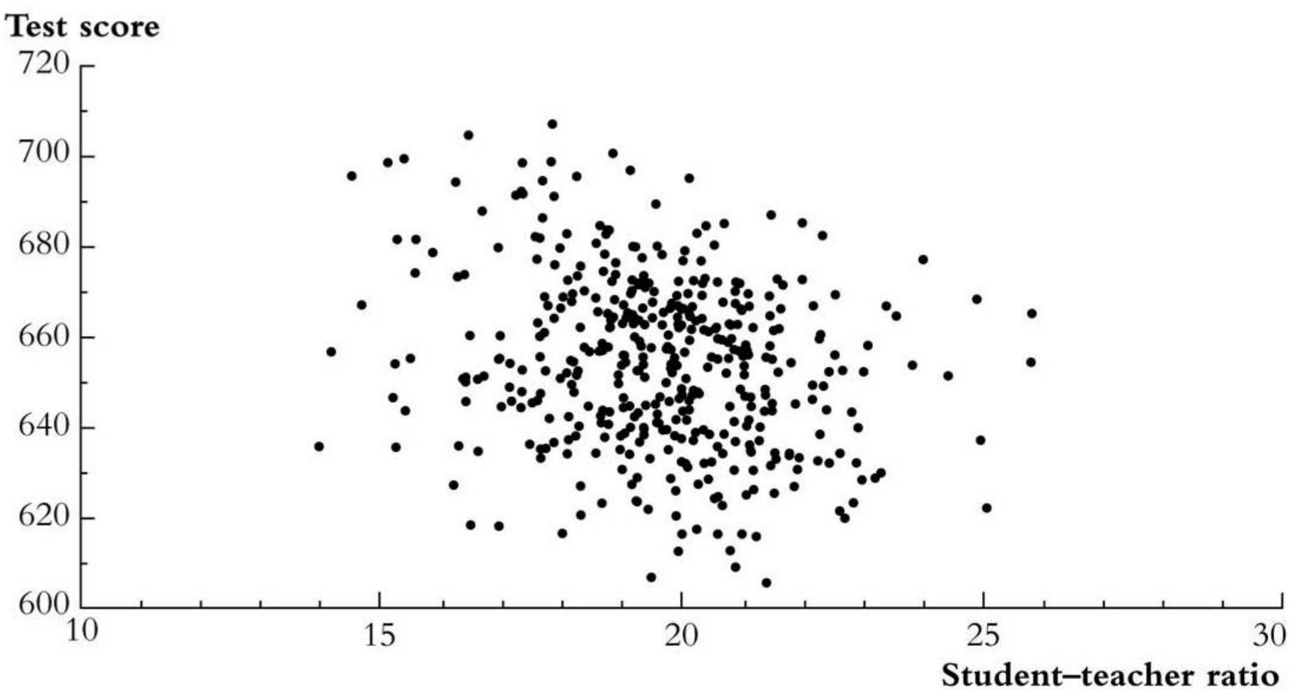
By analogy, **we will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$**, which solves,

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

Mechanics of OLS

The population regression line: $Test\ Score = \beta_0 + \beta_1 STR$

$$\beta_1 = \frac{\Delta \text{Test score}}{\Delta STR} = ??$$



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 (App. 4.2).
- **The result is the OLS estimators 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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

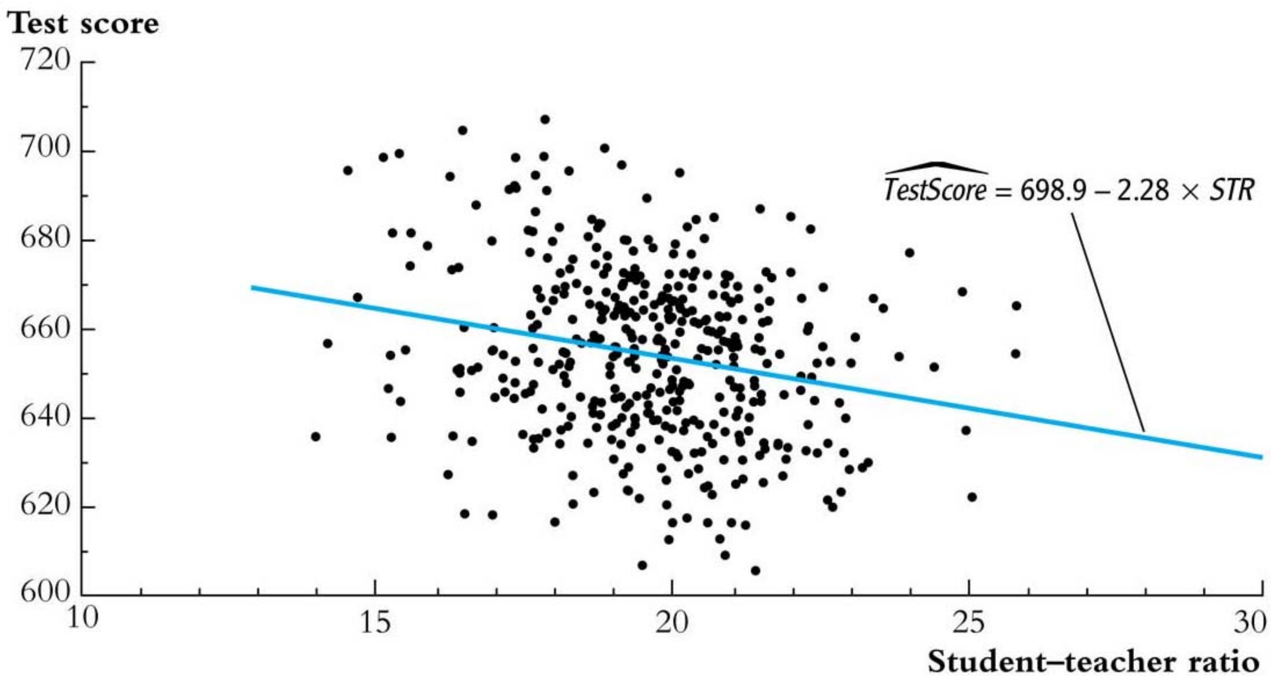
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

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

Application to the California *Test Score – Class Size* data



Estimated slope = $\hat{\beta}_1 = -2.28$

Estimated intercept = $\hat{\beta}_0 = 698.9$

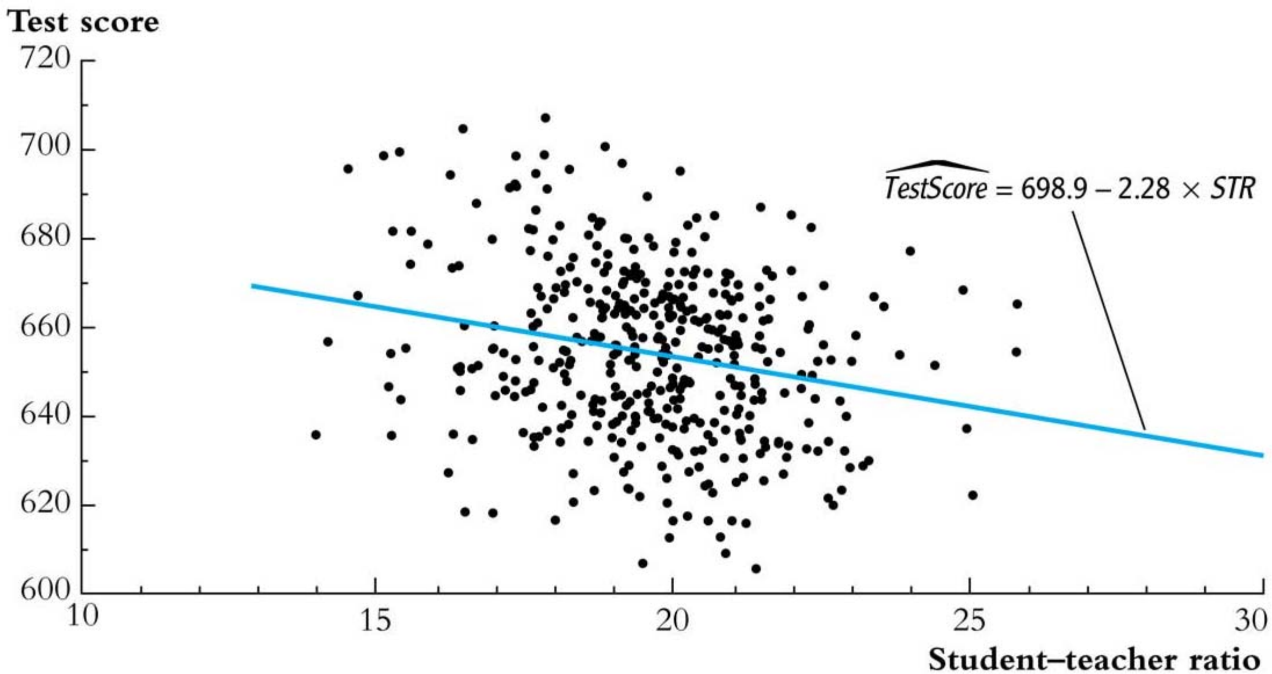
Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 * \text{STR}$

Interpretation of the estimated slope and intercept

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

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\dfrac{\Delta \text{Test score}}{\Delta STR} = -2.28$
- The intercept (taken literally) means 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 intercept 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 *Test Score* = 657.8

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

residual: $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

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

-------------------------------------------------------------------------
        |               Robust
testscr |      Coef.   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
-------------------------------------------------------------------------
```

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

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

Measures of Fit

(Section 4.3)

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*** (***SER***) measures the magnitude of a typical regression residual in the units of *Y*.

The *regression* R^2 is the fraction of the sample variance of Y_i "explained" by the regression.

$Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual

$\Rightarrow$ sample var (Y) = sample var($\hat{Y}_i$) + sample var($\hat{u}_i$) (*why?*)

$\Rightarrow$ total sum of squares = "explained" SS + "residual" SS

Definition of R^2:
$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \le R^2 \le 1$
- For regression with a single X, R^2 = the square of the correlation coefficient between X and Y

The Standard Error of the Regression (SER)

The *SER* measures the spread of the distribution of u. The *SER* is (almost) the sample standard deviation of the OLS residuals:

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

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

(the second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$).

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

The *SER*:

- has the units of u, which are the units of Y
- measures the 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*:

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

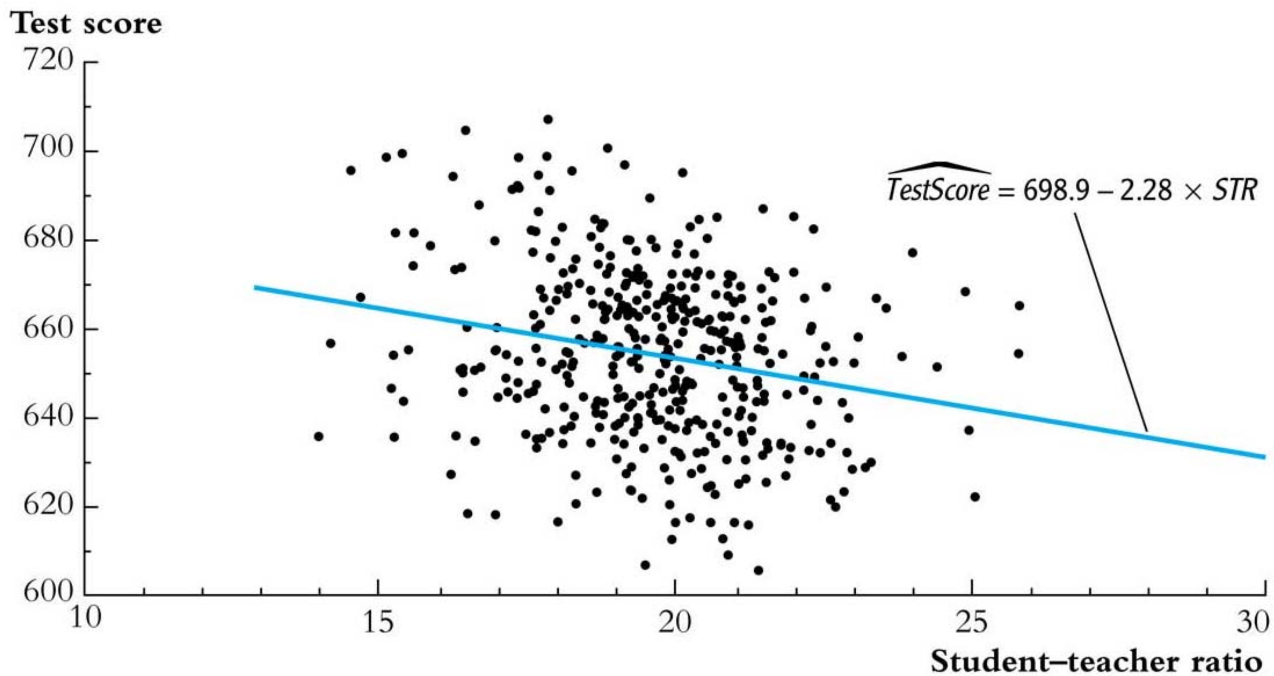
This measures the same thing as the *SER* – the minor difference is division by $1/n$ instead of $1/(n–2)$.

Technical note: why divide by n–2 instead of n–1?

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

- Division by n–2 is a "degrees of freedom" correction – just like division by n–1 in s_Y^2, except that for the *SER*, two parameters have been estimated (β_0 and β_1, by $\hat{\beta}_0$ and $\hat{\beta}_1$), whereas in s_Y^2 only one has been estimated (μ_Y, by $\bar{Y}$).
- When n is large, it makes negligible difference whether n, n–1, or n–2 are used – although the conventional formula uses n–2 when there is a single regressor.
- For details, see Section 17.4

Example of the R^2 and the *SER*



$$\widehat{TestScore} = 698.9 - 2.28 * STR, R^2 = 0.05, SER = 18.6$$

STR explains only a small fraction of the variation in test scores. Does this make sense? Does this mean the STR is unimportant in a policy sense?

The Least Squares Assumptions
(SW Section 4.4)

What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it? Under what conditions is it an unbiased estimator of the true population parameters?

To answer these questions, we need to make some assumptions about how Y and X are related to each other, and about how they are collected (the sampling scheme)

These assumptions – there are three – are known as the Least Squares Assumptions.

The Least Squares Assumptions

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

1. The conditional distribution of u given X has mean zero, that is, $E(u|X = x) = 0$.

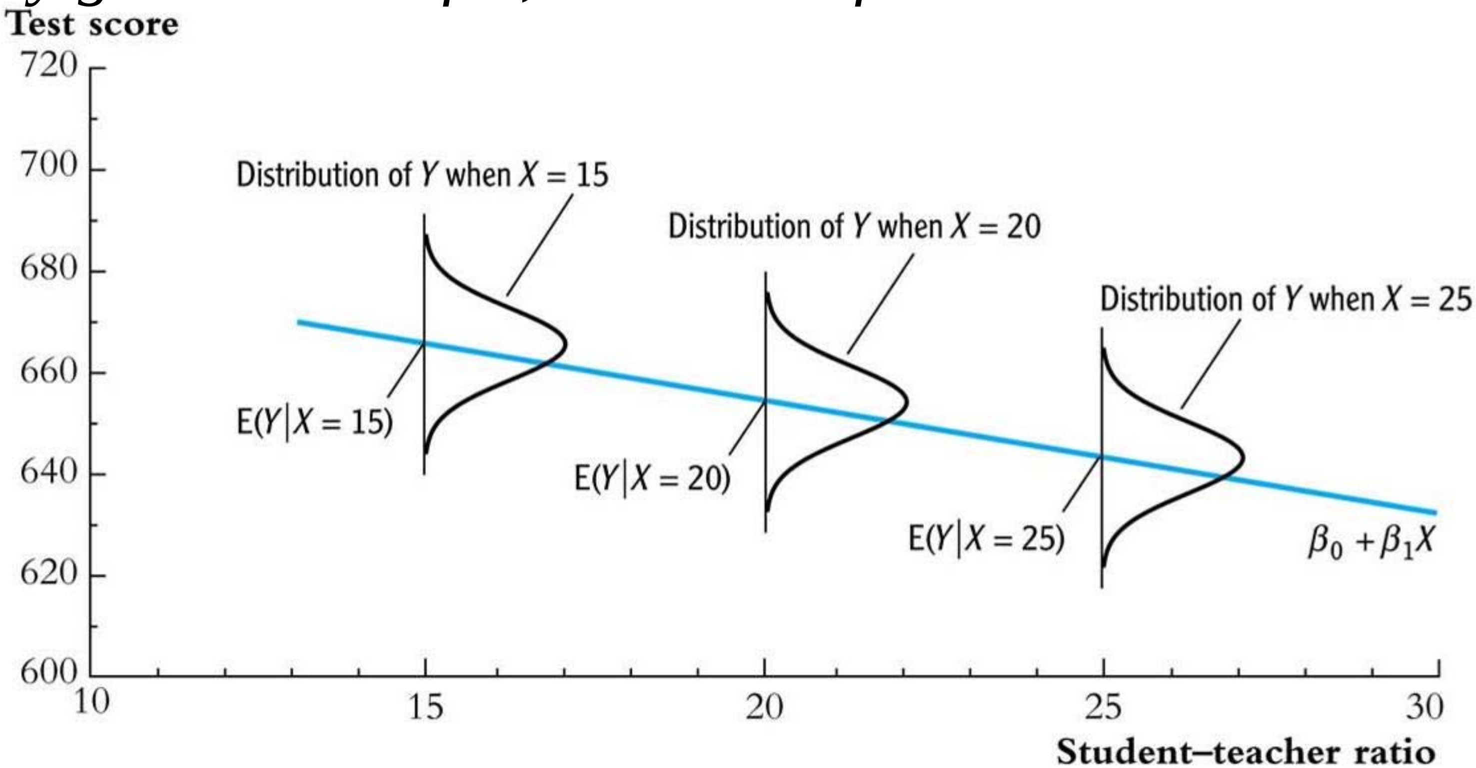
   This implies that $\hat{\beta}_1$ is unbiased

2. (X_i,Y_i), $i = 1,\ldots,n$, are i.i.d.
   - *This is true if X, Y are collected by simple random sampling*
   - *This delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$*

3. Large outliers in X and/or Y are rare.
   - *Technically, X and Y have finite fourth moments*
   - *Outliers can result in meaningless values of $\hat{\beta}_1$*

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

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



Example: $Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$, u_i = other factors

- What are some of these "other factors"?
- Is $E(u|X=x) = 0$ plausible for these other factors?

Least squares assumption #1, ctd.

A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment:

- X is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer – using no information about the individual.
- Because X is assigned randomly, all other individual characteristics – the things that make up u – are independently distributed of X
- Thus, in an ideal randomized controlled experiment, $\boldsymbol{E(u|X = x) = 0}$ (that is, LSA #1 holds)
- In actual experiments, or with observational data, we will need to think hard about whether $E(u|X = x) = 0$ holds.

Least squares assumption #2: (X_i, Y_i), $i = 1,\ldots,n$ are i.i.d.

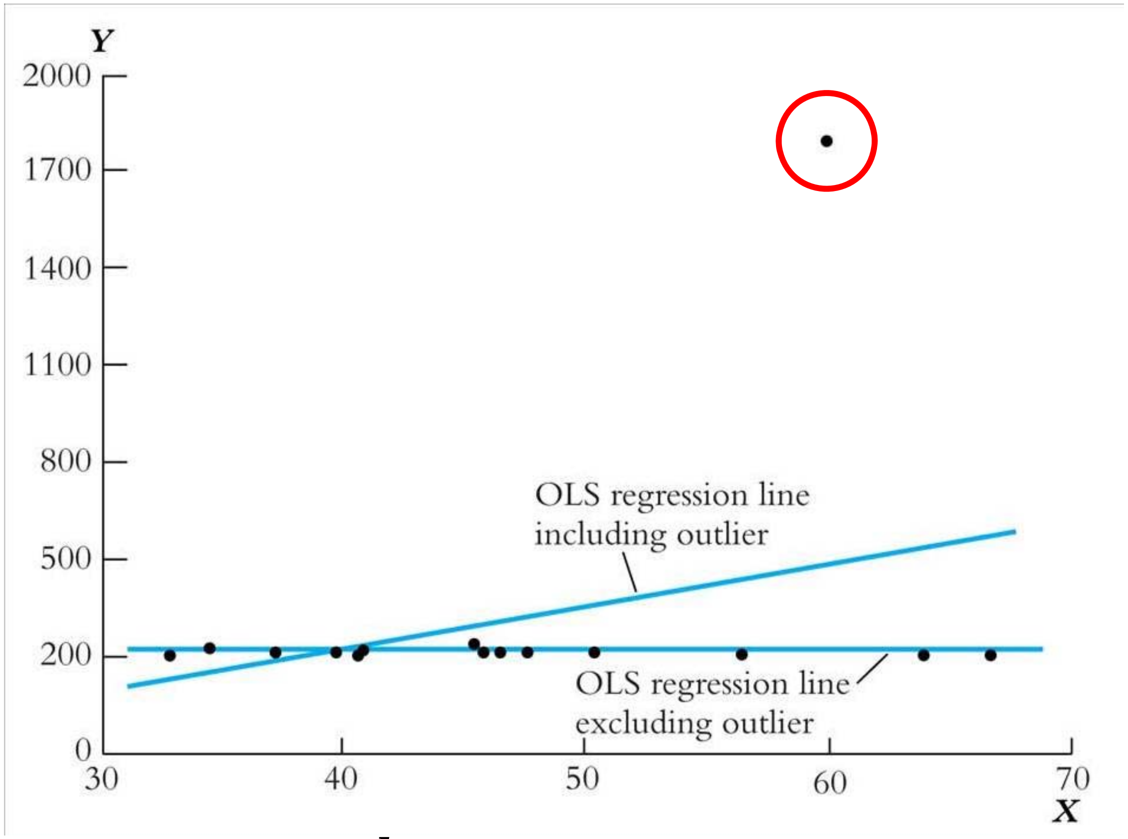
This arises automatically if the entity (individual, district) is sampled by simple random sampling: the entity is selected then, for that entity, X and Y are observed (recorded).

The main place we will encounter non-i.i.d. sampling is when data are recorded over time ("time series data") – this will introduce some extra complications.

Least squares assumption #3: *Large outliers are rare* *Technical statement:* $E(X^4) < \infty$ and $E(Y^4) < \infty$

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; *STR*, family income, etc. satisfy this too).
- However, the substance of this assumption is that a large outlier can strongly influence the results

OLS can be sensitive to an outlier:



- *Is the lone point an outlier in X or Y?*
- In practice, outliers often are data glitches (coding/recording problems) – so check your data for outliers! The easiest way is to produce a scatterplot.

The Sampling Distribution of the OLS Estimator

(SW Section 4.5)

The OLS estimator is computed from a sample of data; a different sample gives a different value of $\hat{\beta}_1$. This is the source of the "sampling uncertainty" of $\hat{\beta}_1$. We want to:

- 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
- All these require figuring out the sampling distribution of the OLS estimator. Two steps to get there…
  - Probability framework for linear regression
  - Distribution of the OLS estimator

Probability Framework for Linear Regression

The probability framework for linear regression is summarized by the three least squares assumptions.

Population

The group of interest (ex: all possible school districts)

Random variables*: *Y*, *X

Ex: (*Test Score*, *STR*)

***Joint distribution of* (*Y*, *X*)**

The population regression function is linear

$E(u|X) = 0$ (1st Least Squares Assumption)

X, Y have finite fourth moments (3rd L.S.A.)

Data Collection by simple random sampling:

$\{(X_i, Y_i)\}$, $i = 1,\ldots, n$, are i.i.d. (2nd L.S.A.)

The Sampling Distribution of $\hat{\beta}_1$

Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.

- What is $E(\hat{\beta}_1)$? (where is it centered?)
  - If $E(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
- What is $\text{var}(\hat{\beta}_1)$? (measure of sampling uncertainty)
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - It can be very complicated in general
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - It turns out to be relatively simple – in large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$

Some preliminary algebra:

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

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$$

so $$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$$

Thus,

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

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

so $$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Now $$\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\sum_{i=1}^{n}(X_i - \bar{X})\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\left(\sum_{i=1}^{n}X_i\right) - n\bar{X}\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i$$

Substitute $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ into the expression for $\hat{\beta}_1 - \beta_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

so

$$\boxed{\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Now we can calculate $E(\hat{\beta}_1)$ and $var(\hat{\beta}_1)$:

$$E(\hat{\beta}_1) - \beta_1 = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

$$= E\left\{E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \Bigg| X_1,\ldots,X_n\right\}$$

$= 0$ **because $E(u_i|X_i=x) = 0$ by LSA #1**

- Thus LSA #1 implies that $E(\hat{\beta}_1) = \beta_1$
- That is, $\hat{\beta}_1$ **is an unbiased estimator of β_1.**
- For details see App. 4.3

Next calculate var($\hat{\beta}_1$):

write

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i$. If n is large, $s_X^2 \approx \sigma_X^2$ and $\frac{n-1}{n} \approx 1$, so

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n}v_i}{\sigma_X^2},$$

where $v_i = (X_i - \bar{X})u_i$ (see App. 4.3). Thus,

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

so $$\text{var}(\hat{\beta}_1 - \beta_1) = \text{var}(\hat{\beta}_1)$$

$$= \frac{\text{var}(v)/n}{(\sigma_X^2)^2}$$

so

$$\boxed{\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}} \quad .$$

Summary so far

- $\hat{\beta}_1$ is unbiased: $E(\hat{\beta}_1) = \beta_1$ – just like $\bar{Y}$!
- $\text{var}(\hat{\beta}_1)$ is inversely proportional to n – just like $\bar{Y}$!

What is the sampling distribution of $\hat{\beta}_1$?

The exact sampling distribution is complicated – it depends on the population distribution of (Y, X) – but when n is large we get some simple (and good) approximations:

(1) Because var($\hat{\beta}_1$) $\propto$ $1/n$ and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$

(2) When n is large, the sampling distribution of $\hat{\beta}_1$ is well approximated by a normal distribution (CLT)

Recall the ***CLT***: suppose $\{v_i\}$, $i = 1,\ldots, n$ is i.i.d. with $E(v) = 0$ and var(v) = σ^2. Then, when n is large, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0,\sigma_v^2/n)$.

Large-*n* approximation to the distribution of $\hat{\beta}_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}, \text{ where } v_i = (X_i - \bar{X})u_i$$

- When n is large, $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$, which is i.i.d. (*why*?) and var(v_i) $< \infty$ (*why*?). So, by the CLT, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0, \sigma_v^2/n)$.
- Thus, for n large, $\hat{\beta}_1$ is approximately distributed

$$\boxed{\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right), \text{ where } v_i = (X_i - \mu_X)u_i}$$

The larger the variance of *X*, the smaller the variance of $\hat{\beta}_1$

The math

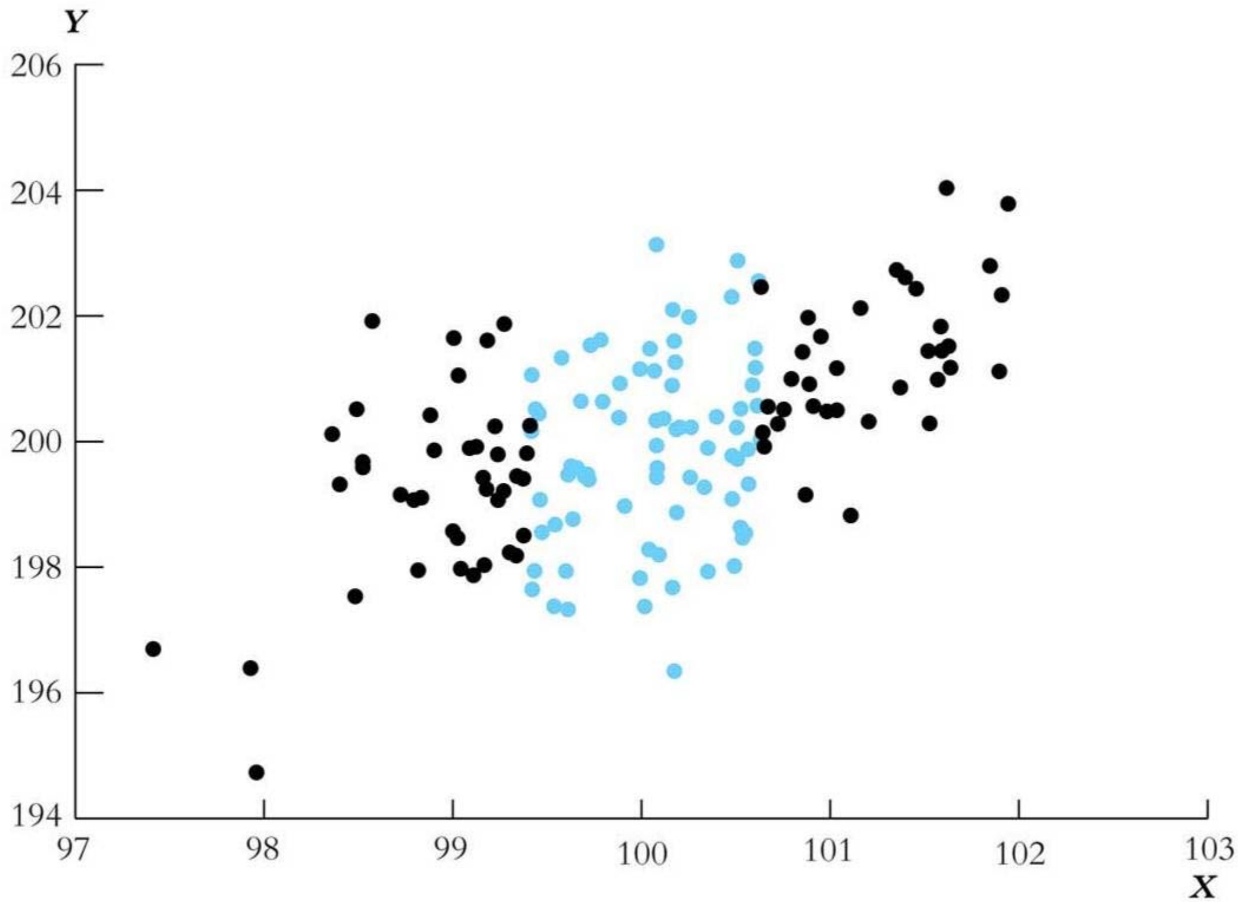
$$\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}$$

where $\sigma_X^2 = \text{var}(X_i)$. The variance of X appears in its square in the denominator – so increasing the spread of X decreases the variance of β_1.

The intuition

If there is more variation in X, then there is more information in the data that you can use to fit the regression line. This is most easily seen in a figure…

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$



There are the same number of black and blue dots – using which would you get a more accurate regression line?

Summary of the sampling distribution of $\hat{\beta}_1$:

If the three Least Squares Assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:
  - $E(\hat{\beta}_1) = \beta_1$ (that is, $\hat{\beta}_1$ is unbiased)
  - $\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4} \propto \frac{1}{n}$.
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (X,u)
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (that is, $\hat{\beta}_1$ is consistent)
- When n is large, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0,1)$ (CLT)
- *This parallels the sampling distribution of* $\bar{Y}$.

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 $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\operatorname{var}[(X_i - \mu_X)u_i]}{[\operatorname{var}(X_i)]^2}. \tag{4.21}$$

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

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\operatorname{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left(\frac{\mu_X}{E(X_i^2)}\right)X_i. \tag{4.22}$$

We are now ready to turn to hypothesis tests & confidence intervals…