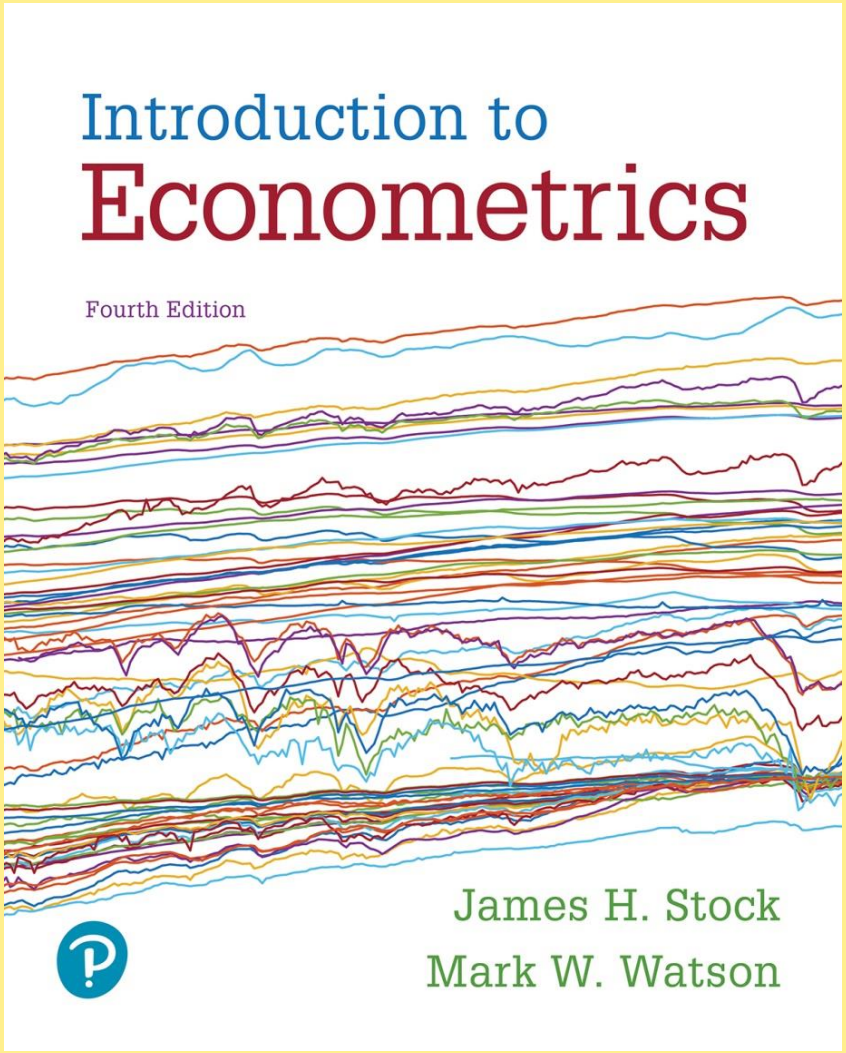
Introduction to Econometrics

Fourth Edition



Chapter 4

Linear Regression with One Regressor

Outline

1. The population linear regression model
2. The ordinary least squares (OLS) estimator and the sample regression line
3. Measures of fit of the sample regression
4. The least squares assumptions for causal inference
5. The sampling distribution of the OLS estimator
6. The least squares assumptions for prediction

Linear regression lets us estimate the population regression line and its slope

- The population regression line is the **expected value** of Y given X.
- The slope is the **difference** in the **expected values** of Y, for two values of X that differ by one unit
- The estimated regression can be used either for:
  - **causal inference** (learning about the causal effect on Y of a change in X)
  - **prediction** (predicting the value of Y given X, for an observation not in the data set)
- **Causal inference** and **prediction** place different requirements on the data – but both use the same regression toolkit.

The problem of statistical inference for linear regression is, 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?
    - Answer: ordinary least squares (OLS).
  - What are advantages and disadvantages of OLS?
- Hypothesis testing:
  - How to test whether the slope is zero?
- Confidence intervals:
  - How to construct a confidence interval for the slope?

Class size is measured by the "student–teacher ratio".

Smaller class size means that children get more attention from teachers, but also more teachers, which means higher cost for running school.

We can write this as a mathematical relationship using the Greek letter beta, $\beta_{ClassSize}$, where the subscript *ClassSize* distinguishes the effect of changing the class size from other effects. Thus,

$$\beta_{ClassSize} = \frac{Change\ in\ Test\ Score}{Change\ in\ Class\ Size} = \frac{\Delta_{Test\ Score}}{\Delta_{Class\ Size}}$$

where the Greek letter Δ (delta) stands for "change in."

That is, $\beta_{ClassSize}$ is the change in the test score that results from changing the class size divided by the change in the class size.

The Linear Regression Model (SW Section 4.1)

The ***population regression line***:

$$Test\ Score = \beta_0 + \beta_1 STR$$

β_1 = slope of population regression line

Why are β_0 and β_1 "population" parameters?

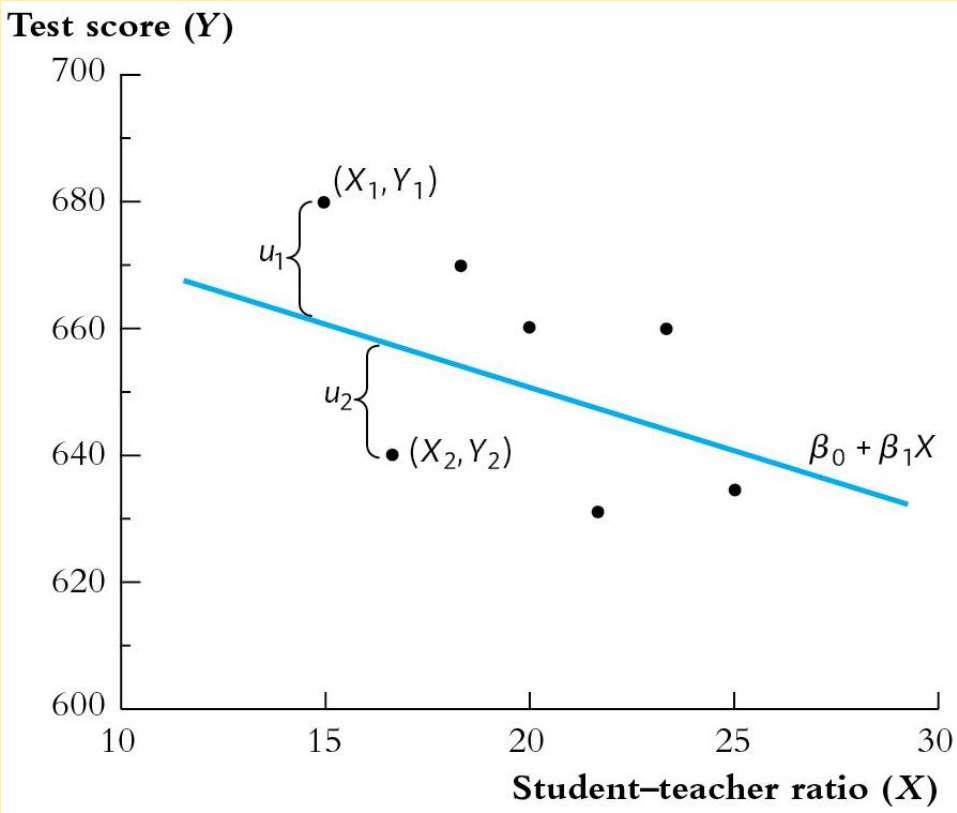
- We would like to know the population value of β_1 .
- We don't know β_1, so must estimate it using data.

The Population Linear Regression Model

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

- We have *n* observations, (X_i, Y_i), $i = 1, .., n$.
- 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. In general, these omitted factors are other factors that influence Y, other than the variable X. The regression error also includes error in the measurement of Y.

The population regression model in a picture: Observations on *Y and X (n = 7); the population regression line; and the* regression error (the "error term")



Derivation of OLS estimator

$\hat{\beta}_0$ and $\hat{\beta}_1$ from

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

(1) Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of the errors. That is, select $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$S = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)$$

(2) Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of the absolute errors. That is, select $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$S = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|$$

(3) Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the sum of the squared errors. That is, select $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$S = \sum_{i=1}^{n} e_1^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$S = \sum_{i=1}^{n} e_1^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) = 0 \qquad \sum_{i=1}^{n} \left(Y_i \right) = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

More simply (dividing both sides by n before the second step):

$$\Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \qquad \text{(N1)}$$

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \quad (N1)$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = \frac{\partial \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(-X_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)(X_i) = 0$$

Or, expanding the sum out:

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad (N2)$$

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \quad (N1)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad (N2)$$

Equations (N1) and (N2) are known as the normal equations.

Notice **that** there are two normal equations and two unknowns, **so** we can solve these two equations for those parameters and **that** will provide our OLS estimators.

Equation (N1) suggests **that** an estimator for the intercept of the regression line is

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \ (*1)$$

Notice also **that** the normal equation (N1) implies **that** the ***regression line*** passes through the point $(\bar{X}, \bar{Y})$.

$$\sum_{i=1}^{n} Y_i = n + \hat{\beta}_1 \sum_{i=1}^{n} X_i \Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \quad (N1)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad (N2)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \ (*1)$$

Multiplying equation (*1) by $\sum_{i=1}^{n} X_i$ and subtracting n times equation (N2) and solving in terms of $\hat{\beta}_1$ yields an estimator for the slope of the regression line:

$$\hat{\beta}_1 = \frac{n\sum_i^n Y_i X_i - \sum_i^n Y_i \sum_i^n X_i}{n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)^2}, \ (*2)$$

$$\hat{\beta}_1 = \frac{n\sum_i^n Y_i X_i - \sum_i^n Y_i \sum_i^n X_i}{n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)^2}, \text{ (*2)}$$

This expression can be written in many ways, but two of the more useful versions are:

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

$$\hat{\beta}_1 = \frac{n\sum_i^n Y_i X_i - \sum_i^n Y_i \sum_i^n X_i}{n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)^2}, \quad (*2)$$

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

where the second equality follows from the fact that:

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

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

$$\text{å}_{i=1}^n (Y_i - \bar{Y})\bar{X} = \bar{X}(\text{å}_{i=1}^n Y_i + n\bar{Y}) = \bar{X}\text{å}_{i=1}^n Y_i + n\bar{X}\bar{Y}$$

$$= n\bar{X}(1/n)\text{å}_{i=1}^n Y_i + n\bar{X}\bar{Y} = n\bar{X}\bar{Y} - n\bar{X}\bar{Y} = 0.$$

$$\hat{\beta}_1 = \frac{n\sum_i^n Y_i X_i - \sum_i^n Y_i \sum_i^n X_i}{n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)^2}, \quad (*2)$$

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

Similarly, using the same logic, we could write this expression as

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

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

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)$$

$$= \sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\bar{X}^2$$

$$= \left(\sum_{i=1}^{n} X_i^2\right) - 2n\bar{X}\bar{X} + n\bar{X}^2$$

$$= \left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2$$

Notice also that

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

which is the ratio of the estimated covariance between Y and X and the estimated variance of X.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \text{ (*1)}$$

$$\hat{\beta}_1 = \frac{n\sum_i^n Y_i X_i - \sum_i^n Y_i \sum_i^n X_i}{n\sum_i^n X_i^2 - \left(\sum_i^n X_i\right)^2}, \text{ (*2)}$$

Thus, equations (*1) and (*2) in their various forms provide estimators for the ***slope*** and ***intercept*** of the population regression line.

Notice that no other estimators will make the sum of squared errors smaller than the ordinary least squares estimator.

The Ordinary Least Squares Estimator (SW Section 4.2)

How can we estimate β_0 and β_1 from data?

Recall that was the least squares estimator of μ_Y: $\bar{Y}$ solves,

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

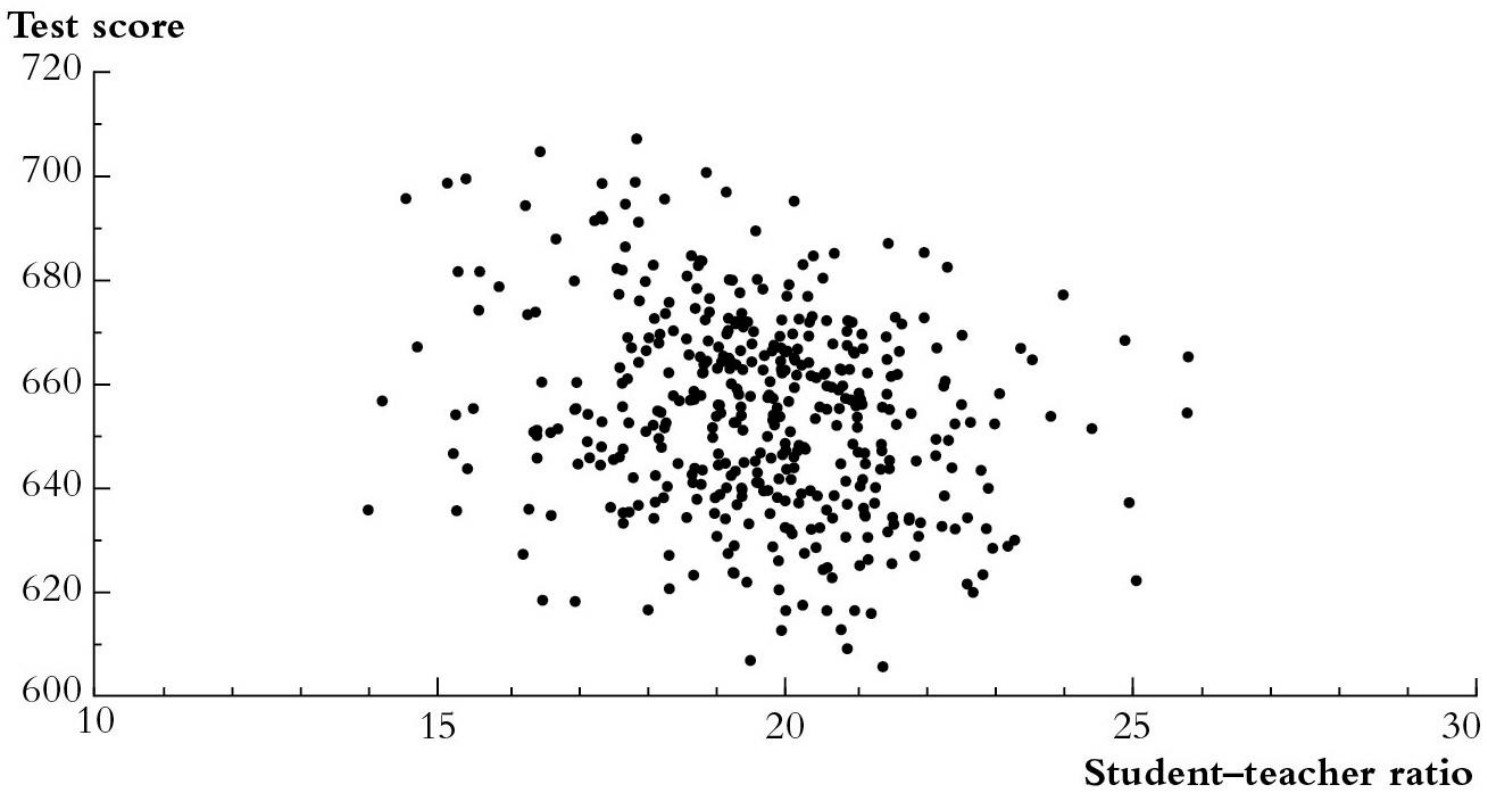
By analogy, **we will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters β_0 and β_1.** The OLS estimator solves,

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

Mechanics of OLS

The population regression line: $E(Test\,Score|STR) = \beta_0 + \beta_1 STR$

$$\beta_1 = slope = ??$$



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.**

Key Concept 4.2: The OLS Estimator, Predicted Values, and Residuals

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

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (4.8)$$

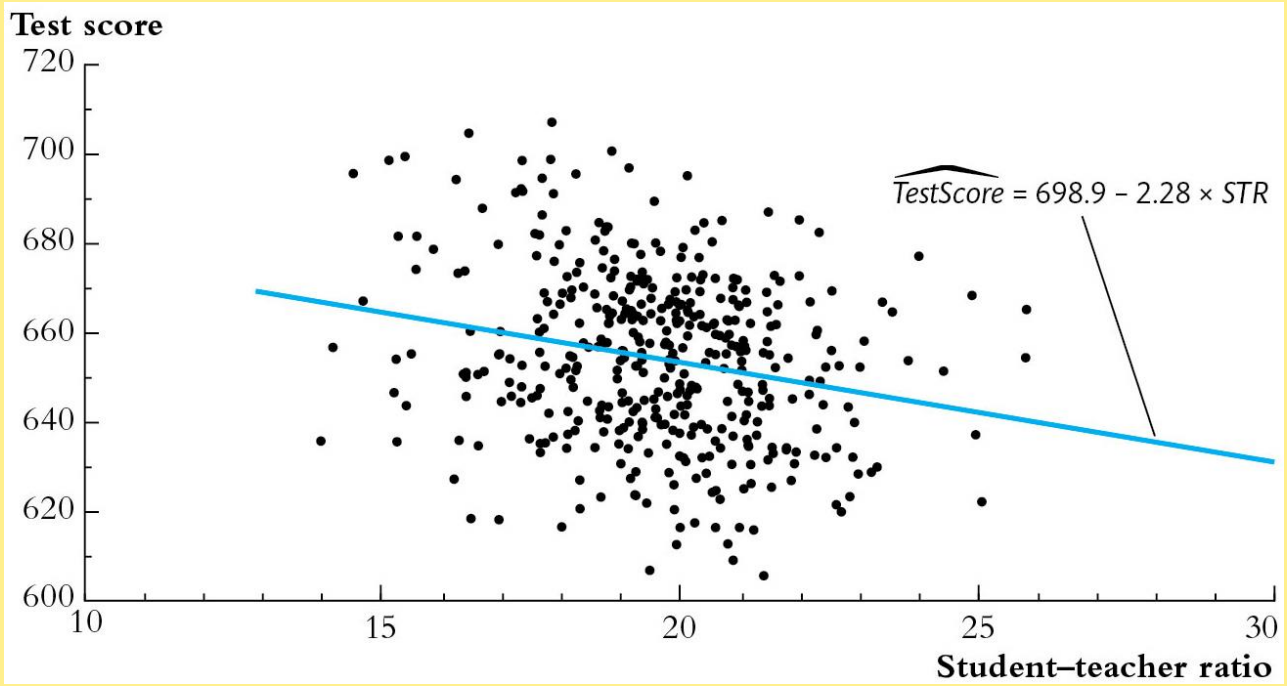
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

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

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

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

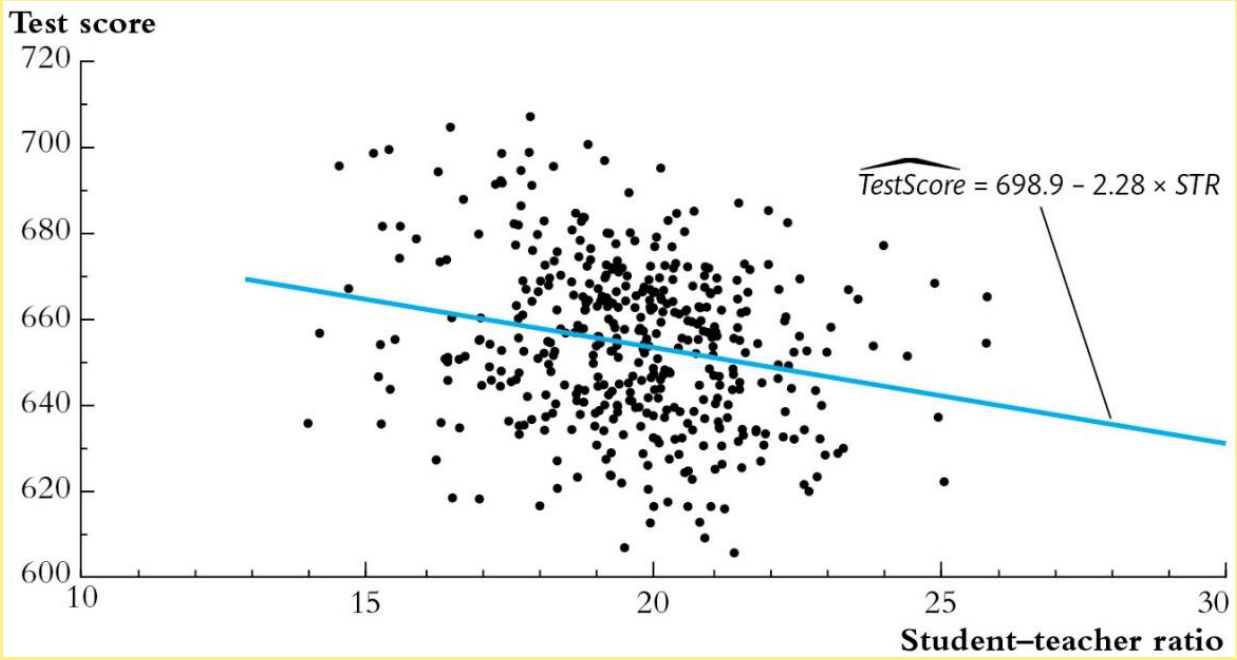


- Estimated slope $= \hat{\beta}_1 = -2.28$
- Estimated intercept $= \hat{\beta}_0 = 698.9$
- Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Interpretation of the estimated slope and intercept

- $\widehat{TestScore} = 698.9 - 2.28 \times STR$
- **Districts with one more student per teacher on average have test scores that are 2.28 points lower.**
- That is, $\frac{\Delta E(Test\ score|STR)}{\Delta STR} = -2.28$
- 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. But 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
------------------------------------------------------------------------------
```

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

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

Measures of Fit (SW Section 4.3)

Two regression statistics provide complementary measures of how well the regression line "fits" or explains the data:

- 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

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

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

$$\rightarrow \text{total sum of squares} = \text{"explained" SS} + \text{"residual" SS}$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

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

$$Model\ SS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

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*:

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

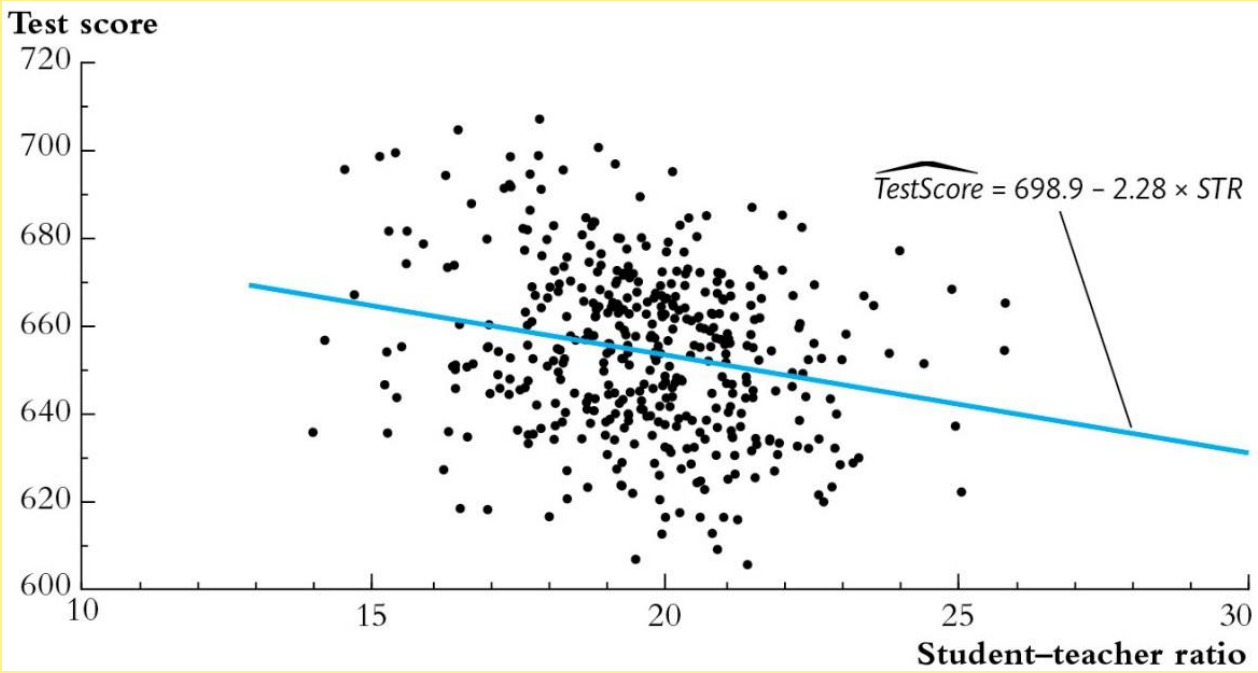
This measures the same thing as the *SER* – the minor difference is division by 1/*n* instead of 1/(*n* – 2).

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, 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 $\overline{Y}$).
- When n is large, it doesn't matter 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 18.4

Example of the R^2 and the *SER*



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

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?

Degrees of freedom and adjusted R^2

If we have n observations and estimate three regression parameters, $y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\varepsilon_i$, then there are only (n-3) residuals to vary, **because** given any (n-3) residuals, the remaining residuals can be obtained by solving equations as follows: 3 equations, 3 unknowns

$$\sum \hat{\varepsilon}_i = 0, \sum x_{1i}\hat{\varepsilon}_i = 0, \sum x_{2i}\hat{\varepsilon}_i = 0$$

Here we say that the degree of freedom in this estimation is n-3.

The estimate of the residual variance $\hat{\sigma}^2_{\varepsilon_i}$ is given by the residual sum of squares (RSS) over degrees of freedom.

$$\hat{\sigma}^2_{\varepsilon_i} = \frac{RSS}{n-(k+1)}$$

$$\hat{\sigma}_{\varepsilon_i}^2 = \frac{RSS}{n-(k+1)}$$

As we increase the number of explanatory variables, RSS declines, but there is a decrease in the degrees of freedom as well.

What happens to the variance of residuals $\hat{\sigma}_{\varepsilon_i}^2$ depends on the proportionate decrease in the numerator and the denominator.

Thus there will a point when the variance of residuals $\hat{\sigma}_{\varepsilon_i}^2$ will actually **start increasing as** we add more explanatory variables.

It is often suggested that **we should choose the set of variables for which the variance of residuals** $\hat{\sigma}_{\varepsilon_i}^2$ **is the minimum**.

$$\hat{\sigma}_{\varepsilon_i}^2 = \frac{RSS}{n-(k+1)}$$

This is also the reason why, in multiple regression problems, it is customary to report what is known as adjusted R^2, denoted by adj R^2.

The measure R^2 ($R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$) defined earlier keeps on increasing (until it reaches 1.0) as we add extra explanatory variables and thus does not take account of the degrees of freedom problem.

$$\hat{\sigma}_{\varepsilon_i}^2 = \frac{RSS}{n - (k + 1)}$$

The adjusted R^2 is simply R^2 adjusted for degrees of freedom.

It is defined by the relation

$$1 - \bar{R}^2 = \frac{n-1}{n-(k+1)}(1 - R^2)$$

where k is the number of regressors.

We subtract (k+1) from n because we estimate a constant term in addition to the coefficients of these k regressors.

The above equation can be written as follows

$$\frac{(1-\bar{R}^2)\sum(y_i-\bar{y})^2}{n-1} = \frac{(1-R^2)\sum(y_i-\bar{y})^2}{n-k-1} = \hat{\sigma}^2$$

$$\frac{(1-\bar{R}^2)\sum(y_i-\bar{y})^2}{n-1} = \frac{(1-R^2)\sum(y_i-\bar{y})^2}{n-k-1} = \hat{\sigma}^2$$

Since the sum of squared deviations of y_i and n are constant, as we increase the number of regressors included in the equation, $\hat{\sigma}^2$ and (1- adjusted R^2) move in the same direction **as** $\hat{\sigma}^2$ and adjusted R^2 move in the opposite direction.

Thus the set of variables that gives minimum $\hat{\sigma}^2$ is also the set that maximizes adjusted R^2.

$$1-\bar{\boldsymbol{R}}^2 = \frac{n-1}{n-(k+1)}(1-\boldsymbol{R}^2)$$

$$\bar{\boldsymbol{R}}^2 = 1-\frac{(1-\boldsymbol{R}^2)(n-1)}{(n-k-1)}$$

$$\bar{R}^2 = \left\{1 - \left[\frac{(1 - R^2)(n-1)}{(n - k \uparrow -1) \downarrow}\right] \uparrow\right\} \downarrow$$

So, if R^2 does not increase significantly on the addition of a new independent variable, then the value of $\bar{R}^2$ will actually decrease.

$$\bar{R}^2 = \left\{1 - \left[\frac{\downarrow (1 - R^2 \uparrow)(n-1)}{(n - k \uparrow -1) \downarrow}\right] \downarrow\right\} \uparrow$$

On the other hand, if on adding the new independent variable we see a significant increase in R^2 value, then the $\bar{R}^2$ value will also increase.

The Least Squares Assumptions for Causal Inference (SW Section 4.4)

- So far we have treated OLS as a way to draw a straight line through the data on *Y* and *X*. Under what conditions does the slope of this line have a causal interpretation? That is, when will the OLS estimator be unbiased for the causal effect on *Y* of *X*?
- What is the variance of the OLS estimator over repeated samples?
- 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 for Causal Inference.

Definition of Causal Effect

- The causal effect on *Y* of a unit change in *X* is the expected difference in *Y* as measured in a randomized controlled experiment
  - For a binary treatment, the causal effect is the expected difference in means between the treatment and control groups, as discussed in Ch. 3
- With a binary treatment, for the difference in means to measure a causal effect requires random assignment or as-if random assignment.
  - Random assignment ensures that the treatment (*X*) is uncorrelated with all other determinants of *Y*, so that there are no confounding variables
- The least squares assumptions for causal inference generalize the binary treatment case to regression.

The Least Squares Assumptions for Causal Inference

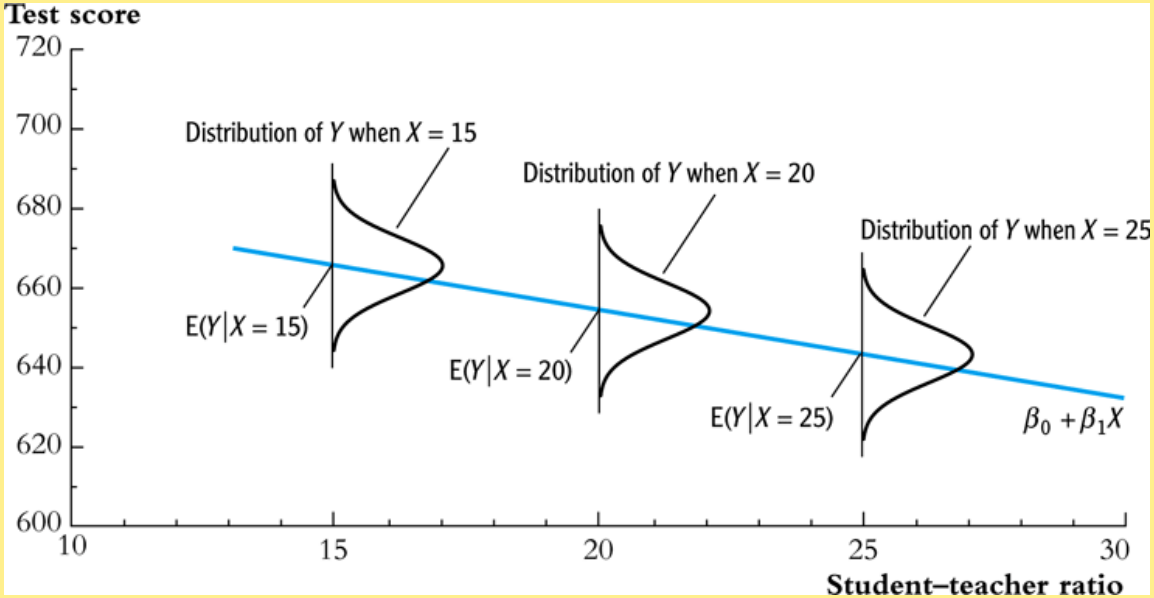
Let β_1 be the causal effect on *Y* of a change in *X:*

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

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 for the causal effect* β_1
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: $E(u|X = x) = 0$ (1 of 2) (note: u and x are not correlated)

When β_1 is the causal effect, 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: $E(u|X = x) = 0$ (2 of 2)

- The benchmark for understanding 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 distributed independently of *X*, so *u* and *X* are independent
- 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:

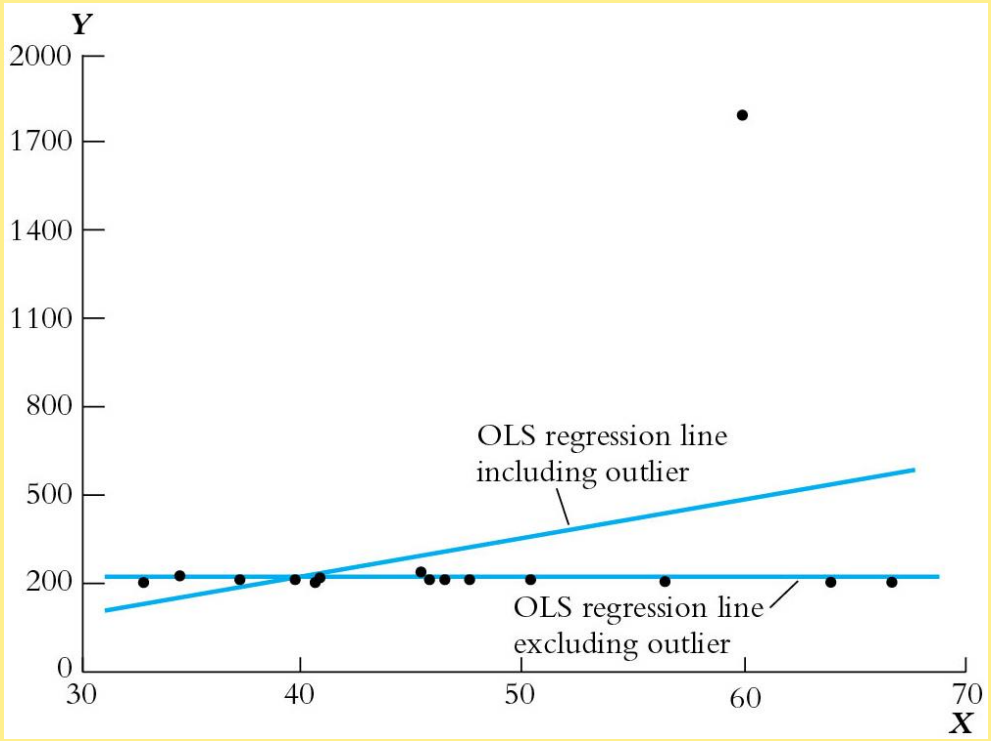
- The entities are selected from the same population, so (X_i, Y_i) are *identically distributed* for all $i = 1, \ldots, n$.
- The entities are selected at random, so the values of (X, Y) for different entities are *independently distributed*.

The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) – we will deal with that complication when we cover panel data.

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.)
- The substance of this assumption is that a large outlier can strongly influence the results – so we need to rule out large outliers.
- Look at your data! If you have a large outlier, is it a typo? Does it belong in your data set? Why is it an outlier?

OLS can be sensitive to an outlier



- *Is the lone point an outlier in X or Y?*
- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. Plot your data!

The Sampling Distribution of the OLS Estimator (SW Section 4.5)

The OLS estimator is computed from a sample of data. A different sample yields 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*). We assume:**

- The population regression function is linear
- $E(u|X) = 0$ (1st Least Squares Assumption)
- X, Y have nonzero finite fourth moments (3rd L.S.A.)

Data Collection by simple random sampling implies:

- $\{(X_i, Y_i)\}$, $i = 1,\ldots, n$, 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 $E(\hat{\beta}_1)$?
  - If $E(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
- What is var($\hat{\beta}_1$)? (measure of sampling uncertainty)
  - We need to derive a formula so we can compute the standard error of β_1.
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - It is very complicated in general
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - In large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$ (1 of 3)

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,

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

The mean and variance of the sampling distribution of $\hat{\beta}_1$ (2 of 3)

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

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

The mean and variance of the sampling distribution of $\hat{\beta}_1$ (3 of 3)

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

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

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

$$E(\hat{\beta}_1 | X_1, \ldots, X_n) = \beta_1 + E\left[\left.\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\right| X_1, \ldots, X_n\right]$$

$$= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})E(u_i | X_1, \ldots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

By the second least squares assumption, u_i is distributed independently of X, so $E(u_i|X_1,\ldots,X_n)=0$.

$= 0$ **because $E(u_i \mid X_i = x) = 0$ 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($\hat{\beta}_1$) (1 of 2)

write

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

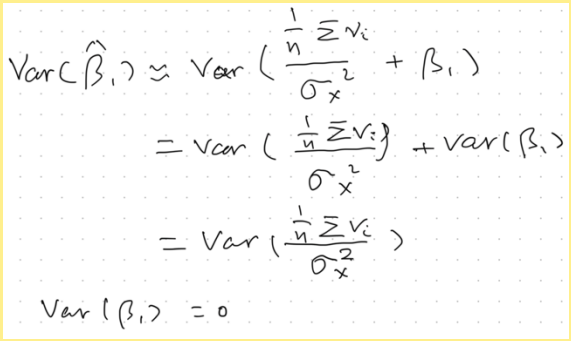
where $v_i = (X_i - \bar{X})u_i$. If n is large, $s_X^2 \approx \sigma_X^2$ and $\frac{n-1}{n} \approx 1$, so

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

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

Next calculate var($\hat{\beta}_1$) (2 of 2)

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



$$E(\hat{\beta}_1 - \beta_1)^2 = var(\hat{\beta}_1)$$

so $$\text{var}(\hat{\beta}_1 - \beta_1) = \text{var}(\hat{\beta}_1) = \text{var}\left[\frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{(s_X^2)^2}\right] = \frac{\text{var}(v_i)/n}{(s_X^2)^2}$$

where the final equality uses assumption 2. Thus,

$$\text{var}(\hat{\beta}_1) = \frac{1}{n}\cdot\frac{\text{var}[(X_i - \mu_x)u_i]}{(s_X^2)^2}.$$

Summary so far

1. $\hat{\beta}_1$ is unbiased: under LSA#1, $E(\hat{\beta}_1) = \beta_1$ - just like $\bar{Y}$!
2. var($\hat{\beta}_1$) is inversely proportional to n - just like $\bar{Y}$!

Now calculate $var(\hat{\beta}_1)$

$$\hat{\beta}_1 \cong \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{var(X_i)} + \beta_1$$

$$var(\hat{\beta}_1) = \frac{var\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right)}{(\sigma_x^2)^2}$$

$$var(\hat{\beta}_1) = \frac{1}{n} \times \frac{var\left((X_i - \bar{X})u_i\right)}{(\sigma_x^2)^2}$$

Because **error term** is independent of X variable, **variance** of X and **variance of error** variable are also independent. Therefore, **variance of** X and the denominator can be cancelled out. Afterwards, it is the same as the usual one.

$$\hat{\beta}1 = \frac{\sum i = 1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x_i and y_i, respectively.

Substitute $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ into the equation for $\hat{\beta}_1$:

$$\hat{\beta}1 = \frac{\sum i = 1^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Expanding the terms:

$$\hat{\beta}1 = \beta_1 + \frac{\sum i = 1^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Thus, $\hat{\beta}_1$ is a combination of the true β_1 and a random noise term involving the errors ϵ_i. Now, let's derive the variance of $\hat{\beta}_1$.

$$\hat{\beta}1 = \beta_1 + \frac{\sum i = 1^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Since ϵ_i are independent and identically distributed random variables with mean 0 and variance σ^2, the variance of $\hat{\beta}_1$ comes from the second term in the above equation (the noise term).

We know that for a linear combination of independent random variables, the variance of the sum is the sum of the variances. Therefore, the variance of $\hat{\beta}_1$ is:

$$\mathrm{Var}(\hat{\beta}1) = \mathrm{Var}\left(\frac{\sum i = 1^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$$

Since the ϵ_is are independent, their variances sum up:

$$\mathrm{Var}(\hat{\beta}1) = \frac{\sum i = 1^n (x_i - \bar{x})^2 \cdot \mathrm{Var}(\epsilon_i)}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2}$$

Since $\mathrm{Var}(\epsilon_i) = \sigma^2$, we have:

$$\mathrm{Var}(\hat{\beta}1) = \frac{\sigma^2 \sum i = 1^n (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2}$$

Simplifying:

$$\mathrm{Var}(\hat{\beta}1) = \frac{\sigma^2}{\sum i = 1^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y}) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1)$$

Variance of $\bar{y}$:

The sample mean $\bar{y}$ is an average of the dependent variable values, so its variance is:

$$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n}$$

Thus:

$$\mathrm{Var}(\hat{\beta}0) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum i = 1^n (x_i - \bar{x})^2}$$

For a single explanatory variable OLS regression, degrees of Freedom for Variance of Residuals = n – 2

$$var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

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 $\text{var}(\hat{\beta}_1) \propto 1/n$ 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 $\text{var}(v) = \sigma^2$. 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 $\text{var}(v_i) < \infty$ (*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, $\hat{\beta}_1$s approximately distributed

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

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

The math

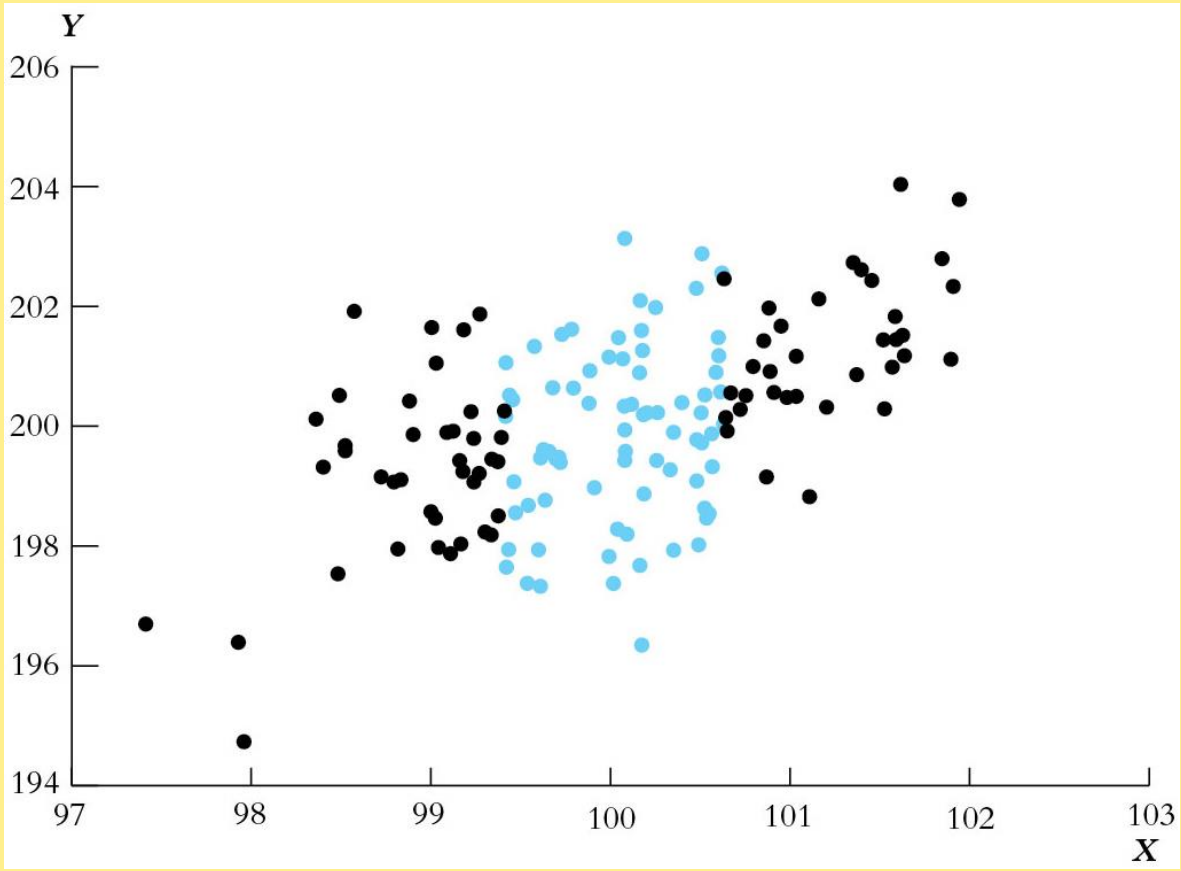
$$E(\hat{\beta}_1 - \beta_1)^2 = var(\hat{\beta}_1) \quad \text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}$$

Where $\sigma_X^2 = \text{var}(X_i)$. The variance of X appears (squared) 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 $\hat{\beta}_1$



The number of black and blue dots is the same. 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)
  - $\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4} \quad \propto \frac{1}{n}.$
- 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* $\bar{Y}$.

Key Concept 4.4: 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

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

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

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

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

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}_0)^2, \tag{3-1}$$

where $\mathbf{b}_0$ denotes the choice for the coefficient vector. In matrix terms, minimizing the sum of squares in (3-1) requires us to choose $\mathbf{b}_0$ to

$$\text{Minimize}_{\mathbf{b}_0}\ S(\mathbf{b}_0) = \mathbf{e}_0'\mathbf{e}_0 = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)'(\mathbf{y} - \mathbf{X}\mathbf{b}_0). \tag{3-2}$$

Expanding this gives

$$\mathbf{e}_0'\mathbf{e}_0 = \mathbf{y}'\mathbf{y} - \mathbf{b}_0'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0 \tag{3-3}$$

or

$$S(\mathbf{b}_0) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}_0 + \mathbf{b}_0'\mathbf{X}'\mathbf{X}\mathbf{b}_0.$$

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b}_0)}{\partial \mathbf{b}_0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_0 = \mathbf{0}.^2 \tag{3-4}$$

Let **b** be the solution. Then, after manipulating (3-4), we find that **b** satisfies the **least squares normal equations**,

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{3-5}$$

If the inverse of $\mathbf{X}'\mathbf{X}$ exists, which follows from the full column rank assumption (Assumption A2 in Section 2.3), then the solution is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3-6}$$

The two assumptions imply that

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_1\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n \mid \mathbf{X}] \\ E[\varepsilon_2\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_2\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n \mid \mathbf{X}] \\ \vdots & \vdots & \vdots & \vdots \\ E[\varepsilon_n\varepsilon_1 \mid \mathbf{X}] & E[\varepsilon_n\varepsilon_2 \mid \mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n \mid \mathbf{X}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix},$$

which we summarize in Assumption 4:

$$\boxed{\text{ASSUMPTION:} \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\mathbf{I}.} \tag{2-9}$$

By using the variance decomposition formula in (B-69), we find

$$\text{Var}[\boldsymbol{\varepsilon}] = E[\text{Var}[\boldsymbol{\varepsilon} \mid \mathbf{X}]] + \text{Var}[E[\boldsymbol{\varepsilon} \mid \mathbf{X}]] = \sigma^2\mathbf{I}.$$

- **The sampling error** is defined as $\mathbf{b} - \boldsymbol{\beta}$. It too can be related to $\boldsymbol{\varepsilon}$ as follows.

$$\begin{aligned}
\mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \quad \text{(by (1.2.5))} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \quad \text{(since } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ by Assumption 1.1)} \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \qquad (1.2.14)
\end{aligned}$$

$$\begin{aligned}
&Var(b) \\
&= E\{(b - b)(b - b)'\} \\
&= E\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\} \\
&= (X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

The Least Squares Assumptions for Prediction (SW Appendix 4.4) (1 of 2)

- Prediction entails using an estimation sample to estimate a prediction model, then using that model to predict the value of *Y* for an observation *not* in the estimation sample.
  - Prediction requires good out-of-sample performance.
- For prediction, β_1 is simply the slope of the population regression line (the conditional expectation of *Y* given *X*), which in general is not the causal effect.
- The critical LSA for Prediction is that the out-of-sample ("OOS") observation for which you want to predict *Y* comes from the same distribution as the data used to estimate the model.
  - This replaces LSA#1 for Causal Inference

The Least Squares Assumptions for Prediction (SW Appendix 4.4) (2 of 2)

1. The out of sample observation (X^{OOS}, Y^{OOS}) is drawn from the same distribution as the estimation sample (X_i, Y_i), $i = 1,\ldots,n$
   - *This ensures that the regression line fit using the estimation sample also applies to the out-of-sample data to be predicted.*
2. (X_i, Y_i), $i = 1,\ldots,n$ are i.i.d.
   - *This is the same as LSA#2 for causal inference*
3. Large outliers in X and/or Y are rare (X and Y have four moments)
   - *This is the same as LSA#3 for causal inference*