Applied Statistics and Econometrics

Lecture 4

Saul Lach

September 2017

Outline of Lecture 4

1. **Linear regression model with single regressor (SW 4.1)**
2. Estimation of parameters in linear regression model (SW 4.2)
3. Measures of fit (SW 4.3)
4. Properties of OLS
    1. The Least Squares assumptions (SW 4.4)
    2. Sampling distribution of OLS (SW 4.5)

A policy question

- **Test scores and class size.**
- A school district superintendent must decide whether to hire 40 new teachers.
- She faces a trade-off between costs of hiring and its benefits:
  - Costs: hiring increases expenditures by,say, $1,800,000.
  - Benefits: hiring reduces the student-per-teacher (STR) ratio by 2, from 22 to 20.
- The question is: how does a decrease in STR affect performance (measured by test scores).

A personal question

- **Wages and education.**
- You want to know what salary to expect after you graduate from college.
- You may also want to know whether to continue your education with a master degree or not.
- You face a trade-off between the cost (effort and money) of obtaining a master degree and its benefits.
- You are interested in answering the question: How does a MA degree affect wages?

Answers

- We propose to give a quantitative answer to these questions based on the following model:

$$Testscore = \beta_0 + \beta_1 \times STR$$

or

$$Wages = \gamma_0 + \gamma_1 \times educ$$

where *educ* is years of education.

- These equations are examples of straight lines in the $(Testscore, STR)$ $[(Wages, educ)]$ plane where β_0 $[\gamma_0]$ is the intercept and β_1 $[\gamma_1]$ is the slope.

Answers are given by slope (and intercept)

- These models will answer the questions posed:

$$\begin{aligned} \Delta Testscore &= \beta_1 \times \Delta STR \\ \Delta Wages &= \gamma_1 \times \Delta educ \end{aligned}$$

- When STR decreases by 2 the change in scores is

$$\Delta Testscore = \beta_1 \times (-2)$$

- When educ is 16 years (at graduation) wages will be

$$\gamma_0 + \gamma_1 16$$

and going for an MA degree (another year) will change wages by

$$\Delta Wages = \gamma_1 \times 1$$

- The slope of the linear equation **is the change in Testscore (or wages) when STR (or educ) is changed by $\Delta = 1$ units.**

Qualifing the answers

- As a model to predict testscores $Testscore = \beta_0 + \beta_1 STR$ is incomplete since we suspect there are other factors affecting scores (e.g., teacher quality, family income, etc.).
- Since these factors vary across school districts, this equation cannot hold for each individual district.
- We view equation as saying something about the relationship between test scores and STR **on average across all districts**.
- The model for **each** district incorporates these other factors which we lump together:

$$Testscore = \beta_0 + \beta_1 STR + \text{other factors}$$

  - In Lecture 6 we will "open up" this additional term.
- Similarly,

$$Wages = \gamma_0 + \gamma_1 \times educ + \text{other factors}$$

- These are examples of the **linear regression model with a single regressor**.

Linear Regression Model with a single regressor

- We introduce general notation for the model

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{Population Regression}} + \underbrace{u}_{\text{Other Factors}}$$

- The term $\beta_0 + \beta_1 X$ is the **population regression** line or function.
- u is called the **regression error** term and it captures all other things generating differences in Y between units having the same value of X.
  - For example, Test scores (Y) may differ because of different teacher quality and/or family income among schools having the same STR (X).
- Without the u all units having the same value of X are predicted to have the same Y which is unlikely in most realistic situations.

Notation and terminology

- The linear regression model in the **population** is

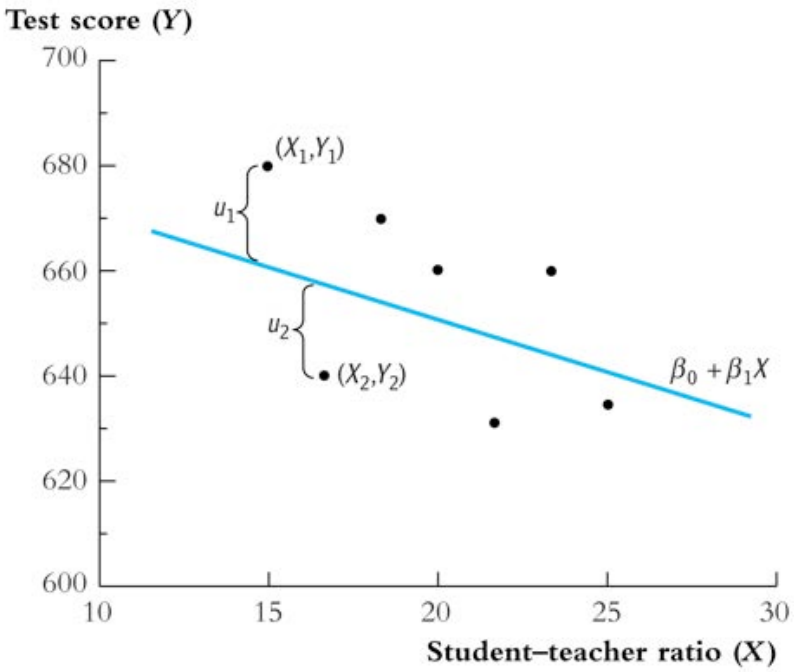
$$Y = \beta_0 + \beta_1 X + u$$

- The linear regression model in a **sample** of n observations is

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

- X is the independent variable or **regressor** (observed)
- Y is the dependent variable (observed)
- β_0 is the (unknown) intercept
- β_1 is the (unknown) slope
- u is the regression error (unobserved)

Terminology in a picture



Black dots are the data points (X_i, Y_i); blue line is the population regression line; deviation between data and line is the regression error u_i

Interpretation of beta1 in general case

- In general, β_1 is the change in Y when X changes by one unit (holding u constant).
- When X changes by Δ units, the change in Y is $\beta_1 \times \Delta$.
- If X is continuous we can examine small (infinitesimal) changes,

$$\frac{\partial Y}{\partial X} = \beta_1$$

- In all cases we change X and hold u constant.

Interpretation of beta0 in general case

- β_0 is the value Y takes when $X = 0$ (and $u = 0$).
- Often this does not make much economic sense.

What do we want to do?

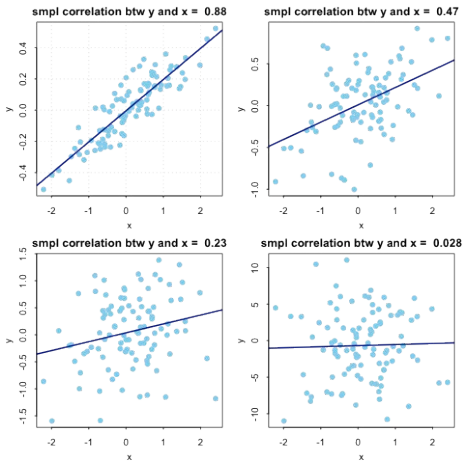
- We want to use the data on Y and X to estimate the unknown parameters β_0 and β_1.
- Knowledge of β_1 (and β_0) will enable us to answer many questions of interest (see examples at the beginning of this lecture).
- How do we do it?

Where are we?

1. Linear regression model with single regressor (SW 4.1)
2. **Estimation of parameters in linear regression model (SW 4.2)**
3. Measures of fit (SW 4.3)
4. Properties of OLS
    1. The Least Squares assumptions (SW 4.4)
    2. Sampling distribution of OLS (SW 4.5)

How to draw a line in the scatterplot?

- The data on X and Y are a cloud of points in a scatterplot of Y and X.



- The unknown population regression line passes through these points.
- We want to estimate such a line (i.e., its intercept β_0 and slope β_1).
- The slope of the line reflects the strength of the linear association between X and Y.

How to draw a line in the scatterplot?

- There are an infinite number of lines that one could possibly draw.
- How do we select one of them?
- Need some criterion. For example, choose the line that best fits the cloud of points ...what does "best fit" mean?
- We will use the **least squares** (LS) criterion: minimize squared deviations between observed points and the line.

Method of Least Squares- History

- Carl Friedrich Gauss is credited with developing the fundamentals of the basis for least-squares analysis in 1795 at the age of eighteen.
- But Adrien Marie Legendre, a french mathematician, was the first to publish the method.

The Least Squares method

- We encountered the LS method in Lecture 2: $\bar{Y}$ is the "least squares" estimator of $E(Y) = \mu_Y$.
- Indeed $\bar{Y}$ minimizes the squared deviations between observations Y_i and the chosen (flat) line at level m

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

- We find the optimal value of m by setting derivative wrt m to zero and solving:

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

$$\Longrightarrow \quad \hat{m} = \frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$$

- m is a flat line in the (Y,X) plane. LS chooses the best value for the "height" of this line.

The Least Squares method

- Suppose now we allow the line to have a slope.
- Generically the line is written as

$$b_0 + b_1 X$$

  - When we do not allow for a slope (as in previous slide) we have $b_0 = m$ and $b_1 = 0$.
- Given b_0 and b_1 and a value for X_i, the **predicted value** of Y_i is $\hat{Y}_i$ defined by

$$\hat{Y}_i = b_0 + b_1 X_i$$

- The Least Squares criterion is **to select values of b_0 and b_1** that give **predicted** values of Y that are "**closest in a LS sense**" to the **observed** data points
  - "**closest in a LS sense**" means that the value of the squared deviations between the data points and their predicted value is smallest.

The Least Squares method for the linear regression model

- The linear regression model in the **population** is

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

- Let

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

  SSR stands for sum of squared residuals.
- The **Ordinary Least Squares (OLS)** estimators of the intercept and slope are the values of b_0 and b_1 that minimize the SSR

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

- These minimizers – the solution to this minimization problem – are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.
- Compare this formulation to previous one without slope.

Ordinary Least Squares (OLS)

- 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 $(b_0 + b_1 X_i)$.
- The **residuals** are the deviations between data points Y_i and its predicted value based on b_0 and b_1 defined for each observation i : .

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i, \qquad i = 1, \ldots, n$$

- So, **OLS minimizes the sum of squared residuals**

$$\min_{b_0, b_1} SSR(b_0, b_1) = \min_{b_0, b_1} \sum_{i=1}^{n} \hat{u}_i^2$$

OLS predicted values and residuals

- In general, we are interested in the **OLS predicted values**

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

  i.e., the predicted values of Y based on the OLS estimator.
- And in the corresponding **OLS residuals**

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i, \qquad i = 1, \ldots, n$$

  based on the OLS estimator.
- When we talk about predicted values and residuals we mean OLS predicted values and residuals (unless stated otherwise).

Numerical example

- LS chooses b_0 and b_1 to minimize the sum of squared residuals

$$\min_{b_0,b_1} SSR(b_0, b_1) = \min_{b_0,b_1} \sum_{i=1}^{n} \hat{u}_i^2$$

Assume: $b_0 = -0.6$ and $b_1 = 0.14$

Observed data

y	x
-0.032	-0.626
-0.885	0.184
-7.024	-0.836
-2.677	1.595
1.539	0.330
1.785	-0.820
-5.943	0.487

Predictions

$\hat{Y}_i$
$-0.6 + 0.14 \times (-0.626) = -0.688$
$-0.6 + 0.14 \times (0.184) = -0.574$
$-0.6 + 0.14 \times (-0.836) = -0.717$
$-0.6 + 0.14 \times (1.595) = -0.377$
$-0.6 + 0.14 \times (0.330) = -0.554$
$-0.6 + 0.14 \times (-0.820) = -0.715$
$-0.6 + 0.14 \times (-0.487) = -0.668$

Numerical example

Predictions and squares

$\hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
-0.688	0.430
-0.574	0.097
-0.717	39.778
-0.377	5.290
-0.554	4.381
-0.751	6.250
-0.668	27.826

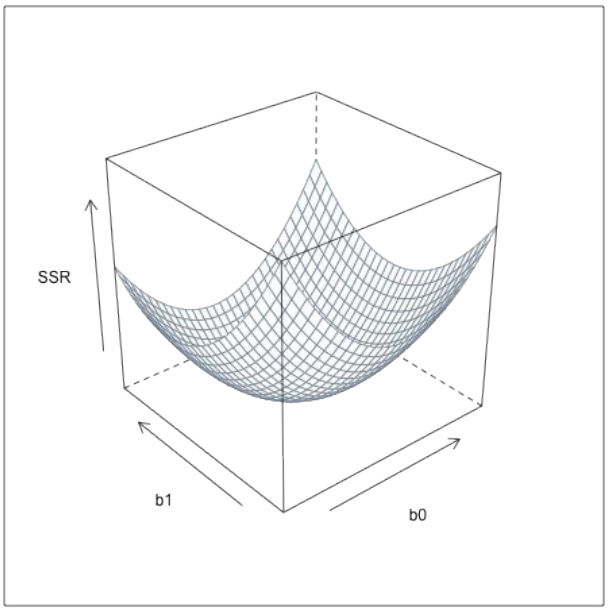
Sum of squares

The sum of the squares for the line with $b_0 = -0.6$ and $b_1 = 0.14$ is

$$\begin{aligned} SSR =& 0.430 + 0.097 + 39.778 \\ &+ 5.290 + 4.381 + 6.25 + 29.826 \\ =& 84.05 \end{aligned}$$

Solution to minimization problem

- One option is to try all combinations of values for b_0 and b_1, compute SSR and choose that combination generating the smallest SSR.
- Graphing the SSRs of all these combinations gives



Solution to minimization problem for general function

- Finding the optimal values this way is cumbersome. We can do better (and faster).
- The general problem is to minimize a function $f(x)$.
- You can think of x as a single variable. Similar arguments apply when $f(\cdot)$ is a function of two or more variables (as in our case).
- If the function is strictly convex, the necessary and sufficient condition for x_0 to be a minimizer of $f(x)$ is

$$\frac{d}{dx} f(x_0) = 0$$

  - The minimum value of a function $f(x)$ is denoted

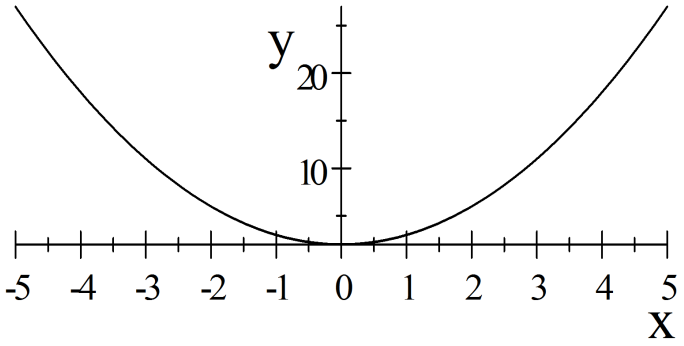
$$\min_x f(x)$$

  - The argument for which the function achieves its minimum is denoted

$$x_0 = \arg\min_x f(x)$$

Example

- $$f(x) = 2 + x^2 \qquad \text{and} \qquad \frac{d}{dx} f(x) = 2x$$
- Equating derivative to zero gives $x^0 = 0$, so that
$$\min_x f(x) = 2 \text{ and } \arg\min_x f(x) = 0$$



Back to regression model

- The function f is the SSR function which is a function of 2 parameters b_0 and b_1,
$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$
- We find the minimizers as in example
$$\frac{\partial \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{\partial b_0} = 0 \Longrightarrow -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = 0$$
$$\frac{\partial \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{\partial b_1} = 0 \Longrightarrow -2 \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i)^2 = 0$$
- These 2 equations are the **first order conditions (FOC)** for the minimum.

Solving the FOC

- The FOC are:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \quad (1)$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0 \quad (2)$$

Two equations in two unknowns (b_0 and b_1).

- Solve for b_0 in (1), to obtain

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Substitute $\hat{\beta}_0$ into (2) and solve for b_1 to obtain

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

Solution

- We can rewrite previous solutions as:

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

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

where $s_{xy} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}$ and $s_x^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ (note that we divide both expressions by the same number $n-1$ so that it cancells out).

- The underlying assumption for the existence of a solution is that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \neq 0$$

(*When does this happen?*)

- If the FOCs can be solved, the FOCs are sufficient for the minimum (why?).

Key concepts

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).

Algebraic properties of OLS

- Note that the FOCs can be written in terms of the OLS residuals as

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

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

- Because OLS satisfies these two equations we will always find, in any sample, that:
  - **The OLS residuals always add up to (and average) zero.**
  - **The sample (not the population) correlation between the residuals and the regressor X is always zero.**
- These properties of OLS hold **by definition**.

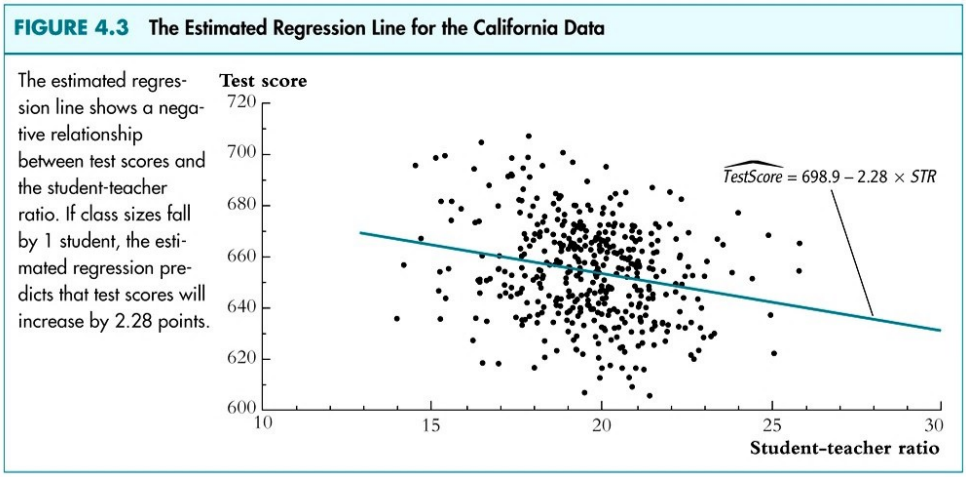
Practical application

- Using data on X and Y we compute $\hat{\beta}_0$ and $\hat{\beta}_1$.This is done using a computer program (Stata in our case but also R, Excel, etc.).
- The output of the regression is often written as:

$$\widehat{testscore} = 698.9329 - 2.2798 \times STR$$

1. Estimated slope: $\hat{\beta}_1 = -2.2798$
2. Estimated intercept: $\hat{\beta}_0 = 698.933$
3. $\widehat{TestScore}$ denotes the estimated regression line.

Graphic representation



Line is estimated regression line $698.9 - 2.28 \times STR$.

Interpretation of estimated parameters

- The estimated slope of -2.28 means than an **increase** in the STR by one student is, on average, **associated with a decrease of 2.28** points on the test.
- If the STR were to **decrease** by 2 students this will, on average, be **associated with an increase of 4.56 points** on the test.
- The intercept (taken literally) means that districts with zero students per teacher would have a (predicted) test score of 698.9....which is non-sense.
  - Extrapolation of the estimated line outside the range of the data.
  - Here, and in most cases, the intercept is not economically meaningful.

How to do it with Stata

- The command is reg (or regress): reg Y X.
- Column "Coef." shows $\hat{\beta}_1$ and $\hat{\beta}_0$.

```
. reg testsc str
```

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

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-2.279808	.4798256	-4.75	0.000	-3.22298	-1.336637
_cons	698.933	9.467491	73.82	0.000	680.3231	717.5428

Where are we?

1. Linear regression model with single regressor (SW 4.1)
2. Estimation of parameters in linear regression model (SW 4.2)
3. **Measures of fit (SW 4.3)**
4. Properties of OLS
    1. The Least Squares assumptions (SW 4.4)
    2. Sampling distribution of OLS (SW 4.5)

Measures of fit (SW 4.3)

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

1. The regression R^2.
2. The standard error of the regression (SER).

Measures of fit: R squared

- R^2 measures the fraction of the variance of Y that is "explained" or "accounted for" by X.
- It is based on the following decomposition

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

- The Explained Sum of Squares (ESS) is the sum of the squared deviations of the predicted values of Y_i from its average,

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

- The Total Sum of Squares (TSS) is the squared deviations of Y_i from its average,

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

Sum of squares in Stata

- ESS and TSS appear in Stata's regression output as SS (sum of squares) of the "Model" (7794.11004) and "Total" (152109.594).

```
. reg testsc str
```

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

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-2.279808	.4798256	-4.75	0.000	-3.22298	-1.336637
_cons	698.933	9.467491	73.82	0.000	680.3231	717.5428

Measures of fit: R squared

- It can be shown that

$$\underbrace{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}_{TSS}=\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}_{ESS}+\underbrace{\sum_{i=1}^{n}\hat{u}_i^2}_{SSR}$$

- In Stata's output, SSR in under the "Residual" SS (144315.484).
- R^2 is defined by

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

  - $0\leq R^2\leq 1$.
  - $R^2=0$ means $ESS=0$, no fit.
  - $R^2=1$ means $ESS=TSS$, perfect fit.
- Check in Stata's output that $R^2=\frac{7794.11004}{152109.594}=0.051\,24$.
- For regression with a single regressor (the case here), R^2 is the square of the correlation coefficient between X and Y.

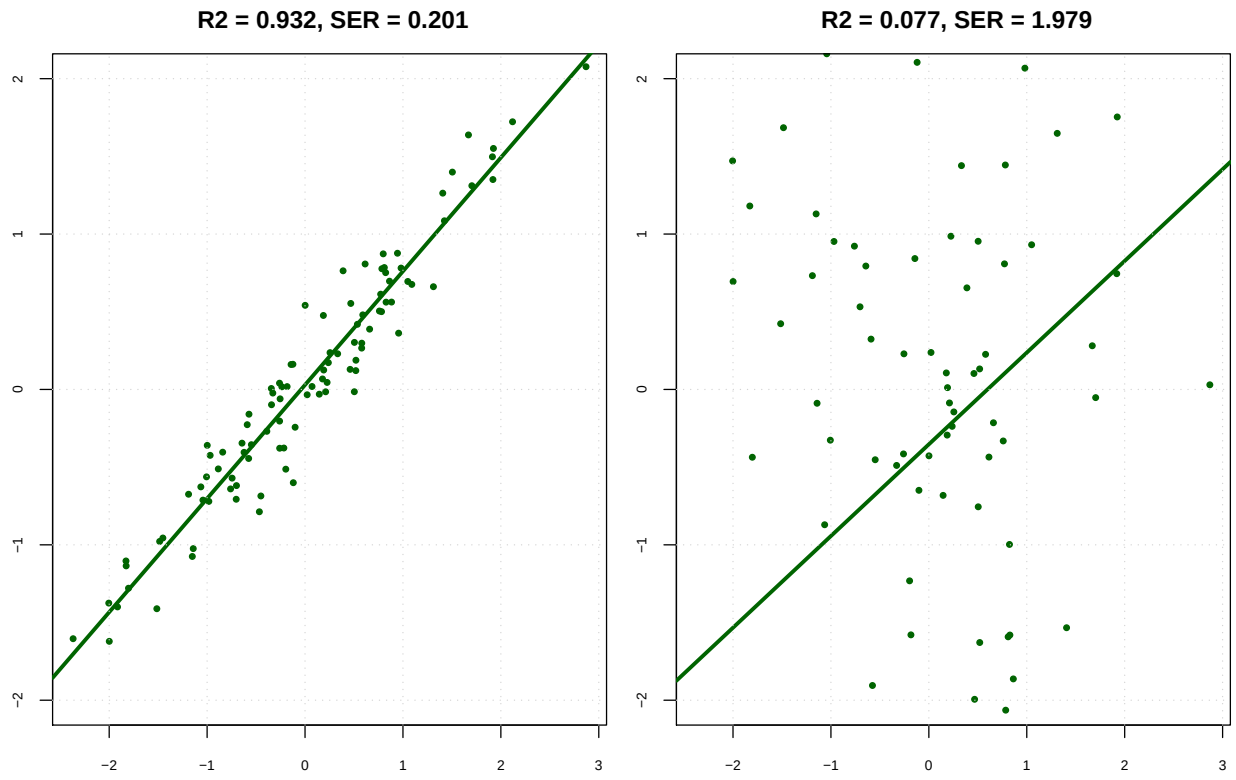
Measures of fit: SER (RMSE)

- The Standard Error of the Regression (SER) is (almost) the sample standard deviation of the OLS residuals:

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

  (the second equality holds because $\bar{\hat{u}}=0$).
- SER has the units of u, which are the units of Y.
- SER measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line).
- Stata calls SER the root mean squared error (RMSE). Check in Stata's output that $RMSE=\sqrt{\frac{144315.484}{420-2}}=18.581$.
  - Some authors (e.g., SW) divide by n instead of $n-2$ but this does not really matter when n is large.

Graphical illustration



Where are we?

1. Linear regression model with single regressor (SW 4.1)
2. Estimation of parameters in linear regression model (SW 4.2)
3. Measures of fit (SW 4.3)
4. **Properties of OLS**
   1. **The Least Squares assumptions (SW 4.4)**
   2. Sampling distribution of OLS (SW 4.5)

Properties of OLS

- The OLS regression line is an estimate computed using a sample of data; a different sample would give a different value of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables because they are functions of r.v.'s (recall that X_i and Y_i are r.v.'s).
- As such, they have a distribution which quantifies the sampling uncertainly associated with the estimator.
- We are interested in knowing the **sampling distribution of the OLS estimator**; at the very least, its **expected value and variance**.

Properties of OLS

- In particular, we want to know whether $E\left(\hat{\beta}_1\right) = \beta_1$? That is, is the OLS estimator unbiased for β_1?
- We also want to know whether the variance of $\hat{\beta}_1$ is small (in relative terms).
- To do this we need some assumptions about the way Y and X are related to each other and about how the data were sampled.
- We first discuss these assumptions which are known as the **Least Squares Assumptions**, and then move on to derive the sampling distribution of OLS.
- Before doing this, next slide clarifies the probability framework for linear regression.

Probability framework for linear regression

- **Population:** the group of observations of interest (e.g., all possible school districts).
- **Random variables:** variables Y and X relevant to the analysis of interest (e.g., test scores, STR).
- These random variables are characterized by a **joint distribution** which is unknown**.** An object of interest in this joint distribution is the conditional expectation of Y given X, $E(Y|X)$.
- **Data collection and simple random sampling:** Choose n units (entities) at random from the population of interest, and observe (record) X and Y for each unit.
  - Simple random sampling implies that $\{(X_i, Y_i)\}, i = 1, \ldots, n$, are independently and identically distributed (i.i.d.).
  - That is, (X_i, Y_i) are distributed independently of (X_j, Y_j) for different observations i and j.

Least Squares Assumptions (SW Section 4.4)

- The linear regression model with a single regressor is

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

- The three **least squares assumptions** are:

Assumption #1 The conditional distribution of u given X has mean zero. That is, for each x

$$E(u|X = x) = 0$$

Assumption #2 (Y_i, X_i) are i.i.d.

Assumption #3 Large outliers in Y and X are unlikely.

Least Square Assumption #1: mean-independence

- LSA #1:
$$E(u|X = x) = 0$$
means that the average effect of all other factors that make up u does not vary with X.
- The assumption states that this conditional expected value is **zero** but this is just a normalization. What matters is that is **constant – it does not vary with** X.
- This is the meaning of **mean-independence.**
- LSA #1 implies, for example, that $E(FamilyIncome|STR) =$ constant, when income is part of u in the testscore-STR regression model. Is this likely to be correct?

Least Squares Assumption #1: mean-independence and correlation

- Mean-independence implies zero covariance (and correlation), but the converse is not true:
$$\begin{aligned} E(u|X) &= 0 \Rightarrow Cov(X, u) = 0 \\ Cov(X, u) &= 0 \not\Rightarrow E(u|X) = 0 \end{aligned}$$
- This means that if X and u are correlated then u cannot be mean-independent of X.
- It is therefore convenient to discuss the mean-independence assumption in terms of possible correlations between X and u.
- We will often see through the course that lack of correlation between X and u is a very strong and questionable assumption.
  - For example, isn't it likely that family income and STR are negatively correlated?

Randomized experiments and mean-independence

- A benchmark for thinking about the validity of 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).
- The **key point** is that because X is assigned randomly, all other individual characteristics – the things that make up u – are uncorrelated with X.
- Thus, in an ideal randomized controlled experiment, $E(u|X = x) = 0$ holds by design of the experiment.
- In actual experiments, or with observational data on X and Y, we will need to think hard about whether we can view X as **if** randomly assigned so that the assumption $E(u|X = x) = 0$ holds.

Implications of Mean-independence assumption (LSA #1)

- If we just write

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

  without assuming anything about $E(u|X)$ then β_1 is not necessarily the causal effect of X on Y since changes in X may trigger changes in the factors that make up u and therefore the changes in Y reflect both causal and indirect effects of X.
- Assuming $E(u|X) = 0$ means that, on average, u does not change with X and thus the change in Y reflects **only** the change in X.
- Thus, under LSA #1, OLS is estimating the **causal effect of X on Y** justifying our goal of estimating β_1.

Implications of Mean-independence assumption (LSA #1)

- Assuming $E(u|X) = 0$ also implies that the population regression line $\beta_0 + \beta_1 X$ equals the conditional expectation,

$$\begin{aligned} E(Y|X) &= E[(\beta_0 + \beta_1 X)|X] + E(u|X) \\ &= \beta_0 + \beta_1 X \end{aligned}$$

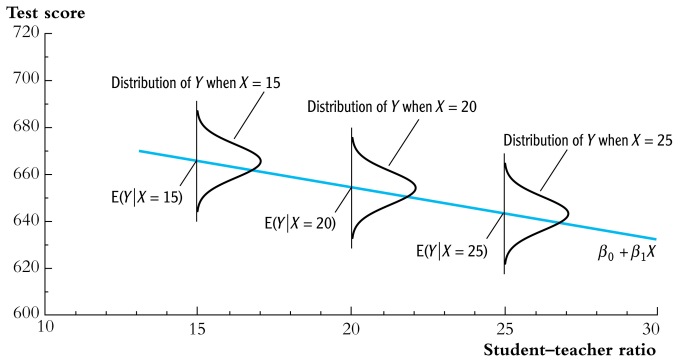
  further justifying our goal of estimating β_1 (and β_0).
- Thus, under LSA#1, β_1 is the causal effect of a unit change in X on the conditional expectation of Y

$$\begin{aligned} E(Y|X &= x+1) - E(Y|X = x) \\ &= (\beta_0 + \beta_1 (x+1)) - (\beta_0 + \beta_1 x) \\ &= \beta_1 \end{aligned}$$

- When X is continuous we can examine the effect of small (infinitesimal) changes,

$$\frac{\partial E(Y|X)}{\partial X} = \beta_1$$

Least Squares Assumption #1: mean-independence or conditional zero-mean



- Picture shows that distribution for u for each x is the same: $E(u|X = x)$ does not change with x.
- Since $E(u|X = x) = 0$, the population line is the C.E. $E(Y|X = x) = \beta_0 + \beta_1 x$.
- What do you think happens if $E(u|X = x)$ is constant but different from zero (say $E(u|X = x) = \alpha_0$ for each x)?

Least Squares Assumption 2: i.i.d.

- (Y_i, X_i) are i.i.d. (identically, independent distributed).
- This arises automatically if the unit (individual, district) is sampled by simple random sampling.
  - Units are selected **randomly** from the **same** population.
  - Thus, their $(X_i, Y_i)'s$ are drawn independently and from the same population.
  - Independence is across observations, not between X and Y!
- The main case when non-i.i.d. sampling occurs is when data are recorded over time ("time series data").

Least Squares Assumption 3: no large outliers

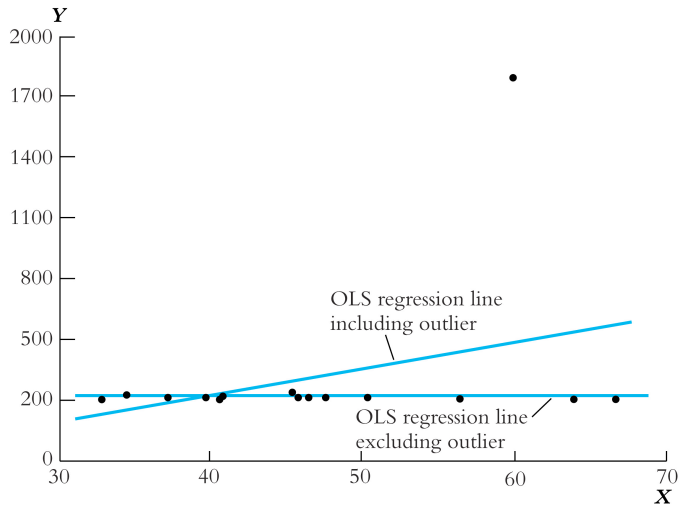
- Large outliers – extreme values – in Y and X are unlikely.
- On a technical level, we assume that they have finite fourth moments (kurtosis),

$$E\left(Y^4\right) < \infty, \qquad E\left(X^4\right) < \infty$$

- Finite kurtosis occurs when a variable is bounded.
- Most economic data are bounded or drawn from distributions with finite fourth moments.
  - Standardized test scores automatically satisfy this,
  - STR, family income, etc. satisfy this too.
- The substance of this assumption is to rule out large outliers that can strongly influence the results.

Least Squares Assumption 3: no large outliers

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 (e.g., plot it)!



Where are we?

1. Linear regression model with single regressor (SW 4.1)
2. Estimation of parameters in linear regression model (SW 4.2)
3. Measures of fit (SW 4.3)
4. **Properties of OLS**
    1. The Least Squares assumptions (SW 4.4)
    2. **Sampling distribution of OLS (SW 4.5)**

Sampling distribution of the OLS estimator (SW 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:
  - Find out $E(\hat{\beta}_1)$: where is the distribution of $\hat{\beta}_1$ centered?
  - Find out $\text{Var}(\hat{\beta}_1)$: quantify the sampling uncertainty associated with $\hat{\beta}_1$.
  - And in general, what is the distribution of $\hat{\beta}_1$ in small samples?
  - And in large samples?

Preliminary algebra 1

Given $Y_i = \beta_0 + \beta_1 X_i + u_i$ and taking means on both sides, noting that

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

we can express the model in mean deviation form

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u}). \tag{3}$$

Substituting (1) into the expression for $\hat{\beta}_1$, we get

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{4}$$

Preliminary algebra 2

We have that

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

Preliminary algebra 3

We can put both results together to write the OLS estimator as

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

Let

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

Then

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n}\omega_i u_i \tag{5}$$

Expected value of the OLS estimator

Taking expectations conditional on all the observed $X_i's$ on both sides of (5) gives

$$\begin{aligned} E\left[\hat{\beta}_1|X_1,\ldots,X_n\right] &= E\left[\beta_1|X_1,\ldots,X_n\right] + E\left[\sum_{i=1}^{n}\omega_i u_i|X_1,\ldots,X_n\right] \\ &= \beta_1 + \sum_{i=1}^{n} E\left[\omega_i u_i|X_1,\ldots,X_n\right] \\ &= \beta_1 + \sum_{i=1}^{n} \omega_i E\left[u_i|X_1,\ldots,X_n\right] \\ & \beta_1 + \sum_{i=1}^{n} \omega_i E\left[u_i|X_i\right] \\ &= \beta_1 \text{ under LSA \# 1} \end{aligned}$$

OLS is unbiased

- Under LSA #1,
$$E\left[\hat{\beta}_1|X_1,\ldots,X_n\right] = \beta_1$$
- And since this expectation is constant for any values of the $X's$ we also have
$$E\left[\hat{\beta}_1\right] = \beta_1$$
The **OLS estimator is unbiased under the Linear Assumption #1**.
- Details in Appendix 4.3

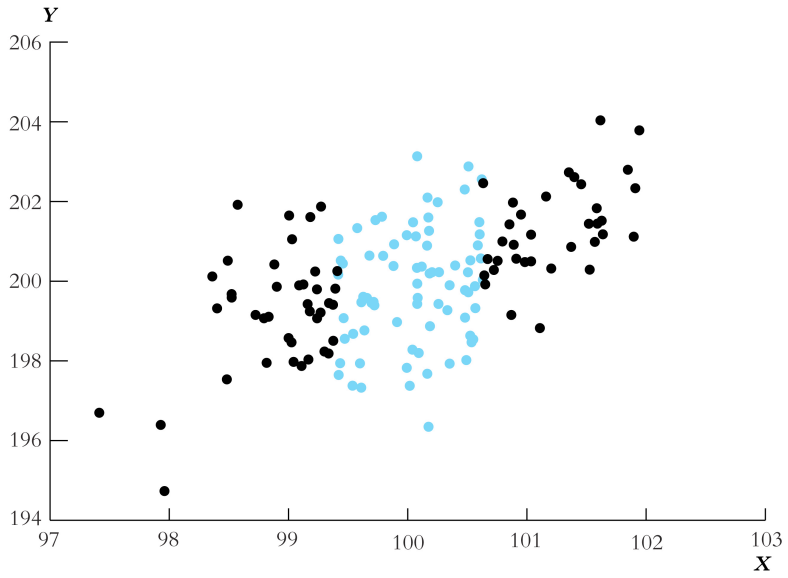
Variance of the OLS estimator

- Start with
$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
- By the LLN in large samples $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is (probabilistically) close to $\sigma_X^2 = Var(X)$ and $(X_i - \bar{X})u_i$ is also close to $(X_i - \mu_X)u_i$, implying
$$\hat{\beta}_1 - \beta_1 \approx \frac{\sum_{i=1}^{n}(X_i - \mu_X)u_i}{n\sigma_X^2}$$
- So that
$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{Var\left(\sum_{i=1}^{n}(X_i - \mu_X)u_i\right)}{n^2\left(\sigma_X^2\right)^2} = \frac{nVar\left((X_i - \mu_X)u_i\right)}{n^2\left(\sigma_X^2\right)^2} \\ &= \frac{1}{n} \times \frac{Var\left((X_i - \mu_X)u_i\right)}{\left(\sigma_X^2\right)^2} \end{aligned}$$

Variance of the OLS estimator

- $$Var(\hat{\beta}_1) = \frac{1}{n} \times \frac{Var\left((X_i - \mu_X)u_i\right)}{\left(\sigma_X^2\right)^2}$$
- The variance of $\hat{\beta}_1$ is inversely proportional to n – just like $Var(\bar{Y})$.
- The larger the variance of X, the smaller the variance of $\hat{\beta}_1$

Larger variance of X, smaller variance of the OLS estimator



If there is more variation in X, then there is more information in the data that you can use to pin down the regression line.

Large sample distribution of the OLS estimator

- 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) approximation:

$$\hat{\beta}_1 \xrightarrow{d} N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$$

where

$$\sigma^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \frac{1}{n} \times \frac{Var((X_i - \mu_X)u_i)}{(\sigma^2_X)^2}$$

- We usually use the **asymptotic** distribution of (standardized) $\hat{\beta}_1$ given by:

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0,1)$$

$N(0,1)$ is a good approximation to the true (unknown) distribution.

Consistency of OLS

- When the Least Squares assumptions hold, we can show that the OLS estimators (probabilistically) approaches the true parameter β_1 as the sample size increases,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1$$

- This implies that **OLS is a consistent for** β_1.

Parallels between asymptotic distribution of OLS and sample mean

$\hat{\beta}_1$

- $E[\hat{\beta}_1] = \beta_1$
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$
- $\hat{\beta}_1 \xrightarrow{d} N(\beta_1, \sigma^2_{\hat{\beta}_1})$
- $\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{Var((X_i - \mu_X)u_i)}{(\sigma^2_X)^2}$

$\bar{Y}$

- $E[\bar{Y}] = \mu_Y$
- $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y} \xrightarrow{d} N(\mu_Y, \sigma^2_{\bar{Y}})$
- $\sigma^2_{\bar{Y}} = \frac{\sigma^2_Y}{n}$

Summary of the sampling distribution of OLS

If the three LS assumptions hold, then:

1. The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(that is, } \hat{\beta}_1 \text{ is unbiased)}$$

$$Var(\hat{\beta}_1) = \frac{1}{n} \times \frac{Var((X_i - \mu_X)u_i)}{(\sigma_X^2)^2} \propto \frac{1}{n}$$

   1. Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (Y, X).

2. $\hat{\beta}_1$ is a consistent estimator of β_1 :

$$\hat{\beta}_1 \xrightarrow{p} \beta_1$$

3. When n is large, approximately (or asymptotically)

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_1)}} \sim N(0, 1)$$

Key concept

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

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

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

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