Ch02. The Simple Regression Model

Ping Yu

HKU Business School
The University of Hong Kong

Definition of the Simple Regression Model

Definition of the Simple Regression Model

- The simple linear regression (SLR) model is also called two-variable linear regression model or bivariate linear regression model.
- The SLR model is usually written as

$$y = \beta_0 + \beta_1 x + u,^1$$

where β_0 is called the intercept (parameter) or the constant term, and β_1 is called the slope (parameter).
- As shown in the next slide, β_1 is of main interest in econometrics.

y	x	u
Dependent variable	Independent Variable	Error Term
Explained variable	Explanatory variable	Disturbance
Response variable	Control variable	Unobservable
Predicted variable	Predictor variable	
Regressand	Regressor	
	Covariate	

Table: Terminology for SLR

[1] How to understand this equation? It means that for each individual in the population, if given the values of x (education) and u (ability), then the value of y (wage) is determined.

Interpretation of the SLR Model

- The SLR model tries to "explain variable y in terms of variable x" or "study how y varies with changes in x":
$$\frac{dy}{dx}(=\beta_1+\frac{\partial u}{\partial x}).$$
- The total derivative $\frac{dy}{dx}$ is the total effect of increasing x by one unit on y:
$$\begin{aligned}\frac{dy}{dx} &= \text{direct effect}+\text{indirect effect}\\ &= \frac{\partial y}{\partial x}+\frac{\partial y}{\partial u}\frac{\partial u}{\partial x}=\frac{\partial(\beta_0+\beta_1x+u)}{\partial x}+\frac{\partial(\beta_0+\beta_1x+u)}{\partial u}\frac{\partial u}{\partial x}=\beta_1+\frac{\partial u}{\partial x},\end{aligned}$$
where $\frac{\partial y}{\partial x}$ means "by how much does y change if **only** x is increased by one unit?".
- ∂ in partial derivative is the counterpart of d in derivative (e.g., $\frac{dy}{dx}$), where d is the first letter of "delta" (Δ) which usually means a small change in mathematics.
- From the definition of causal effect, the causal effect of x on y is $\frac{\partial y}{\partial x}=\beta_1$, which is equal to $\frac{dy}{dx}$ only if $\frac{\partial u}{\partial x}=0$.
- If $\frac{\partial u}{\partial x}\neq 0$, then $\frac{dy}{dx}$ includes the "indirect" effect of x on y, $\frac{\partial u}{\partial x}$.
- In summary, $\frac{dy}{dx}$ is the causal effect of x on y only if $\frac{\partial u}{\partial x}=0$.
- The simple linear regression model is rarely applicable in practice but its discussion is useful for pedagogical reasons.

Two SLR Examples

- Example (Soybean Yield and Fertilizer):

$$yield = \beta_0 + \beta_1 fertilizer + u,$$

where β_1 measures the effect of fertilizer on yield, holding all other factors fixed, and u contains factors such as rainfall, land quality, presence of parasites,$\cdots$

- Example (A Simple Wage Equation):

$$wage = \beta_0 + \beta_1 educ + u,$$

where β_1 measures the change in hourly wage given another year of education, holding all other factors fixed, and u contains factors such as labor force experience, innate ability, tenure with current employer,[2] work ethic,$\cdots$

[2] What is the difference between tenure and experience?

(*) When Can We Estimate a Causal Effect?

- Although when $\frac{\partial u}{\partial x} = 0$, $\frac{dy}{dx}$ has a causal interpretation for **each individual**, it hardly holds in practice.
  - Technically, because we usually can observe only **one** pair of (x, y) for each individual, we cannot identify the individual causal effect which requires y values for at least two x values.
- So, we are usually interested in the **average** causal effect, which in our setup is also β_1 (because the individual causal effect is β_1).
- We can estimate β_1 under the (conditional) mean independence assumption:

$$E[u|x] = 0.^3$$

  - The explanatory variable must not contain information about the mean of the unobserved factors.
  - Intuition: for average causal effects, we need only that the average of u (rather than u itself) does not change as x changes.
- $E[u|x] = 0$ implies $Cov(x, u) = 0$ [proof not required, $Cov(x, u)$ will be defined later]. So in practice, we just argue why x and u are correlated to invalidate a causal interpretation of our estimator.
- Three Questions: (i) How is $E[u|x]$ defined? (ii) Why is $E[u|x] = c$ $(\neq 0)$ enough? (iii) Why can the random assignments in Chapter 1 generate causal effects?

[3] It is called the zero conditional mean assumption in the textbook.

[Review] Mean

- For a random variable (r.v.) X, the mean (or expectation) of X, denoted as $E[X]$ (or $E(X)$) and sometimes μ_X (or simply μ), is a weighted average of all possible values of X.
- For example, in the population, proportion p (e.g., 17% in US) individuals are college graduates, and the remaining are not. Define $X = 1(\text{college graduate})$, where $1(\cdot)$ is the indicator function which equals 1 when the statement in the parenthesis is true and zero otherwise.
- The distribution of X is

$$X = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1-p, \end{cases}$$

so

$$E[X] = 0 \times (1-p) + 1 \times p = p.$$

- For a general discrete r.v. X, $P(X = x_j) = p_j, j = 1, \cdots, J$,[4] where $p_j \geq 0$, and

$$p_1 + \cdots + p_J = \sum_{j=1}^{J} p_j = 1,$$

we have

$$E[X] = \sum_{j=1}^{J} x_j p_j.$$

[4] This is called the probability mass function (pmf) of X.

continue

- The mean of a continuous r.v. can be defined as an approximation of a discrete r.v..
- For a continuous r.v. taking values on (a,b), where a can be $-\infty$ and b can be ∞, [figure here]

$$E[X] \approx \sum_{i=0}^{(b-a)/\Delta-1} \left(a+\left(i+\tfrac{1}{2}\right)\Delta\right) P\left(a+i\Delta < X \le a+(i+1)\Delta\right)^{5}$$
$$\approx \sum_{i=0}^{(b-a)/\Delta-1} \left(a+\left(i+\tfrac{1}{2}\right)\Delta\right) f\left(a+\left(i+\tfrac{1}{2}\right)\Delta\right)\Delta \overset{\Delta\to 0}{\longrightarrow} \int_a^b x f(x)dx.$$

- $\sum \to \int$, $a+\left(i+\frac{1}{2}\right)\Delta \to x$, and $\Delta \to dx$
- For example, if $X \sim N\left(\mu,\sigma^2\right)$, the normal distribution with mean μ and variance σ^2, then $E[X]=\mu$.
- Two Useful Properties: (i) "the mean of the sum is the sum of the mean",

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i];\ (X_i \text{ need not be independent})$$

(ii) for any constants a and b, $E[a+bX] = a+bE[X]$.

[5] $i=0$, $a+i\Delta=a$, and $i=\frac{b-a}{\Delta}-1$, $a+(i+1)\Delta = a+\frac{b-a}{\Delta}\Delta = b$.

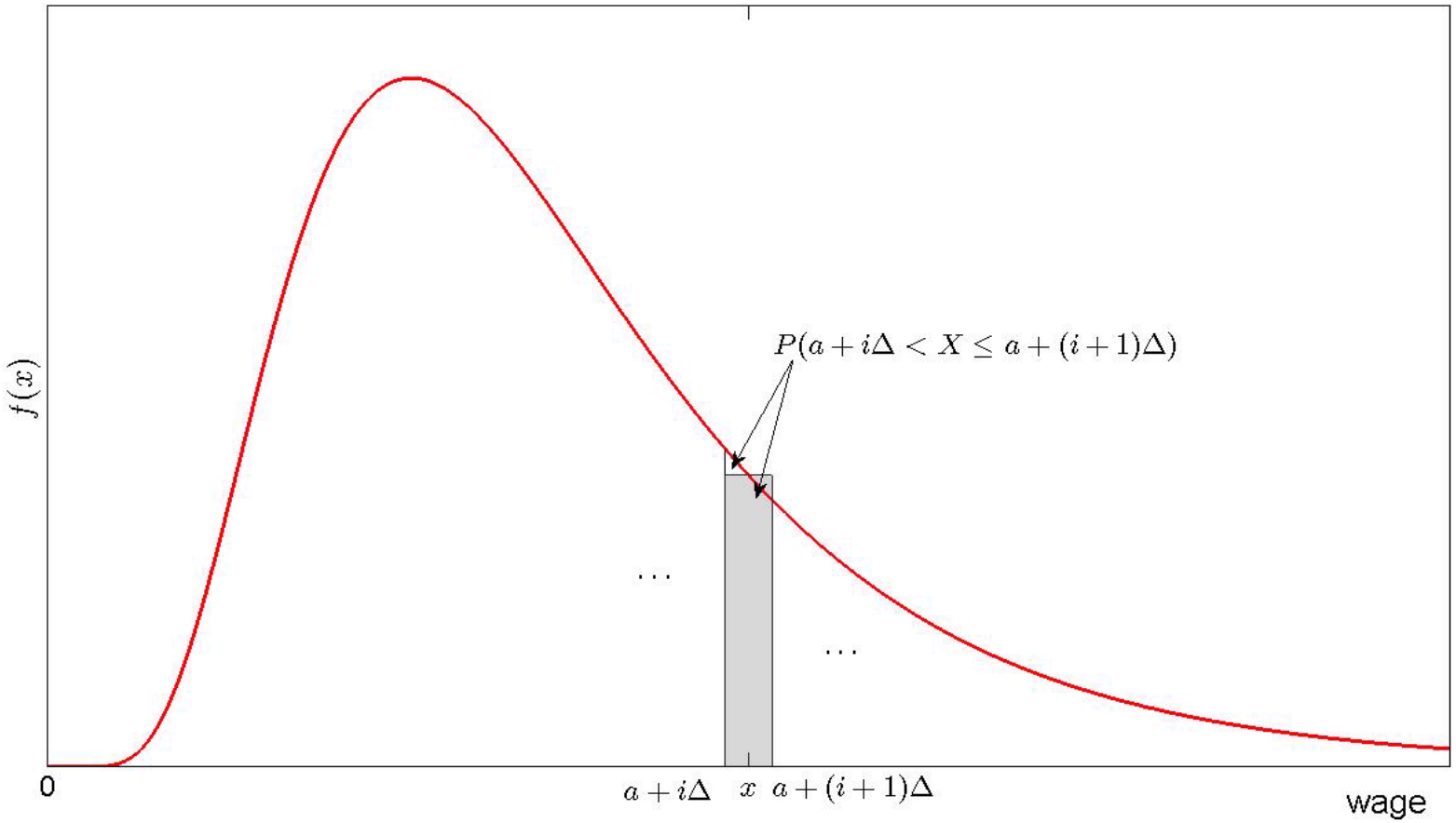


Figure: Probability Density Function (pdf) of Wage: wage $\sim \exp\left(N\left(\mu,\sigma^2\right)\right)$, $a=0, b=\infty$

[Review] Conditional Mean

- For two r.v.'s, Y and X, the conditional mean of Y given $X=x$, denoted as $E[Y|X=x]$ (or $E(Y|X=x)$), is the mean of Y for the (slice of) individuals with $X=x$.
- For example, if Y is the hourly wage, $X=1(\text{college graduate})$, then

$$E[Y|X=1]=\int yf(y|X=1)dy$$

  is the average wage for college graduates, where $f(y|X=1)$ is the density of wage among college graduates.
- The conditional mean $E[Y|X=x]$ can be **any** function of x.

continue

- One Useful Property: $E[g(X)Y|X=x]=g(x)E[Y|X=x]$ for any function $g(\cdot)$, i.e., conditioning on X means X can be treated as a constant.
  - $g(x)$ is similar to b in the second property of mean.
- The two properties of mean can still apply to conditional mean:
  - (i) "the conditional mean of the sum is the sum of the conditional mean",

$$E\left[\sum_{i=1}^{n} Y_i \middle| X=x\right]=\sum_{i=1}^{n} E[Y_i|X=x];$$

  - (ii) for any constants a and b, $E[a+bY|X=x]=a+bE[Y|X=x]$.

Conditional Mean Independence

- Although $E[u|x]$ can be any function of x, conditional mean independence restricts it to be the constant zero.
- $E[u|x]=0$ means for whatever value x takes, the mean of u given the specific x value is zero.
- The zero in $E[u|x]=0$ is just a normalization. If $E[u|x]=c\neq 0$, then redefine $u^*=u-c$, and $\beta_0^*=\beta_0+c$. Now,

$$y=\beta_0+\beta_1 x+u=(\beta_0+c)+\beta_1 x+(u-c)\equiv\beta_0^*+\beta_1 x+u^*,$$

  where $E[u^*|x]=E[u-c|x]=E[u|x]-c=c-c=0$, and $\equiv$ means "defined as".
- So the **key** here is that $E[u|x]$ is a constant, not depending on x.
  - The constant must be $E[u]$ because if the mean of u for each slice of x is c, the mean of u for the entire population must be c.

$E[u|x]=c$ and Causal Interpretation: Return to Schooling

- Recall the wage equation
$$wage=\beta_0+\beta_1 educ+u.$$
- Suppose, for simplicity, $educ=1$(college graduate), and u represents the innate ability. Then, $\frac{dwage}{deduc}$ should be replaced by $E[wage|educ=1]-E[wage|educ=0]$.
- If
$$E[u|educ=1]=c=E[u|educ=0],$$
then
$$\begin{aligned} & E[wage|educ=1]-E[wage|educ=0] \text{ (we can estimate)} \\ = \; & E[\beta_0+\beta_1 educ+u|educ=1]-E[\beta_0+\beta_1 educ+u|educ=0] \\ = \; & (\beta_0+\beta_1)+E[u|educ=1]-\beta_0-E[u|educ=0] \\ = \; & (\beta_0+\beta_1)-\beta_0=\beta_1 \text{ (the target causal effect).} \end{aligned}$$
- Although u is not equal for each individual, averagely, its mean within each group of education level is equal. This is what "all other relevant factors are balanced" in random assignment of x of Chapter 1 means. [If $educ$ is independent of u, then $E[u|educ]=E[u]=c$ for whatever value of $educ$.]
- The conditional mean independence assumption is unlikely to hold here because individuals with more education will also be more intelligent on average.

Causality and Correlation: Polio and Ice-cream

- "By 1910, frequent epidemics became regular events throughout the developed world, primarily in cities during the summer months. At its peak in the 1940s and 1950s, polio would paralyze or kill over half a million people worldwide every year."
  - From Wiki

- Folk legends: (i) A pretty woman causes death? (inauspicious or unlucky?)
  (ii) Old age causes death of my children? (bad omen or natural phenomenon?)
- Donald Trump: More tests, more infections.

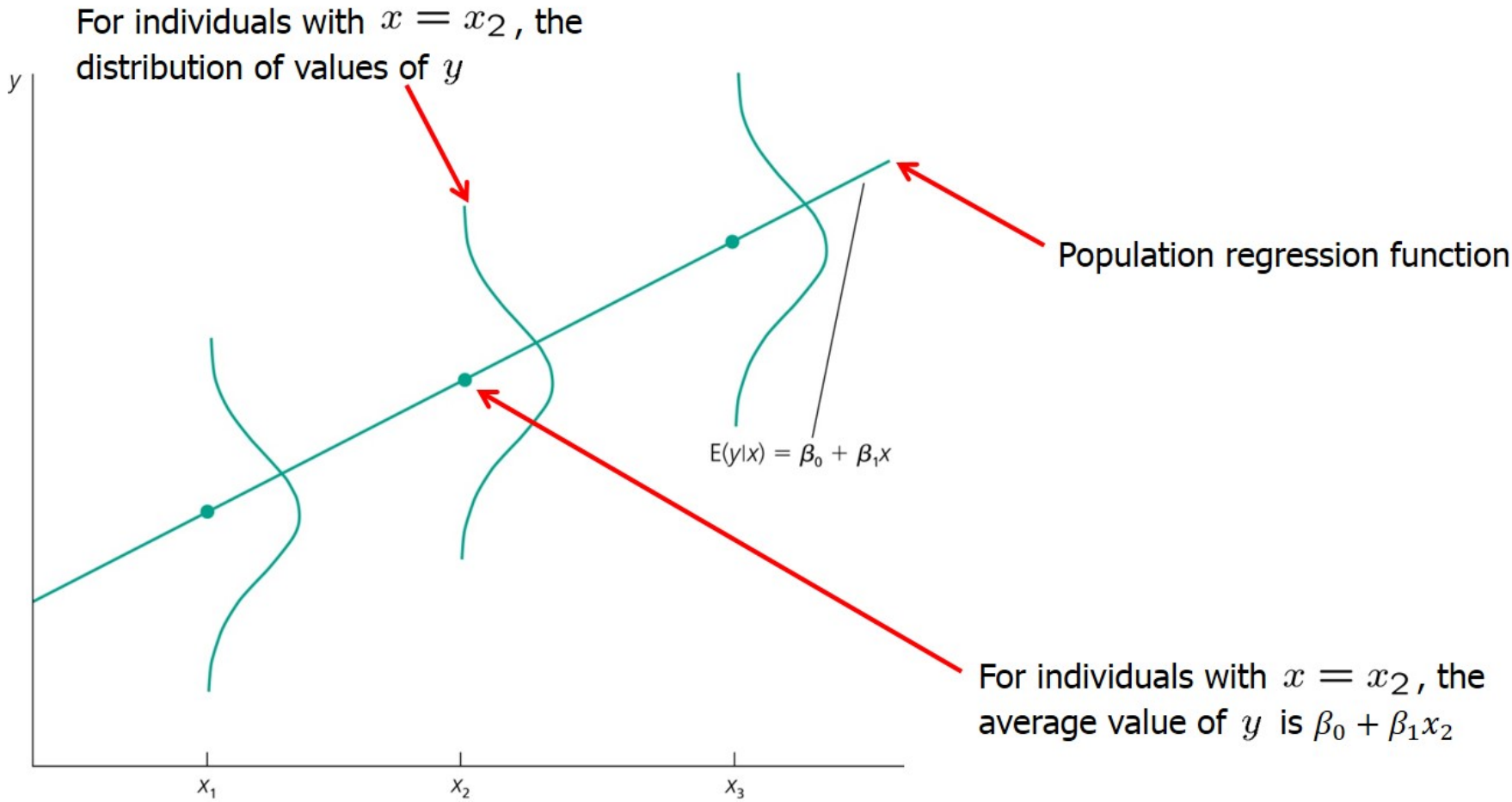
Population Regression Function (PRF)

- Similar as in the return-to-schooling example, the conditional mean independence assumption implies that

$$\begin{aligned} E[y|x] &= E[\beta_0 + \beta_1 x + u|x] \\ &= \beta_0 + \beta_1 x + E[u|x] \\ &= \beta_0 + \beta_1 x. \end{aligned}$$

- This means that the average value of the dependent variable can be expressed as a linear function of the explanatory variable although in general $E[y|x]$ can be any function of x.
- The PRF is **unknown**. It is a theoretical relationship assuming a linear model and conditional mean independence.
- We need to **estimate** the PRF.

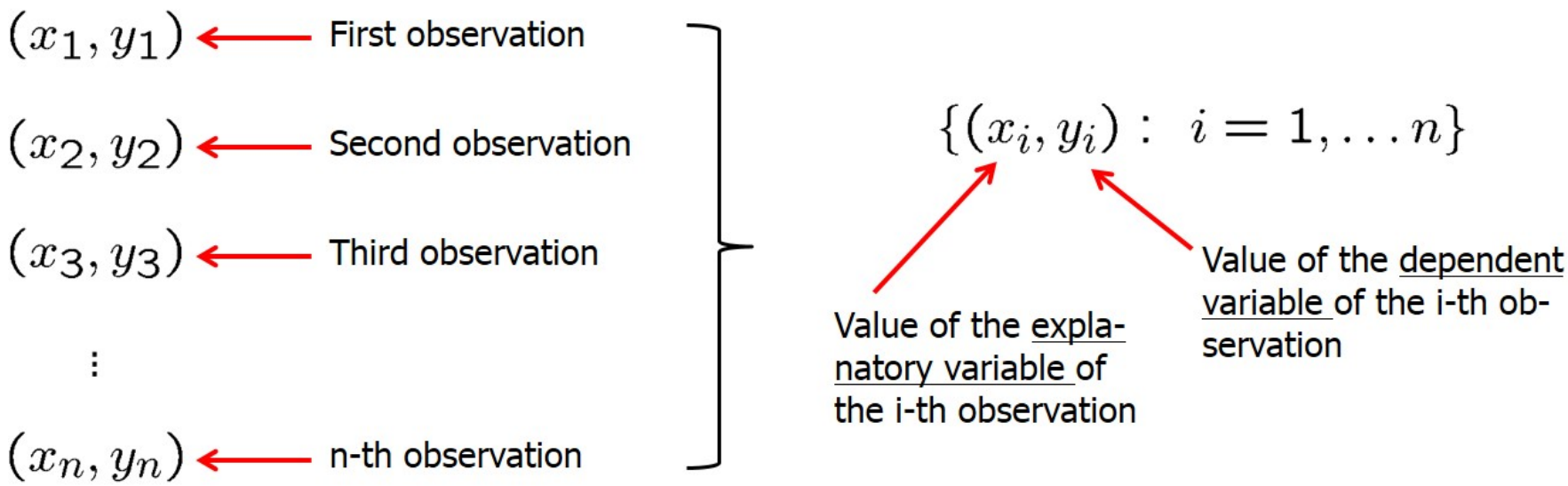
$E[y|x]$ As a Linear Function of x



Deriving the Ordinary Least Squares Estimates

A Random Sample

- In order to estimate the regression model we need data.



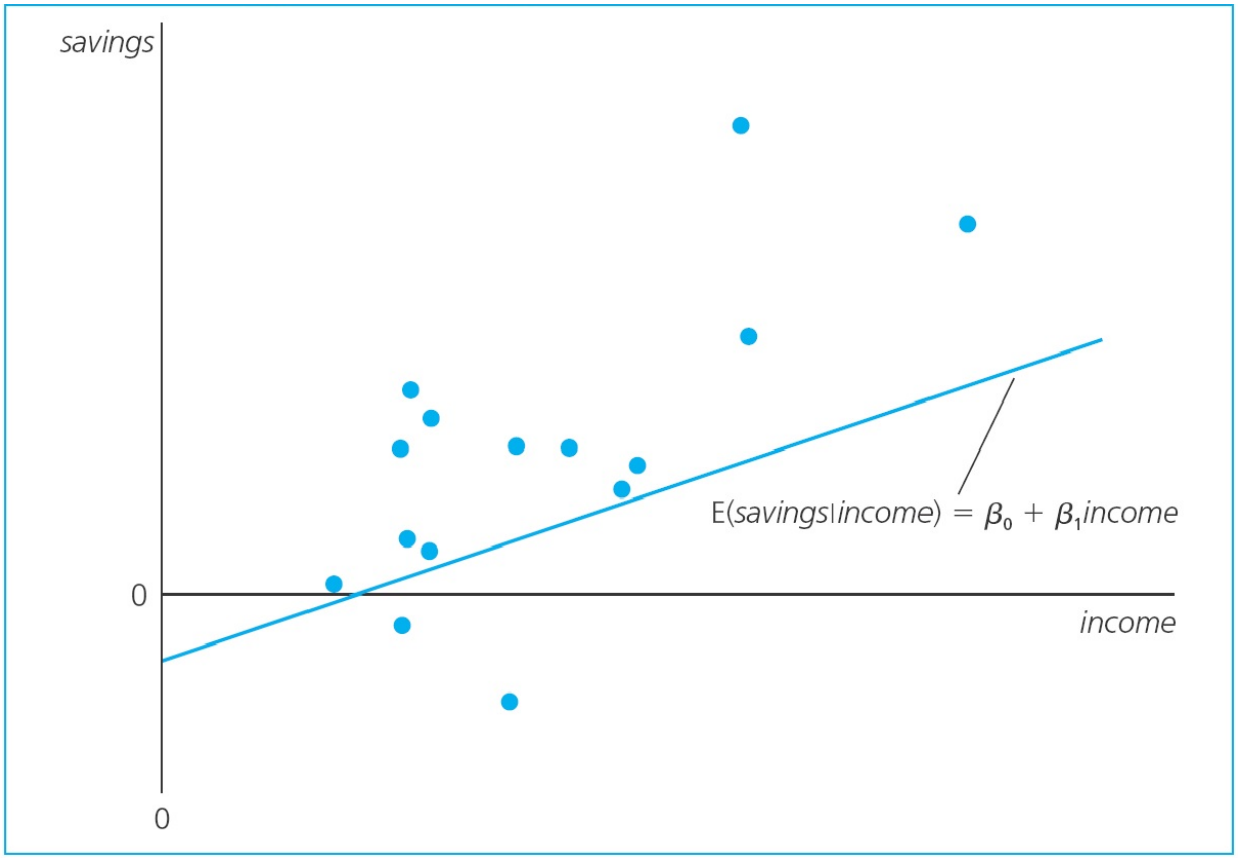


Figure: Scatterplot of Savings and Income for 15 Families, and the Population Regression $E\left[\text{savings}|\text{income}\right] = \beta_0 + \beta_1 \text{income}$

Ordinary Least Squares (OLS) Estimation

- The OLS estimates of $\beta = (\beta_0, \beta_1)$ try to fit as good as possible a regression line through the data points:

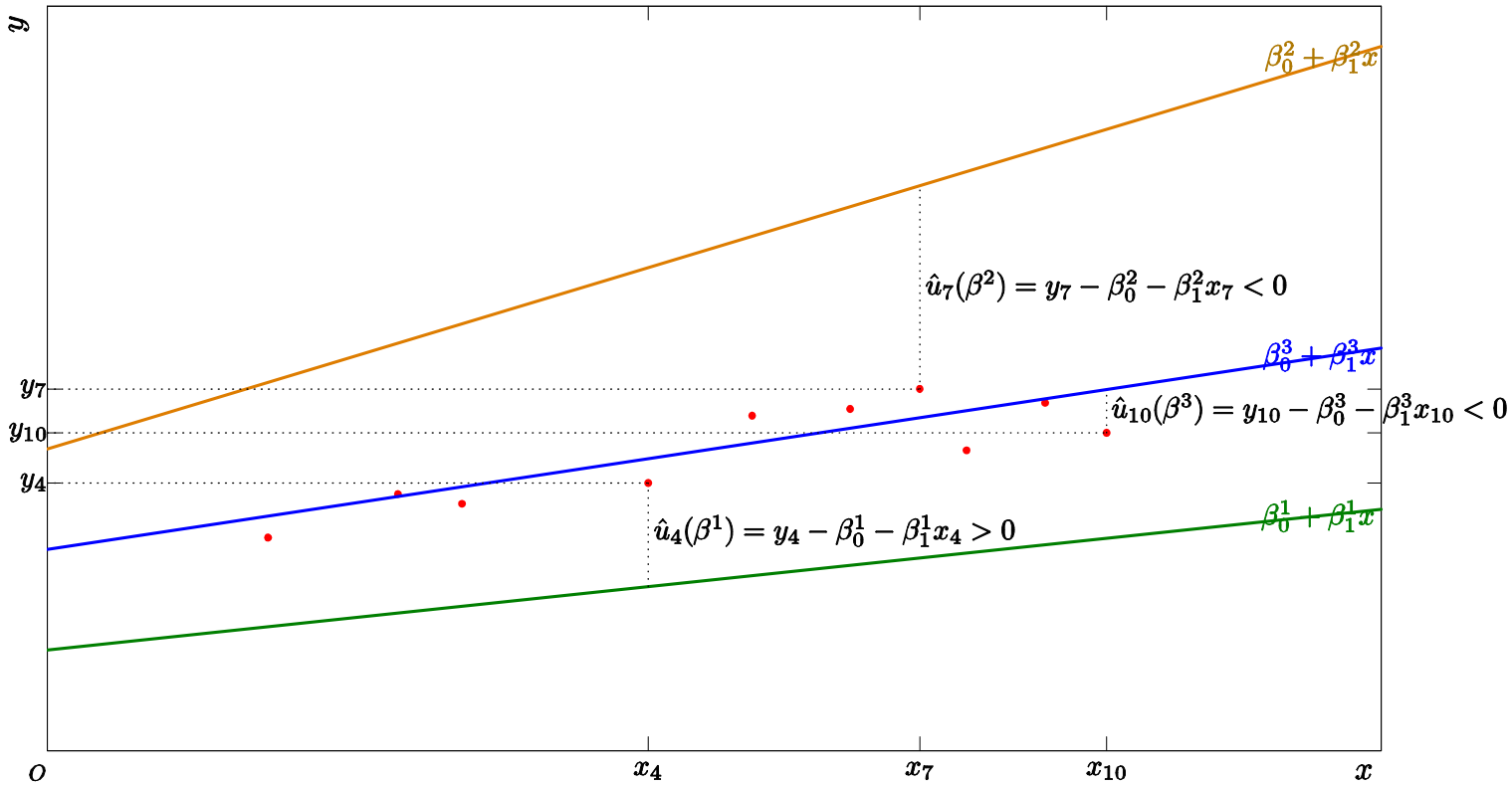


Figure: $\widehat{u}_i(\beta)$ for Three Possible β Values β^1, β^2 and β^3: $n = 10$

What Does "As Good As Possible" Mean?

- Define residuals at arbitrary β as

$$\widehat{u}_i(\beta) = y_i - \beta_0 - \beta_1 x_i.$$

- Minimize the sum of squared residuals [figure here]:

$$\begin{aligned} \min_{\beta_0,\beta_1} SSR(\beta) \quad &\equiv \quad \min_{\beta_0,\beta_1} \sum_{i=1}^{n} \widehat{u}_i(\beta)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\ &\Longrightarrow \quad \widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right), \end{aligned}$$

  where $\widehat{\beta}$ is the solution to the first order conditions (FOCs) for the OLS estimates.

- It turns out that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ and } \widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1,$$

  where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of x, and $\overline{y}$ is similarly defined.

- In modern times, $\widehat{\beta}$ can be easily obtained through standard econometric softwares.

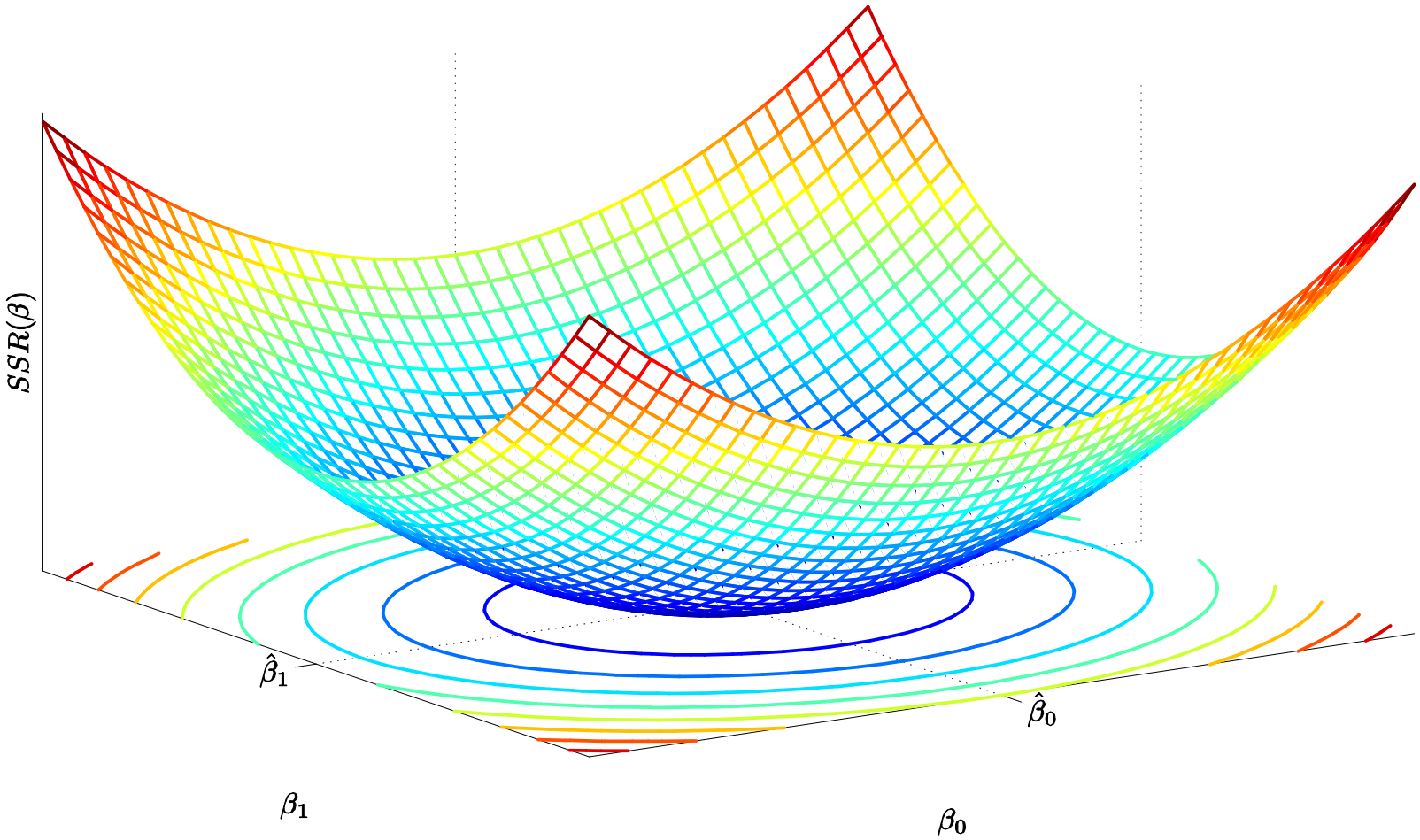


Figure: Objective Functions of OLS Estimation

Derivation of OLS Estimates

- The FOCs are[6]

$$\begin{aligned} -2\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)&=0,\\ -2\sum_{i=1}^{n}x_i\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)&=0, \end{aligned}$$

$\overset{?}{\Longleftrightarrow}$[7]

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)&=0,\\ \frac{1}{n}\sum_{i=1}^{n}x_i\left(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i\right)&=0. \end{aligned}$$

- From the first equation,

$$\overline{y}=\widehat{\beta}_0+\overline{x}\widehat{\beta}_1\Longrightarrow\widehat{\beta}_0=\overline{y}-\overline{x}\widehat{\beta}_1.$$

- Substituting $\widehat{\beta}_0$ into the second equation, we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}x_i\left[y_i-\left(\overline{y}-\overline{x}\widehat{\beta}_1\right)-\widehat{\beta}_1x_i\right]=0\Rightarrow\frac{1}{n}\sum_{i=1}^{n}x_i\left(y_i-\overline{y}\right)-\frac{1}{n}\sum_{i=1}^{n}x_i\left(x_i-\overline{x}\right)\widehat{\beta}_1=0\\ &\Longrightarrow\frac{1}{n}\sum_{i=1}^{n}x_i\left(y_i-\overline{y}\right)=\widehat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}x_i\left(x_i-\overline{x}\right). \end{aligned}$$

[6] Recall that $\frac{dx^2}{dx}=2x$ and $\frac{d(ax+b)}{dx}=a$, so by the chain rule,
$\frac{d\left(y_i-\beta_0-\beta_1x_i\right)^2}{d\beta_0}=2\left(y_i-\beta_0-\beta_1x_i\right)\frac{d\left(y_i-\beta_0-\beta_1x_i\right)}{d\beta_0}=-2\left(y_i-\beta_0-\beta_1x_i\right)$, and
$\frac{d\left(y_i-\beta_0-\beta_1x_i\right)^2}{d\beta_1}=2\left(y_i-\beta_0-\beta_1x_i\right)\frac{d\left(y_i-\beta_0-\beta_1x_i\right)}{d\beta_1}=-2x_i\left(y_i-\beta_0-\beta_1x_i\right)$.

[7] First devide the constant -2 and then multiply the constant $\frac{1}{n}$.

continue

- So

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i (y_i - \overline{y})}{\sum_{i=1}^n x_i (x_i - \overline{x})} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2},$$

where

$$\begin{aligned}\sum_{i=1}^n x_i (y_i - \overline{y}) - \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^n [x_i - (x_i - \overline{x})](y_i - \overline{y}) = \sum_{i=1}^n \overline{x}(y_i - \overline{y}) \\ &= \overline{x} \sum_{i=1}^n (y_i - \overline{y}) \overset{?}{=}^{8} \overline{x}(n\overline{y} - n\overline{y}) = 0,\end{aligned}$$

$$\sum_{i=1}^n x_i (x_i - \overline{x}) - \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n \overline{x}(x_i - \overline{x}) = 0.$$

- Summing the product of two demeaned terms need only demean one of them.

- Alternative Expression for $\widehat{\beta}_1$:

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)}.$$

[8] $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$, so $\sum_{i=1}^n y_i = n\overline{y}$.

[Review] Covariance and Variance

- The population covariance between two r.v.'s X and Y, sometimes denoted as σ_{XY}, is defined as

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

- Intuition: [figure here]
  - If $X > \mu_X$ and $Y > \mu_Y$, then $(X-\mu_X)(Y-\mu_Y) > 0$, which is also true when $X < \mu_X$ and $Y < \mu_Y$. While if $X > \mu_X$ and $Y < \mu_Y$, or vice versa, then $(X-\mu_X)(Y-\mu_Y) < 0$.
  - If $\sigma_{XY} > 0$, then, on average, when X is above/below its mean, Y is also above/below its mean. If $\sigma_{XY} < 0$, then, on average, when X is above/below its mean, Y is below/above its mean.
- A positive covariance indicates that two r.v.'s move in the same direction, while a negative covariance indicates they move in opposite directions.

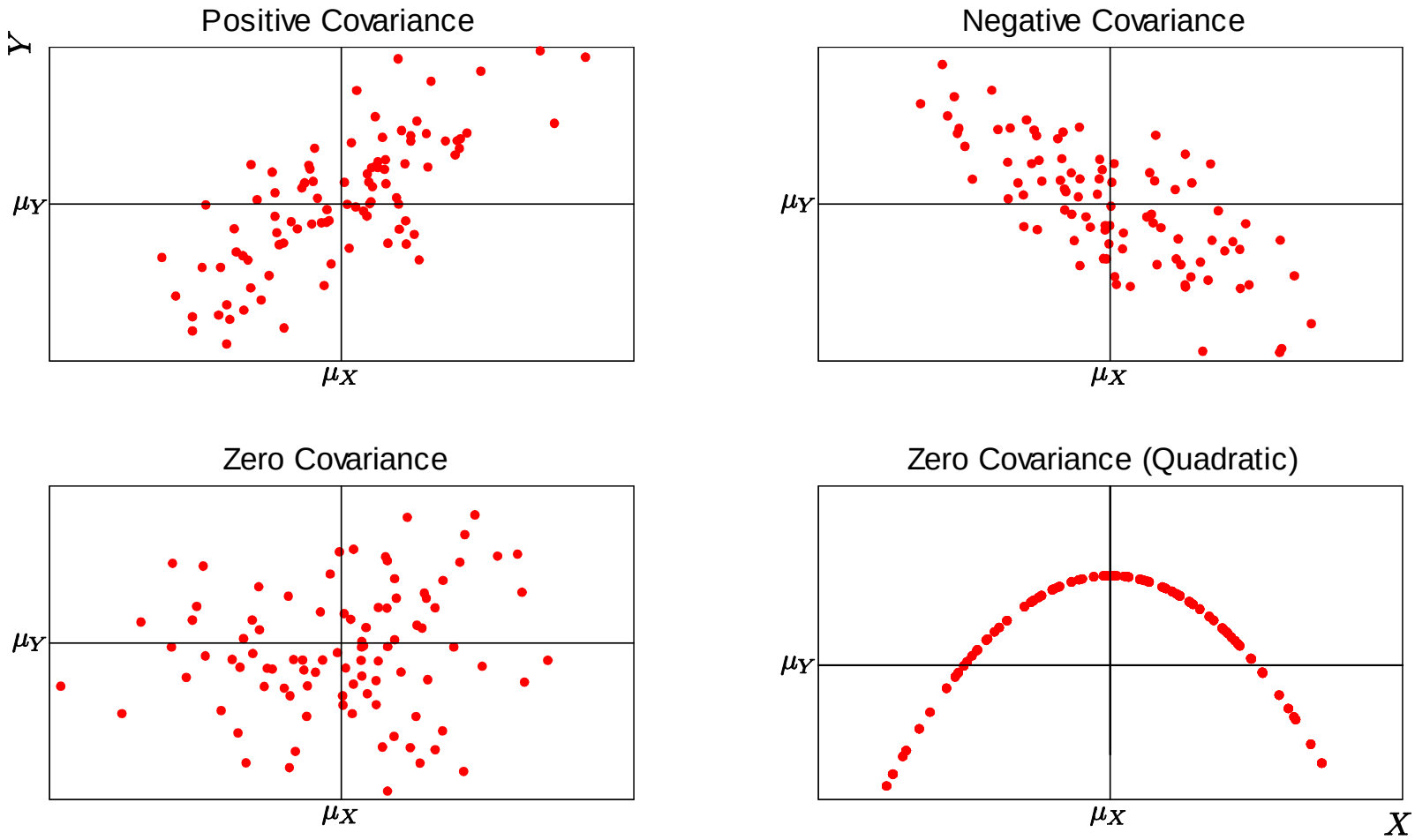


Figure: Positive, Negative an Zero Covariance

continue

- Alternative Expressions of $Cov(X,Y)$:

$$\begin{aligned} Cov(X,Y) &= E[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y] \\ &= E[XY]-\mu_X\mu_Y-\mu_Y\mu_X+\mu_X\mu_Y \\ &= E[XY]-\mu_X\mu_Y \\ &= E[(X-\mu_X)Y]=E[X(Y-\mu_Y)], \end{aligned}$$

  where the last two equalities indicate that demeaning one of X and Y is enough.
- Covariance measures the amount of linear dependence[9] between two r.v.'s.
  - If $E[X]=0$ and $E\left[X^3\right]=0$, then $Cov(X,X^2)=E\left[X^3\right]-E[X]E\left[X^2\right]=0$ although X and X^2 are quadratically related. [Figure here]
- $Var(X)=Cov(X,X)=E\left[(X-\mu_X)^2\right]=E\left[X^2\right]-\mu_X^2$ is the covariance of X with itself, denoted as σ_X^2 or simply σ^2. (we will discuss more on it later)
  - The definition of $Var(X)$ implies $E\left[X^2\right]=Var(X)+E[X]^2$, the second moment is the variance plus the first moment squared.

[9] This is why $\widehat{Cov}(x,y)$ appears in $\widehat{\beta}_1$ which measures the linear relationship between y and x.

[Review] Method of Moments

- The method of moments (MoM) was put forward by Karl Pearson (1857-1936) in 1894. The basic idea is to

$$\text{replace } E[\cdot] \text{ by } \frac{1}{n}\sum_{i=1}^{n}\cdot$$

  So the MoM estimator is often called the sample analog or sample counterpart.
- For example, $E[X]$ can be estimated by the sample mean

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

- $Cov(X,Y)$ can be estimated by the sample covariance

$$\widehat{Cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right).$$

  - Recall that demeaning one of X and Y is enough![10]
- $Var(X)$ can be estimated by the sample variance

$$\widehat{Var}(X)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2.$$

[10] You can use other formulas of $Cov(X,Y)$ to get the same estimator, e.g., $Cov(X,Y)=E[XY]-\mu_X\mu_Y$, so $\widehat{Cov}(X,Y)=\overline{XY}-\overline{X}\cdot\overline{Y}$.

OLS Calculation: A Cooked Numerical Example

	y_i	x_i	$y_i-\overline{y}$	$x_i-\overline{x}$	$(x_i-\overline{x})(y_i-\overline{y})$	$x_i(y_i-\overline{y})$	$(x_i-\overline{x})^2$	$x_i(x_i-\overline{x})$
	4	1	-3	-3	9	-3	9	-3
	5	4	-2	0	0	-8	0	0
	7	5	0	1	0	0	1	5
	12	6	5	2	10	30	4	12
$\sum_{i=1}^4$	28	16	0	0	19	19	14	14
$\frac{1}{4}\sum_{i=1}^4$	7	4	0	0	$4.75>0$	4.75	3.5	3.5

Table: Components of OLS Calculation: $n=4$

- $\widehat{\beta}_1=\frac{\sum_{i=1}^4 x_i(y_i-\overline{y})}{\sum_{i=1}^4 x_i(x_i-\overline{x})}=\frac{\sum_{i=1}^4(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^4(x_i-\overline{x})^2}=\frac{19}{14}=\frac{\frac{1}{4}\sum_{i=1}^4(x_i-\overline{x})(y_i-\overline{y})}{\frac{1}{4}\sum_{i=1}^4(x_i-\overline{x})^2}=\frac{4.75}{3.5}=1.357.$
- $\widehat{\beta}_0=\overline{y}-\overline{x}\widehat{\beta}_1=7-4\times 1.357=1.571.$

History of Ordinary Least Squares

- The least-squares method is usually credited to Gauss (1809), but it was first published as an appendix to Legendre (1805) which is on the paths of comets. Nevertheless, Gauss claimed that he had been using the method since 1795 at the age of 18.

C.F. Gauss (1777-1855), Göttingen

A.-M. Legendre (1752-1833), Éole Normale

CEO Salary and Return on Equity

- We will provide three empirical examples of OLS estimation.
- Suppose the SLR model is

$$salary = \beta_0 + \beta_1 roe + u,$$

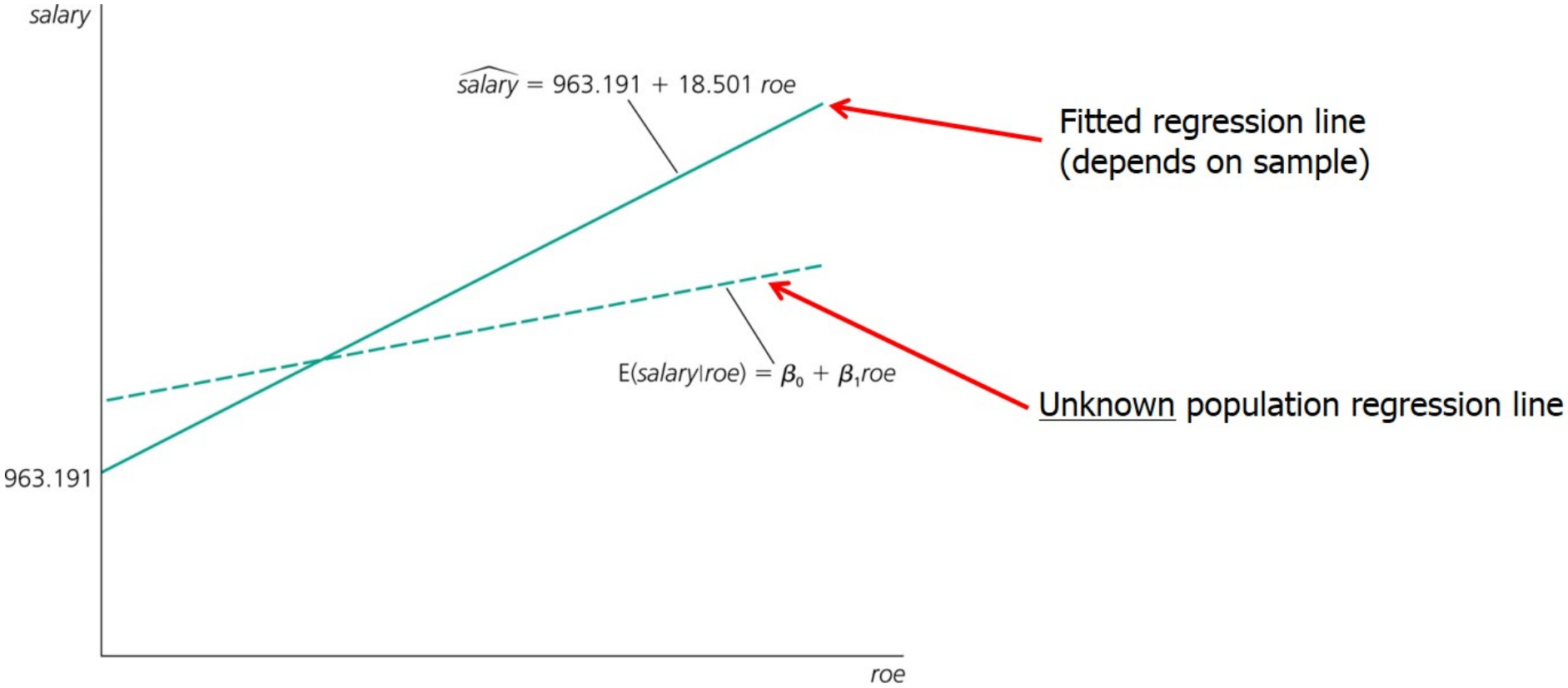
where *salary* is the CEO salary in thousands of dollars, and *roe* is the return on equity of the CEO's firm in percentage.

- The fitted regression is

$$\widehat{salary} = 963.191 + 18.501 roe,$$

where $\widehat{\beta}_1 = 18.501 > 0$, which means that if the return on equity increases by 1 percent, then salary is predicted to change by $\$18,501$. [figure here]

- Even if $roe = 0$, the predicted salary of CEO is $\$963,191$.
- Causal Interpretation of $\widehat{\beta}_1$? Think about what factors are included in u (e.g., market share, sales, tenure, character of the CEO, etc.) and check whether $Cov(x,u) = 0$.


salary
$\widehat{salary} = 963.191 + 18.501\ roe$
Fitted regression line
(depends on sample)
E(salary|roe) = $\beta_0 + \beta_1 roe$
Unknown population regression line
963.191
roe

Wage and Education

- Suppose the SLR model is

$$wage = \beta_0 + \beta_1 educ + u,$$

where *wage* is the hourly wage in dollars, and *educ* is years of education.

- The fitted regression is

$$\widehat{wage} = -0.90 + 0.54 educ,$$

where $\widehat{\beta}_1 = 0.54 > 0$, which means that in the sample, one more year of education was associated with an increase in hourly wage by \$0.54 (which is quite large! e.g., four years college would increase the wage by $\$0.54 \times 4 = \$2.16$ per hour, which implies an increase of $\$2.16 \times 40 \times 52 \approx \$4,500$ per year).

- $\widehat{\beta}_0 = -0.90$ means when $educ = 0$, *wage* is negative. Does this make sense? [figure here]
- Do you think the return to education is constant? (see slide 51 below)
- Causal Interpretation of $\widehat{\beta}_1$? No.

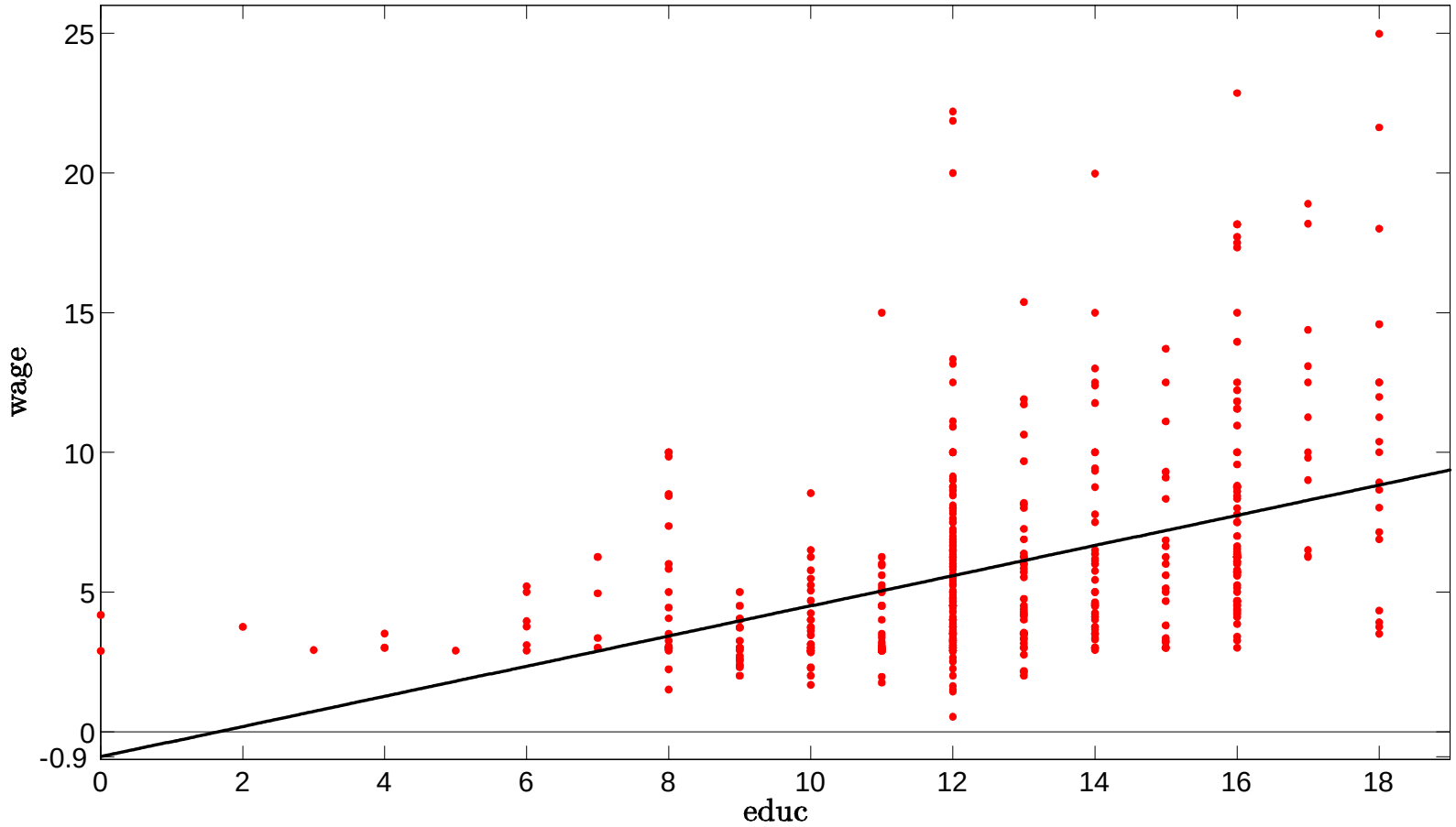


Figure: $\widehat{wage} = -0.90 + 0.54 educ$: only two people have $educ = 0$

Voting Outcomes and Campaign Expenditures (Two Parties)

- Suppose the SLR model is

$$voteA = \beta_0 + \beta_1 shareA + u,$$

where *voteA* is the percentage of vote for candidate *A*, and *shareA* is the percentage of total campaign expenditures spent by candidate *A*.

- The fitted regression is

$$\widehat{voteA} = 26.81 + 0.464 shareA,$$

where $\widehat{\beta}_1 = 0.464 > 0$, which means if candidate *A*'s share of spending increases by one percentage point, he or she receives 0.464 (about one half) percentage points more of the total vote.

- If candidate *A* does not spend any on campaign, then he or she will receive about 26.81% of the total vote. If $shareA = 50$, then $\widehat{voteA}$ is roughly 50.
- Causal Interpretation of $\widehat{\beta}_1$? Maybe OK - *u* includes the quality of the candidates, dollar amounts (not percentage) spent by *A* and *B*, etc.

Properties of OLS on Any Sample of Data

a: Fitted Values and Residuals

- $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \left(\overline{y} - \overline{x}\widehat{\beta}_1\right) + \widehat{\beta}_1 x_i = \overline{y} + \widehat{\beta}_1 (x_i - \overline{x})$ is called the fitted or predicted value at x_i.
- $\widehat{u}_i \equiv \widehat{u}_i\left(\widehat{\beta}\right) = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i = y_i - \widehat{y}_i$ is called the residual, which is the deviation of y_i from the fitted regression line.[11] [figure here]
- $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is called the OLS regression line or sample regression function (SRF).

TABLE 2.2 Fitted Values and Residuals for the First 15 CEOs

obsno	roe	salary	salaryhat	uhat
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606
9	10.5	1237	1157.454	79.54626
10	26.3	833	1449.773	−616.7726
11	25.9	567	1442.372	−875.3721
12	26.8	933	1459.023	−526.0231
13	14.8	1339	1237.009	101.9911
14	22.3	937	1375.768	−438.7678
15	56.3	2011	2004.808	6.191895

x_i

y_i

$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$

$\widehat{u}_i = y_i - \widehat{y}_i$

For example, CEO number 12's salary was \$526,023 lower than predicted using the information on his firm's return on equity

[11] $\widehat{u}_i$ is different from $u_i = y_i - \beta_0 - \beta_1 x_i$. The later is unobservable while the former is a by-product of OLS estimation.

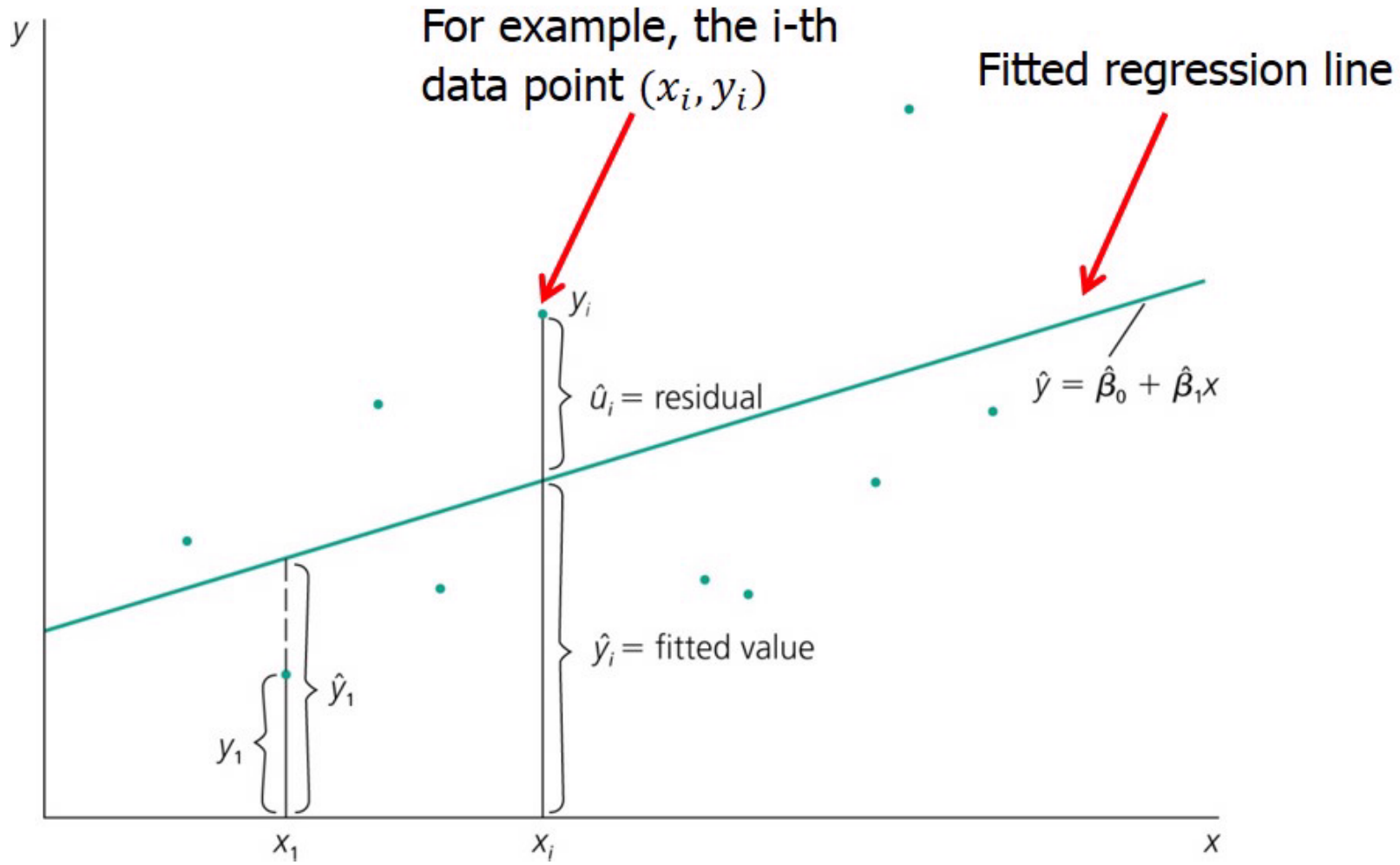


Figure: Fitted Values and Residuals

b: Algebraic Properties of OLS Statistics

- Check the figure above to understand the following properties.
- The key is the two FOCs, and all other results are corollaries.

- $\sum_{i=1}^n \widehat{u}_i = 0$: it must be the case that some residuals are positive and others are negative, so the fitted regression line must lie in the middle of the data points.

  - This property implies $\overline{y} \overset{?}{=} \overline{\widehat{y}} + \overline{\widehat{u}} = \overline{\widehat{y}}$, where ? is because $y_i = \widehat{y}_i + \widehat{u}_i$.

  - This FOC is equivalent to $\overline{y} = \widehat{\beta}_0 + \overline{x}\widehat{\beta}_1$: The fitted regression line passes through $(\overline{x}, \overline{y})$.
- $\sum_{i=1}^n x_i \widehat{u}_i = 0$:

$$\widehat{Cov}(x, \widehat{u}) = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})\left(\widehat{u}_i - \overline{\widehat{u}}\right) = \frac{1}{n}\sum_{i=1}^n x_i\left(\widehat{u}_i - \overline{\widehat{u}}\right) = \frac{1}{n}\sum_{i=1}^n x_i\widehat{u}_i = 0.$$

where the second equality is because we need only demean one of x and $\widehat{u}$, and the second to last equality is because $\overline{\widehat{u}} = 0$.

continue

- These two properties are the sample analogs of $E[u]=0$ and $Cov(x,u)=0$ which are implied by $E[u|x]=0$ [proof not required].
- These two properties imply

$$\sum_{i=1}^{n}\widehat{y}_i\widehat{u}_i=\sum_{i=1}^{n}\left(\widehat{\beta}_0+\widehat{\beta}_1x_i\right)\widehat{u}_i=\widehat{\beta}_0\sum_{i=1}^{n}\widehat{u}_i+\widehat{\beta}_1\sum_{i=1}^{n}x_i\widehat{u}_i=0.$$

  - This means $\widehat{Cov}(\widehat{y},\widehat{u})=n^{-1}\sum_{i=1}^{n}\widehat{y}_i\left(\widehat{u}_i-\overline{\widehat{u}}\right)=n^{-1}\sum_{i=1}^{n}\widehat{y}_i\widehat{u}_i=0.$

The Cooked Numerical Example Revisited

	y_i	x_i	$\widehat{y}_i$	$\widehat{u}_i$	$x_i\widehat{u}_i$	$\widehat{y}_i\widehat{u}_i$
	4	1	2.929	1.071	1.071	3.138
	5	4	7.000	−2.000	−8.000	−14.000
	7	5	8.357	−1.357	−6.785	−11.342
	12	6	9.714	2.286	13.714	22.204
Sum: $\sum_{i=1}^4$	28	16	28	0	0	0
Mean: $\frac{1}{4}\sum_{i=1}^4$	7	4	7	0	0	0

Table: Check Algebraic Properties of OLS Statistics: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ and $\widehat{u}_i = y_i - \widehat{y}_i$

- $\overline{y} = \widehat{\beta}_0 + \overline{x}\widehat{\beta}_1$: $7 = 1.571 + 4 \times 1.357$.

c: Measures of Variation

- How well does the explanatory variable explain the dependent variable?
- Measures of Variation:

$$\begin{aligned} SST &\equiv \sum_{i=1}^{n} (y_i - \bar{y})^2, \\ SSE &\equiv \sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2 \overset{?}{=} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \\ SSR &\equiv \sum_{i=1}^{n} \left(\hat{u}_i - \bar{\hat{u}}\right)^2 \overset{?}{=} \sum_{i=1}^{n} \hat{u}_i^2 = SSR\left(\hat{\beta}\right), \end{aligned}$$

where
SST = total sum of squares, represents total variation in dependent variable,
SSE = explained sum of squares, represents variation explained by regression,
SSR = residual sum of squares, represents variation not explained by regression.

- It can be shown that

$$SST = SSE + SSR.$$

(*) Decomposition of Total Variation

- Note that

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\ &= \sum_{i=1}^{n} [(y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})]^2 \\ &= \sum_{i=1}^{n} [\widehat{u}_i + (\widehat{y}_i - \overline{y})]^2 \\ &= \sum_{i=1}^{n} \widehat{u}_i^2 + 2\sum_{i=1}^{n} \widehat{u}_i (\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 \\ &= SSR + 2\sum_{i=1}^{n} \widehat{u}_i (\widehat{y}_i - \overline{y}) + SSE \\ &= SSR + SSE, \end{aligned}$$

where the last equality is because $\sum_{i=1}^n \widehat{u}_i \widehat{y}_i = 0$ and $\sum_{i=1}^n \widehat{u}_i \overline{y} = \overline{y} \sum_{i=1}^n \widehat{u}_i = 0$.

Goodness-of-Fit

- The R-squared of the regression, also called the coefficient of determination, is defined as
$$R^2 = \frac{SSE}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST}.$$
- R-squared measures the fraction of the total variation that is explained by the regression.
- $0 \le R^2 \le 1$. When $R^2 = 0$? When $R^2 = 1$? [figure here]
  - R^2 tries to explain variation not level; a constant cannot explain variation (but explains only level), so $R^2 = 0$ if only the constant contributes to the regression: if $\widehat{\beta}_1 = 0$, then $\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1 = \overline{y}$, so
$$SSR = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - x_i\widehat{\beta}_1\right)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 = SST.$$
  - R^2 is defined only if there is an intercept;[12] we need to use the constant to absorb the level of y, and then use x_i to measure the variation of y_i:
$$\begin{aligned} SSE &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} \left(\widehat{\beta}_0 + x_i\widehat{\beta}_1 - \overline{y}\right)^2 \\ &= \sum_{i=1}^{n} \left(\overline{y} - \overline{x}\widehat{\beta}_1 + x_i\widehat{\beta}_1 - \overline{y}\right)^2 = \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 = \widehat{\beta}_1^2 SST_x. \end{aligned}$$

[12] See Assignment I, Problem 5(i), for rigorous justification.

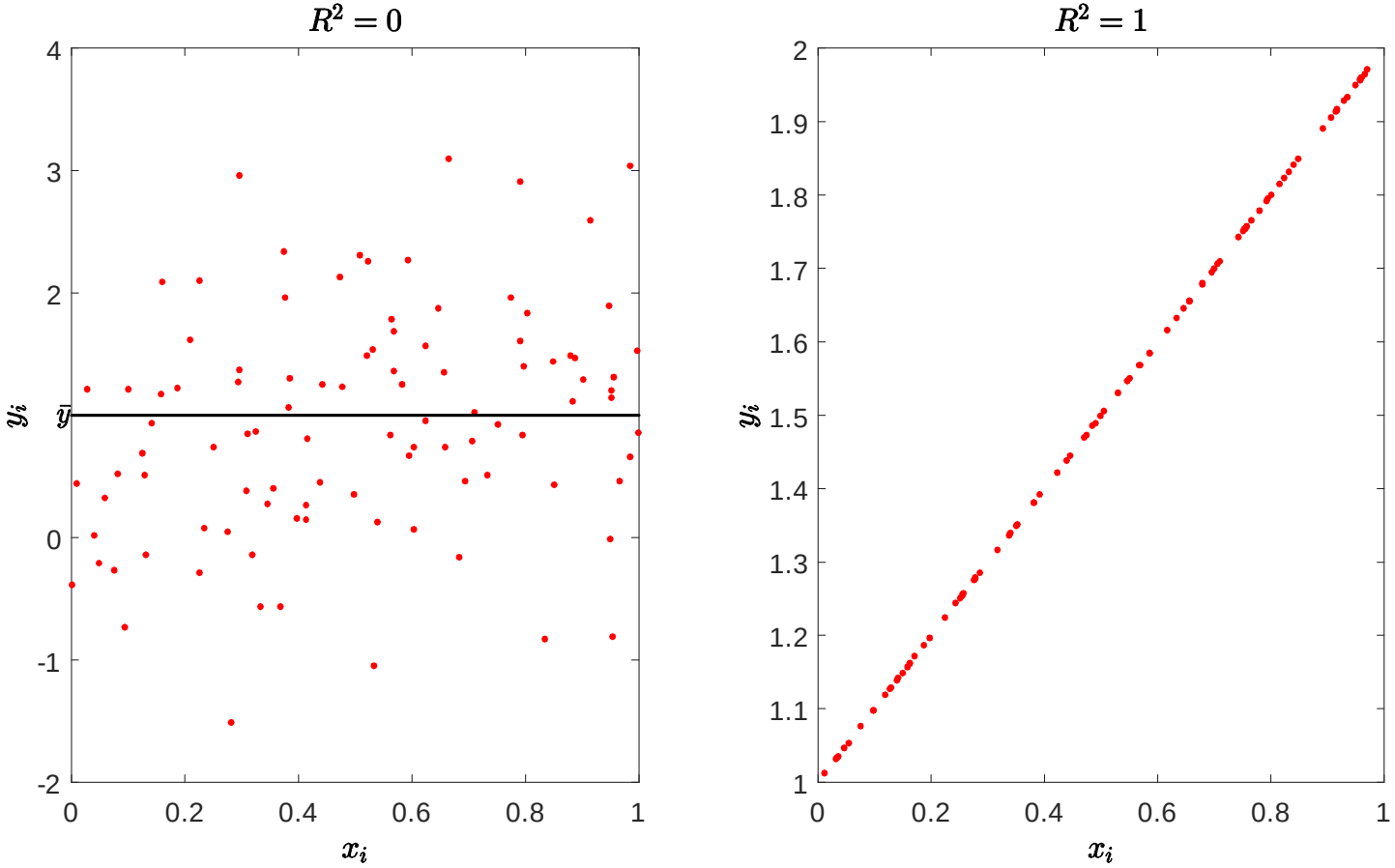


Figure: Data Patterns for $R^2 = 0$ and $R^2 = 1$

- Caution: A high R-squared does not necessarily mean that the regression has a causal interpretation! [check the following two examples]

Two Examples of R-Squared

- CEO Salary and Return on Equity:

$$\begin{aligned} \widehat{salary} &= 963.191 + 18.501 roe, \\ n &= 209, R^2 = 0.0132.^{13} \end{aligned}$$

  - The regression explains only 1.3% of the total variation in salaries.

- Voting Outcomes and Campaign Expenditures:

$$\begin{aligned} \widehat{voteA} &= 26.81 + 0.464 shareA, \\ n &= 173, R^2 = 0.856. \end{aligned}$$

  - The regression explains 85.6% of the total variation in election outcomes.

[13] It is quite standard to have a low R^2 for cross-sectional data because a lot of heterogeneities are contained in u.

Units of Measurement and Functional Form

b: Incorporating Nonlinearities in Simple Regression

- The effects of changing units of measurement on OLS statistics will be discussed in Chapter 6.
- Regression of log wages on years of education:

$$\log(wage) = \beta_0 + \beta_1 educ + u,$$

where $\log(\cdot)$ denotes the natural logarithm. [figure here]

- This is often called semi-log or log-linear regression model.
- This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\partial \log(wage)}{\partial educ} = \frac{1}{wage} \cdot \frac{\partial wage}{\partial educ} = \frac{\partial wage/wage}{\partial educ},$$

where $\partial wage/wage$ is the proportional change of wage. [see the next slide for math review]

- Or,

$$100\beta_1 = \frac{100 \cdot \partial wage/wage}{\partial educ} = \frac{\%\Delta wage}{\Delta educ},$$

where $\%\Delta$ is read as "percentage change of", and Δ is read as "change of".

[Review] Derivative of Logarithmic Functions

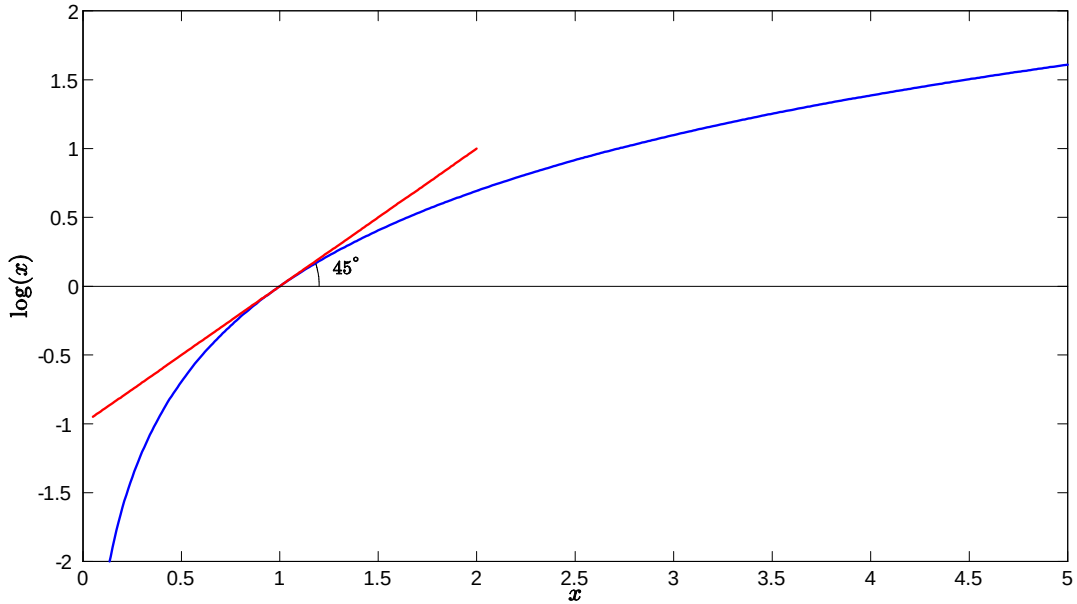


Figure: $\log(x) : x > 0$; *wage* > 0

- Recall that
$$\frac{d\log x}{dx} = \frac{1}{x} \text{ or } d\log x = \frac{dx}{x}.$$
- The derivative gets smaller and smaller as x gets larger and larger: $\lim_{x\to 0} \frac{d\log x}{dx} = \infty$, $\left.\frac{d\log x}{dx}\right|_{x=1} = 1$, $\lim_{x\to\infty} \frac{d\log x}{dx} = 0$.

A Log Wage Equation

- The fitted regression line is

$$\begin{aligned}\widehat{\log(wage)} &= 0.584 + 0.083 educ, \\ n &= 526, R^2 = 0.186,\end{aligned}$$

  which implies

$$\widehat{wage} \approx e^{0.584 + 0.083 educ}.$$

- The wage increases by 8.3% for every additional year of education (= return to education).
- For example, if the current wage is \$10 per hour (which implies that $educ = \frac{\log(10) - 0.584}{0.083}$), and suppose the education is increased by one year. Then

$$\Delta wage = \exp\left(0.584 + 0.083\left(\frac{\log(10) - 0.584}{0.083} + 1\right)\right) - 10 = 0.865 \approx 0.83,$$

  and

$$\frac{\partial wage / wage}{\partial educ} = \frac{+\$0.83/\$10}{+1 \text{ year}} = 0.083 = 8.3\%.$$

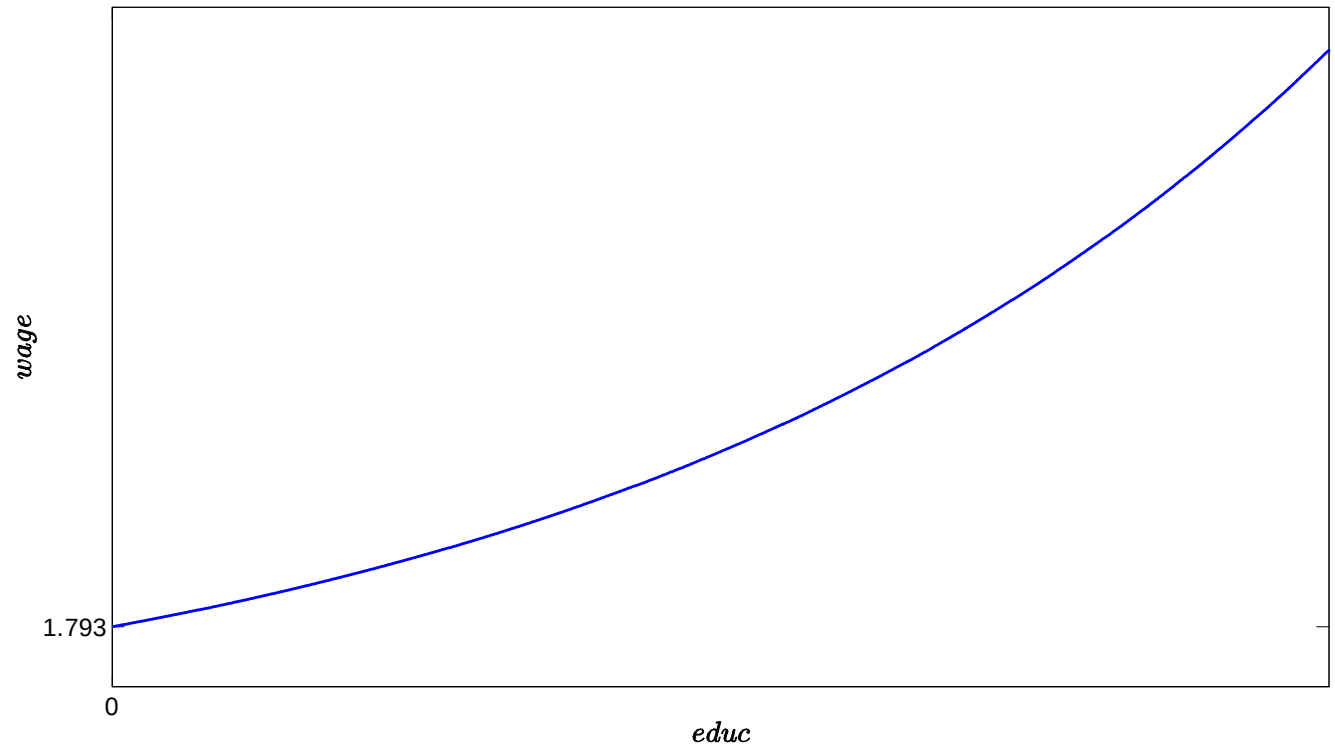


Figure: $wage = \exp(0.584 + 0.083 educ)$

- When the wage level is higher, the increase in wage for one more year of eduction is larger, but the percentage increase of wage is the same.

Constant Elasticity Model

- CEO Salary and Firm Sales:

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u,$$

where *sales* is measured in millions of dollars.

- This changes the interpretation of the regression coefficient:

$$\beta_1 = \frac{\partial \log(salary)}{\partial \log(sales)} = \frac{\partial salary / salary}{\partial sales / sales} = \frac{\%\Delta salary}{\%\Delta sales} = \text{elasticity}.$$

- The log-log form postulates a constant elasticity model, whereas the semi-log form assumes a semi-elasticity model with $100\beta_1$ called the semi-elasticity of y with respect to x: in the log wage equation,

$$\text{elasticity} = \frac{\partial \log(wage)}{\partial \log(educ)} = \frac{\partial \log(wage)}{\partial educ / educ} = \beta_1 \cdot educ,$$

which depends on *educ*. The elasticity is larger for a higher education level.

CEO Salary and Firm Sales

- The fitted regression line is

$$\begin{aligned}\widehat{\log(salary)} &= 4.822 + 0.257\log(sales),\\ n &= 209, R^2 = 0.211,\end{aligned}$$

  which implies

$$\widehat{salary} \approx e^{4.822+0.257\log(sales)} = e^{4.822} sales^{0.257}.$$

- The salary increases by 0.257% for every 1% increase of sales.

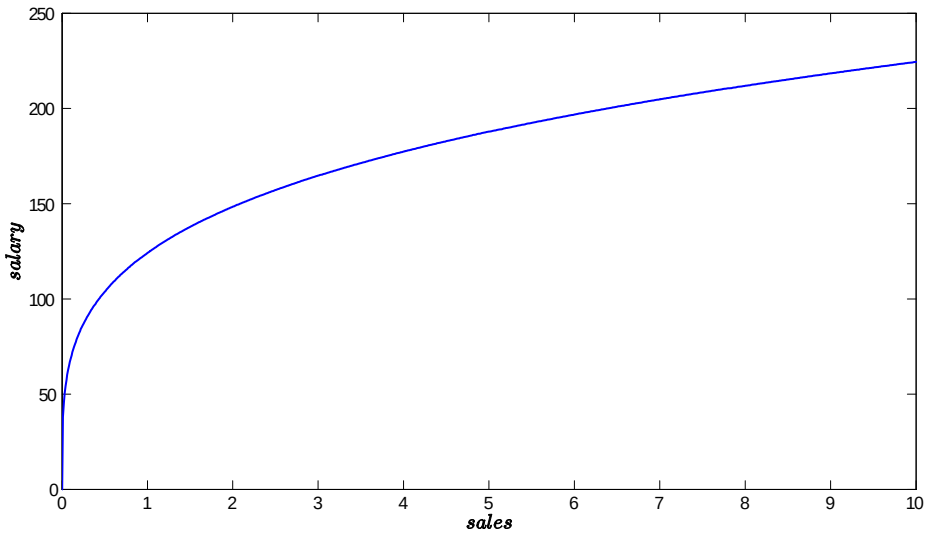


Figure: $salary = e^{4.822} sales^{0.257}$

Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = \frac{\beta_1}{100} \%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Table: Summary of Functional Forms Involving Logarithms

Expected Values and Variances of the OLS Estimators

Statistical Properties of OLS Estimators

- The property such as $\sum_{i=1}^{n}\widehat{u}_i = 0$ is satisfied by any sample of data, i.e., regardless of the values of $\{(x_i, y_i) : i = 1, \cdots, n\}$, this property must satisfy.
- We now treat $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as estimators, i.e., treat them as random variables because they are calculated from a random sample.
- Recall that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ and } \widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1,$$

  where the data $\{(x_i, y_i) : i = 1, \cdots, n\}$ is random and depends on the particular sample that has been drawn.
- Caution: distinguish a random variable and its realization!
- Question: What will the estimators estimate on average and how large is their variability in repeated samples? i.e.,

$$E\left[\widehat{\beta}_0\right] = ?, E\left[\widehat{\beta}_1\right] = ? \text{ and } Var\left(\widehat{\beta}_0\right) = ?, Var\left(\widehat{\beta}_1\right) = ?$$

Standard Assumptions for the SLR Model

- Scientific approach requires assumptions! Science vs. Superstition

- Assumption SLR.1 (Linear in Parameters):

$$y = \beta_0 + \beta_1 x + u.$$

  - In the population, the relationship between y and x is linear.
  - The "linear" in linear regression means "linear in parameter", e.g., $y = \beta_0 + \beta_1 \log(x) + u$ is a linear regression.

- Assumption SLR.2 (Random Sampling): The data $\{(x_i, y_i) : i = 1, \cdots, n\}$ is a random sample drawn from the population, i.e., each data point follows the population equation,

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Discussion of Random Sampling: Wage and Education

- The population consists, for example, of all workers of country A.
- In the population, a linear relationship between wages (or log wages) and years of education holds.
- Draw completely randomly a worker from the population.
- The wage and the years of education of the worker drawn are random because one does not know beforehand which worker is drawn.
- Throw back worker into population and repeat random draw n times.
- The wages and years of education of the sampled workers are used to estimate the linear relationship between wages and education.

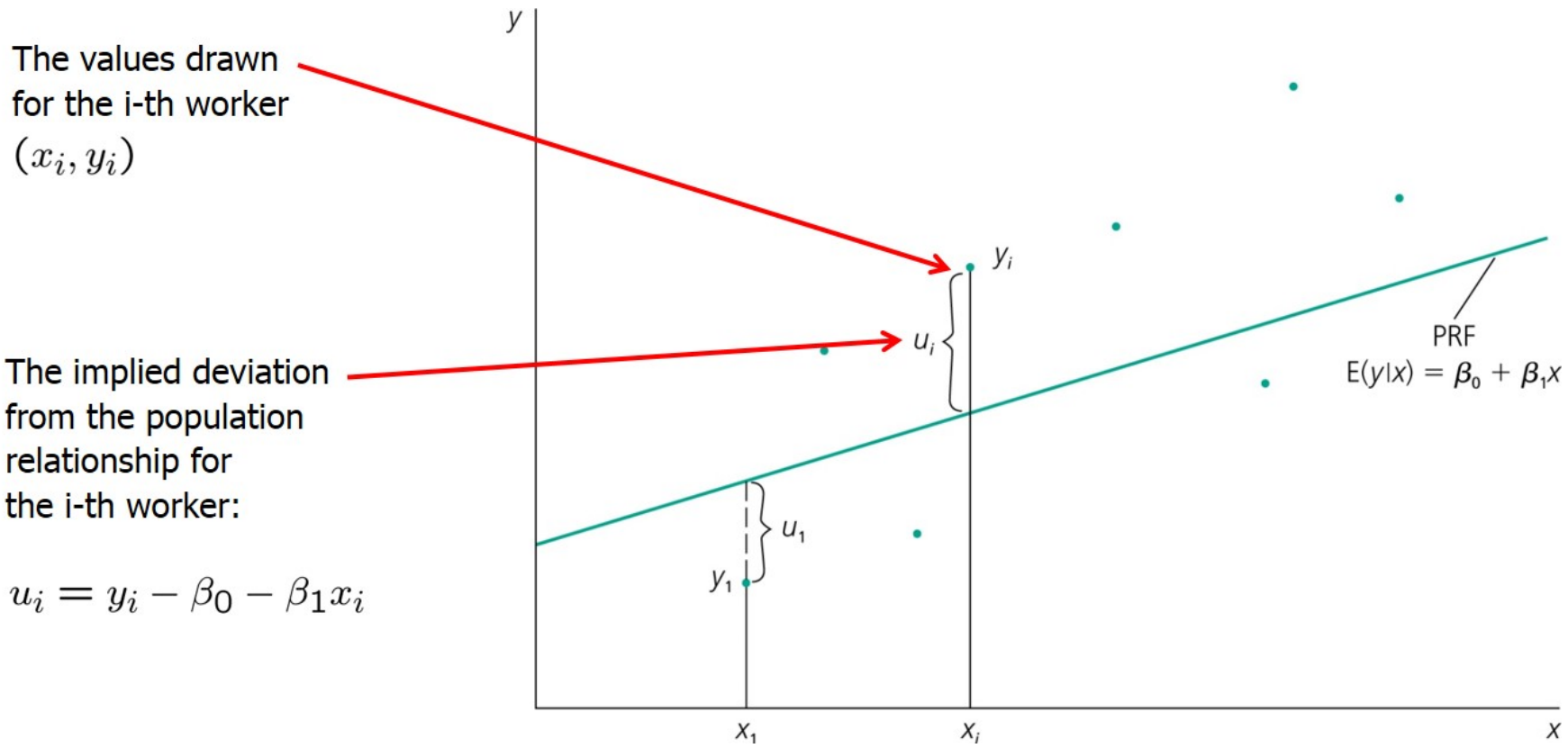


Figure: Graph of $y_i = \beta_0 + \beta_1 x_i + u_i$.

continue

- Assumption SLR.3 (Sample Variation in Explanatory Variable): $\sum_{i=1}^{n}(x_i-\overline{x})^2>0$.
  - The values of the explanatory variables are not all the same (otherwise it would be impossible to study how much the dependent variable changes when the explanatory variable changes one unit - β_1). [figure here]
  - Note that $\sum_{i=1}^{n}(x_i-\overline{x})^2$ is the demonimator of $\widehat{\beta}_1$. If $\sum_{i=1}^{n}(x_i-\overline{x})^2=0$, then $\widehat{\beta}_1$ is not defined.
- Assumption SLR.4 (Zero Conditional Mean):

$$E[u_i|x_i]=0.$$

  - The value of the explanatory variable must contain no information about the mean of the unobserved factors.

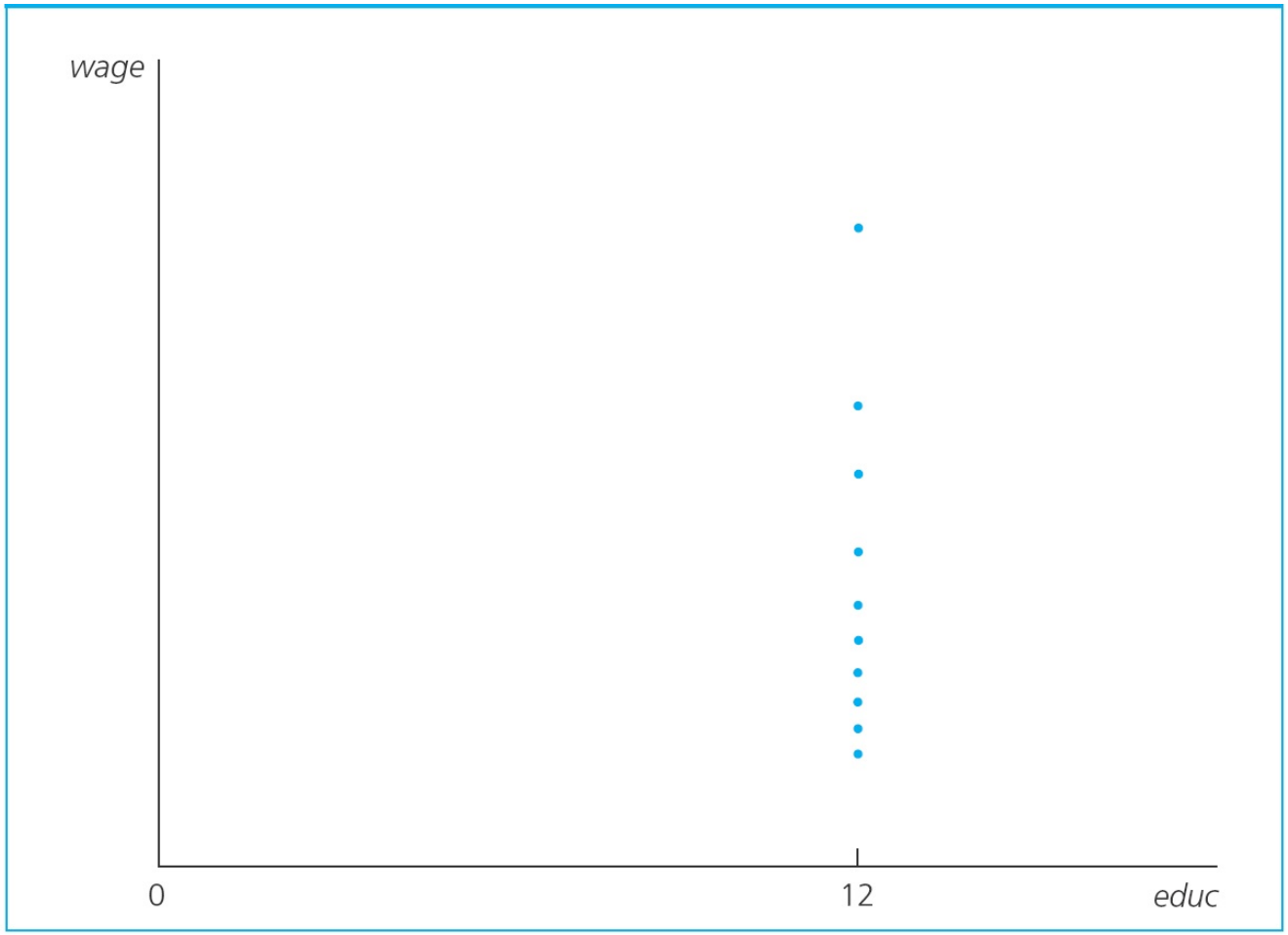


Figure: A Scatterplot of Wage Against Education When $educ_i = 12$ for **All** i

a: Unbiasedness of OLS

- Theorem 2.1: Under assumptions SLR.1-SLR.4,

$$E\left[\widehat{\beta}_0\right] = \beta_0 \text{ and } E\left[\widehat{\beta}_1\right] = \beta_1$$

  for any values of β_0 and β_1.
- How to understand unbiasedness?
- The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw.
- However, on average, they will be equal to the values that characterize the true relationship between y and x in the population.
- "On average" means if sampling was repeated, i.e., if drawing the random sample and doing the estimation was repeated many times.
- In a given sample, estimates may differ considerably from true values.

(*) Proof of Unbiasedness of OLS

Proof.

We always condition on $\{x_i, i = 1, \cdots, n\}$, i.e., the x values can be treated as fixed. Now, the only randomness is from $\{u_i, i = 1, \cdots, n\}$. Note that

$$\begin{aligned}
\widehat{\beta}_1 - \beta_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} - \beta_1 \\
&\overset{SLR.1-2}{=} \frac{\sum_{i=1}^n (x_i - \overline{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (x_i - \overline{x})^2} - \beta_1 \\
&= \beta_0 \frac{\sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} + \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x}) x_i}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{\sum_{i=1}^n (x_i - \overline{x})^2} - \beta_1 \\
&= \frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{\sum_{i=1}^n (x_i - \overline{x})^2},
\end{aligned}$$

where the last equality is because

$$\sum_{i=1}^n (x_i - \overline{x}) = 0 \text{ and } \sum_{i=1}^n (x_i - \overline{x}) x_i = \sum_{i=1}^n (x_i - \overline{x})^2 \overset{SLR.3}{>} 0.$$

Proof continue

Now,

$$E\left[\widehat{\beta}_1\right]-\beta_1 = E\left[\frac{\sum_{i=1}^n (x_i-\overline{x})u_i}{\sum_{i=1}^n (x_i-\overline{x})^2}\right] \overset{\text{(ii)}}{=} \frac{E\left[\sum_{i=1}^n (x_i-\overline{x})u_i\right]}{\sum_{i=1}^n (x_i-\overline{x})^2}$$
$$\overset{\text{(i)}}{=} \frac{\sum_{i=1}^n E\left[(x_i-\overline{x})u_i\right]}{\sum_{i=1}^n (x_i-\overline{x})^2} \overset{\text{(ii)}}{=} \frac{\sum_{i=1}^n (x_i-\overline{x})E\left[u_i\right]}{\sum_{i=1}^n (x_i-\overline{x})^2} \overset{14}{\underset{SLR.2-4}{=}} 0.$$

Further, since $\overline{y}=\beta_0+\beta_1\overline{x}+\overline{u}$,

$$E\left[\widehat{\beta}_0\right] = E\left[\overline{y}-\overline{x}\widehat{\beta}_1\right] = E\left[\beta_0-\left(\widehat{\beta}_1-\beta_1\right)\overline{x}+\overline{u}\right]$$
$$= \beta_0-\overline{x}E\left[\widehat{\beta}_1-\beta_1\right]+E\left[\overline{u}\right]=\beta_0,$$

where the last equality is because $\widehat{\beta}_1$ is unbiased, and $E\left[\overline{u}\right]=\frac{1}{n}\sum_{i=1}^n E\left[u_i\right]=0$ by Assumptions SLR.2 and SLR.4. □

- The key assumption for unbiasedness is Assumption SLR.4.

[14] $E\left[u_i|x_1,\cdots,x_n\right] \overset{SLR.2}{=} E\left[u_i|x_i\right] \overset{SLR.4}{=} 0.$

b: Variances of the OLS Estimators

- Unbiasedness is not the only desirable property of the OLS estimator. [intuition here: gunfire][figure here]
- Depending on the sample, the estimates will be nearer or farther away from the true population values.
- How far can we expect our estimates to be away from the true population values on average (= sampling variability)?
- Sampling variability is measured by the estimator's variances. [see the next slides for review of variance]

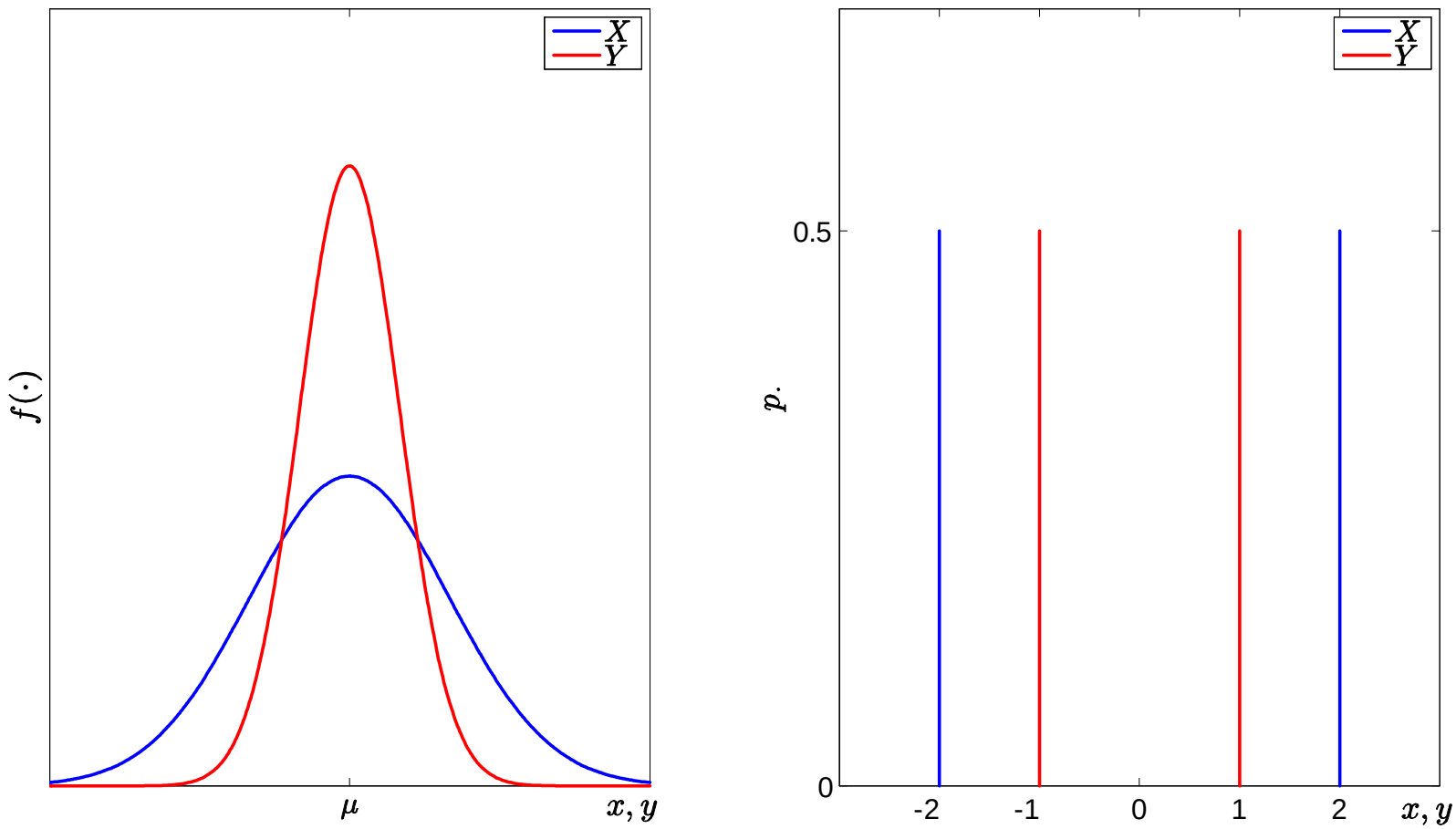


Figure: Random Variables with the Same Mean BUT Different Distributions

[Review] Variance

- Recall that $Var(X) = E\left[(X - E[X])^2\right]$ measures how spreading the distribution of a r.v. X is [figure above].
- For example, consider two random variables X and Y with [figure here]

$$P(X = -2) = P(X = 2) = 1/2 \text{ and } P(Y = -1) = P(Y = 1) = 1/2.$$

- Obviously, X is more spreading than Y (i.e., more probabilities on large values) although both have the mean zero.
- If we check their variances, then indeed,

$$Var(X) = \frac{1}{2}(-2)^2 + \frac{1}{2}2^2 = 4 > 1 = \frac{1}{2}(-1)^2 + \frac{1}{2}1^2 = Var(Y).$$

- The standard deviation of a r.v., denoted as sd(X), is simply the square root of the variance:

$$sd(X) = \sqrt{Var(X)}.$$

- The name "standard deviation" came from Karl Pearson.
- Variance measures the expected squared "deviation" from the mean and has the unit of the squared unit of X.
- By taking $\sqrt{\cdot}$ in sd(X), we get back to the "standard" (original) unit of X.

continue

- If $X = 1(\text{college graduate})$, then $Var(X) = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)$. [intuition here]
- If $X \sim N\left(\mu, \sigma^2\right)$, then $Var(X) = \sigma^2$.
- Two Useful Properties: (i) for **independent** r.v.'s, "the variance of the sum is the sum of the variances",

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i);$$

(ii) for any constants a and b, $Var(a+bX) = b^2 Var(X)$.
- $Var(a+bX) = E\left[(a+bX-a-bE[X])^2\right] = b^2 E\left[(X-E[X])^2\right] = b^2 Var(X)$.

- These two properties imply

$$\begin{aligned} Var(\overline{x}) &= Var\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \overset{\text{(ii)}}{=} \frac{1}{n^2} Var\left(\sum_{i=1}^n x_i\right) \\ &\overset{SLR.2+\text{(i)}}{=} \frac{1}{n^2}\sum_{i=1}^n Var(x_i) \overset{SLR.2}{=} \frac{n Var(x)}{n^2} = \frac{Var(x)}{n}, \end{aligned}$$

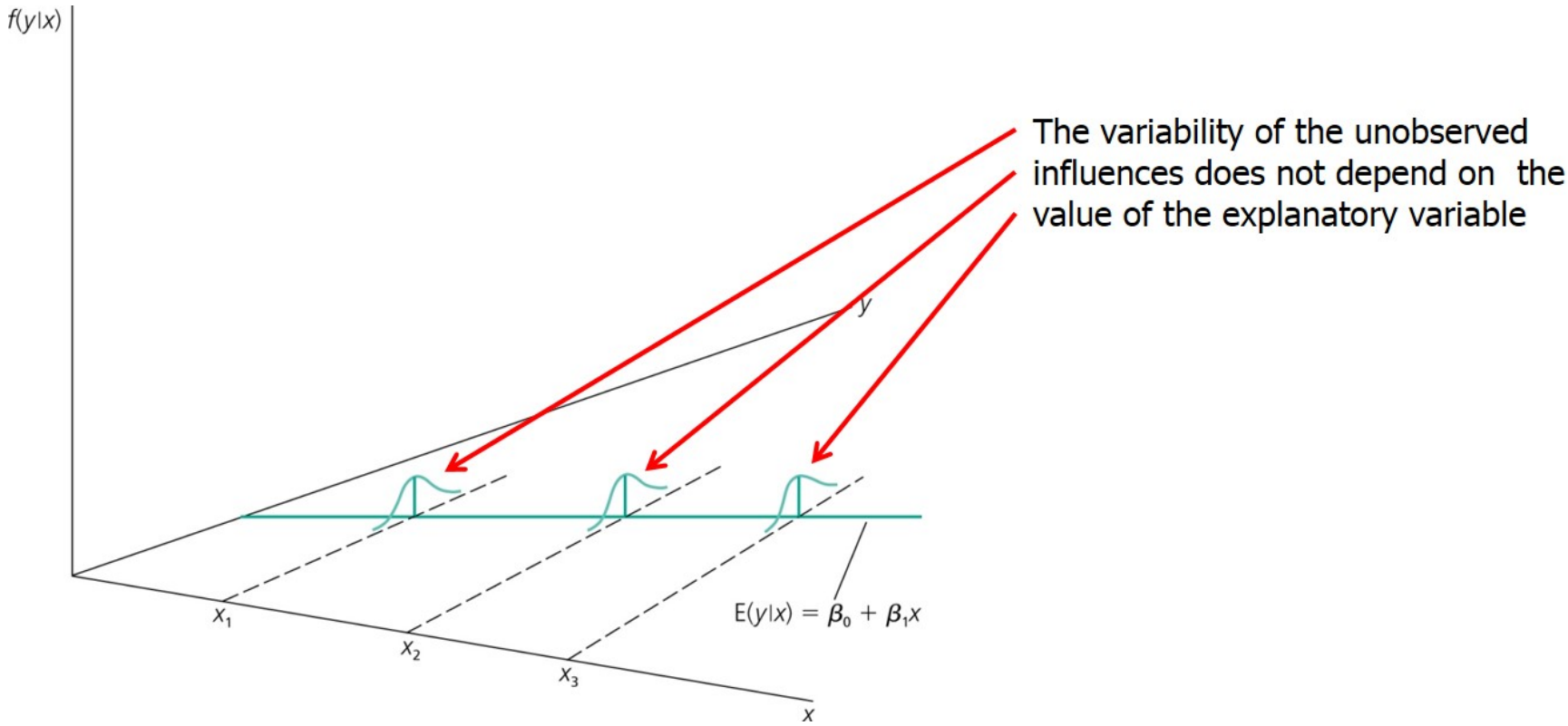
which decreases with n and increases with $Var(x)$.

[Review] Conditional Variance

- As the conditional mean, the conditional variance of Y given $X = x$, denoted as $Var(Y|X = x)$, is the variance of Y for the (slice of) individuals with $X = x$. Rigorously,
$$Var(Y|X = x) = E\left[(Y - E[Y|X = x])^2 |X = x\right].$$
- Apply the second property of variance to the conditional variance to have
$$Var(y_i|x_i) = Var(y_i - \beta_0 - \beta_1 x_i|x_i) = Var(u_i|x_i),$$
where as mentioned in the conditional mean, conditional on x_i, $\beta_0 + \beta_1 x_i$ can be treated as a constant like a in (ii).
- Although $E[y_i|x_i] = \beta_0 + \beta_1 x_i$ is linear in x_i (Assumption SLR1, 2 and 4), $Var(y_i|x_i)$ is assumed not to depend on x_i (Assumption SLR.5 below).

Homoskedasticity

- Assumption SLR.5 (Homoskedasticity): $Var\,(u_i|x_i) = \sigma^2$.
  - The value of the explanatory variable must contain no information about the variability of the unobserved factors.



Heteroskedasticity

- When $Var\left(u_i|x_i\right)$ depends on x_i, the error term is said to exhibit heteroskedasticity.

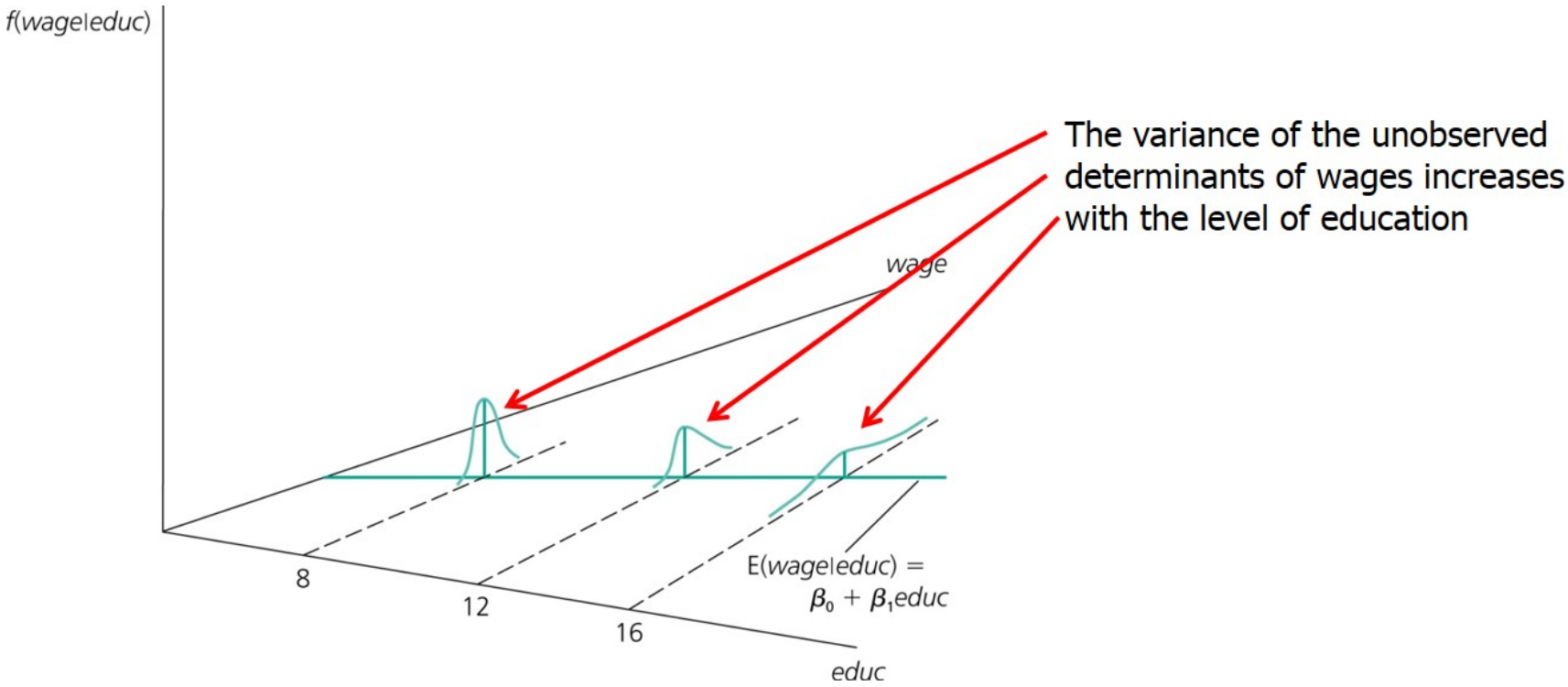


Figure: An Example for Heteroskedasticity: Wage and Education

Variances of OLS Estimators

- Theorem 2.2: Under assumptions SLR.1-SLR.5,

$$\begin{aligned} Var\left(\widehat{\beta}_1\right) &= \frac{\sigma^2}{\sum_{i=1}^n \left(x_i-\overline{x}\right)^2} = \frac{\sigma^2}{SST_x}, \\ Var\left(\widehat{\beta}_0\right) &= \frac{\sigma^2 n^{-1}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n \left(x_i-\overline{x}\right)^2} = \frac{\sigma^2 n^{-1}\sum_{i=1}^n x_i^2}{SST_x}. \end{aligned}$$

- The sampling variability of the estimated regression coefficients will be the higher the larger the variability of the unobserved factors, and the lower, the higher the variation in the explanatory variable ($SST_x = n \cdot \widehat{Var}\left(x\right)$). [figure here]

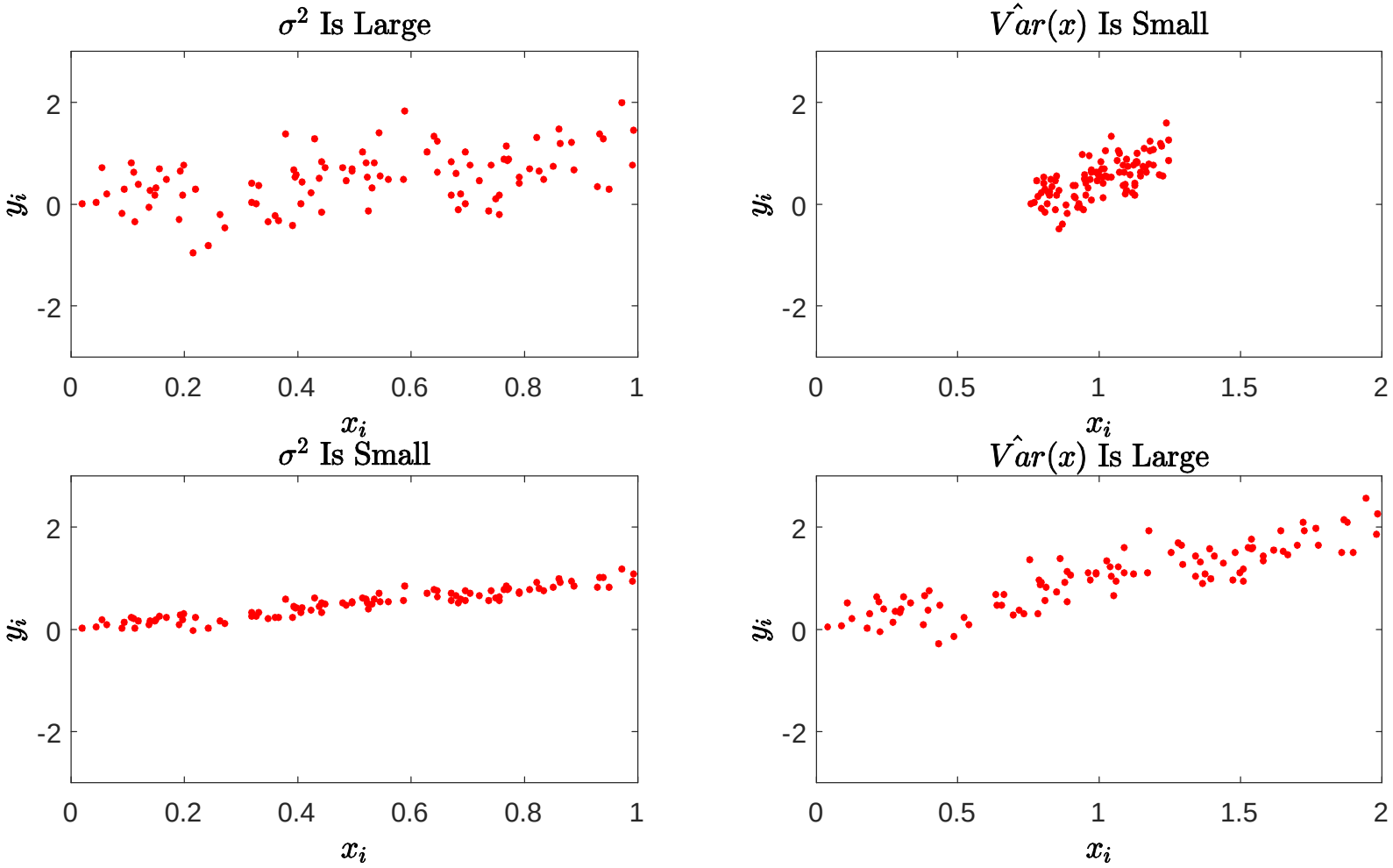


Figure: Relative Difficulty in Identifying β_1

(*) Proof of Theorem 2.2

Proof.

$Var\left(\widehat{\beta}_1\right)$ is more important, so we concentrate on it here. As in the proof of Theorem 2.1, we condition on $\{x_i, i = 1, \cdots, n\}$.

$$\begin{aligned} Var\left(\widehat{\beta}_1\right) &\overset{\text{(ii)}}{=}{}^{15} Var\left(\widehat{\beta}_1 - \beta_1\right) = Var\left(\frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \\ &\overset{\text{(ii)}}{=} \frac{Var\left(\sum_{i=1}^n (x_i - \overline{x}) u_i\right)}{SST_x^2} \overset{SLR.2+\text{(i)}}{=} \frac{\sum_{i=1}^n Var\left((x_i - \overline{x}) u_i\right)}{SST_x^2} \\ &\overset{\text{(ii)}}{=} \frac{\sum_{i=1}^n (x_i - \overline{x})^2 Var\left(u_i\right)}{SST_x^2} \overset{16}{=} \overset{SLR.2+5}{=} \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \sigma^2}{SST_x^2} \\ &= \frac{\sigma^2 SST_x}{SST_x^2} = \frac{\sigma^2}{SST_x}. \end{aligned}$$

□

- The key assumption to get this simple formula of $Var\left(\widehat{\beta}_1\right)$ is Assumption SLR.5.
- The only unknown component of $Var\left(\widehat{\beta}_1\right)$ and $Var\left(\widehat{\beta}_0\right)$ is σ^2 which is estimated below.

[15] Also because $E\left[\widehat{\beta}_1\right] = \beta_1$ and we want to "center" $\widehat{\beta}_1$ at zero.

[16] $Var\left(u_i | x_1, \cdots, x_n\right) \overset{SLR.2}{=} Var\left(u_i | x_i\right) \overset{SLR.5}{=} \sigma^2$.

c: Estimating the Error Variance (assuming homoskedasticity)

- $Var(u_i|x_i) = \sigma^2 \overset{SLR.4+5}{=} Var(u_i)$ [proof not required].
  - The variance of u does not depend on x, i.e., is equal to the unconditional variance.
- The sample analog of $Var(u_i)$ is

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{u}_i - \overline{\widehat{u}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{SSR}{n}.$$

  - Note that

$$\widehat{u}_i = \beta_0 + \beta_1 x_i + u_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i = u_i - \left(\widehat{\beta}_0 - \beta_0\right) - \left(\widehat{\beta}_1 - \beta_1\right)x_i, \tag{1}$$

  so

$$E\left[\widehat{u}_i - u_i\right] = -E\left[\widehat{\beta}_0 - \beta_0\right] - E\left[\widehat{\beta}_1 - \beta_1\right]x_i = 0.$$

  This is why we can use $\widehat{u}_i$ to substitute u_i in the genuine sample analog of $Var(u_i)$, say, $\frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})^2$.
- One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be **biased**.

Unbiased Estimator of σ^2

- An unbiased estimate of the error variance can be obtained by subtracting the number of estimated regression coefficients from the number of observations:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{SSR}{n-2}.$$

  where $n-2$ is called the degree of freedom of $\{\widehat{u}_i\}_{i=1}^n$.
- Effectively, only $n-2$ residuals are used since the other two residuals can be derived from these $n-2$ residuals by solving the two FOCs.
- Theorem 2.3 (Unbiased Estimation of σ^2): Under assumptions SLR.1-SLR.5,

$$E\left[\widehat{\sigma}^2\right] = \sigma^2.$$

(**) Proof of The Unbiasedness of $\widehat{\sigma}^2$

Proof.

For our purpose, we need to derive the formula of $\sum_{i=1}^n \widehat{u}_i^2$. First, averaging (1) we have $0 = \overline{\widehat{u}} = \overline{u} - \left(\widehat{\beta}_0 - \beta_0\right) - \left(\widehat{\beta}_1 - \beta_1\right)\overline{x}$, and subtracting this from (1) we have $\widehat{u}_i = (u_i - \overline{u}) - \left(\widehat{\beta}_1 - \beta_1\right)(x_i - \overline{x})$. Therefore,

$$\sum_{i=1}^n \widehat{u}_i^2 = \sum_{i=1}^n (u_i - \overline{u})^2 + \left(\widehat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^n (x_i - \overline{x})^2 - 2\left(\widehat{\beta}_1 - \beta_1\right) \sum_{i=1}^n (x_i - \overline{x})(u_i - \overline{u}).$$

First,

$$E\left[\sum_{i=1}^n (u_i - \overline{u})^2\right] = E\left[\sum_{i=1}^n u_i^2\right] - nE\left[\overline{u}^2\right] = n\sigma^2 - n\frac{n\sigma^2}{n^2} = (n-1)\sigma^2,$$

where $E[\overline{u}^2] = E[(\sum_{i=1}^n u_i)^2]/n^2 \overset{E[u_i u_j]=0 \text{ if } i\neq j}{=} E[\sum_{i=1}^n u_i^2]/n^2 = n\sigma^2/n^2$. Second, since $E[(\widehat{\beta}_1 - \beta_1^2)] = \sigma^2/SST_x$, the expected value of the second term is σ^2. Finally, since $\widehat{\beta}_1 - \beta_1 = \sum_{i=1}^n (x_i - \overline{x})(u_i - \overline{u})/SST_x$, the third term can be written as $-2(\widehat{\beta}_1 - \beta_1)^2 SST_x$ whose expectation is $-2\sigma^2$. Putting these three terms together gives $E[\sum_{i=1}^n \widehat{u}_i^2] = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$ so that $E[\widehat{\sigma}^2] = \sigma^2$. □

SER and SE

- $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$ is called the standard error of the regression (SER).
- The **estimated** standard deviations of the regression coefficients are called standard errors. They measure how precisely the regression coefficients are estimated:

$$\begin{aligned} se\left(\widehat{\beta}_1\right) &= \sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)} = \sqrt{\frac{\widehat{\sigma}^2}{SST_x}} = \frac{\widehat{\sigma}}{\sqrt{SST_x}}, \\ se\left(\widehat{\beta}_0\right) &= \sqrt{\widehat{Var}\left(\widehat{\beta}_0\right)} = \sqrt{\frac{\widehat{\sigma}^2 n^{-1}\sum_{i=1}^n x_i^2}{SST_x}} \\ &= \frac{\widehat{\sigma}}{\sqrt{SST_x/n^{-1}\sum_{i=1}^n x_i^2}}, \end{aligned}$$

i.e., we plug in $\widehat{\sigma}^2$ for the unknown σ^2.