

Cross-section data (multiple units at a "given point" in time)

- random sample

- ↳ best chance to learn about population of interest
- ↳ each draw is statistically "independent" of each other
- ↳ each unit in population has "same" chance of appearing in sample



"independent and identically distributed (i. i. d.)"

- ceteris paribus (is crucial to establish causality)

- ↳ If hold other relevant factors(u) fixed,
- ↳ then changes in one variable "cause" changes in another variable.
- ↳ but, it is "impossible" to truly hold all other factors fixed (especially in non-experimental data).
- ↳ however, if x is independent of other relevant factors (randomly assigned, $E(u|x) = E(u)$), then simple regression analysis gives us "good" estimate of causal effect.

Issues before "simple regression model"

1. What is functional relationship between x and y?
2. How to allow other factors other than x to affect y? (usually, there is no "exact" relationship b/w x and y)
3. How to capture ceteris paribus relationship between x and y?

$$y = \beta_0 + \beta_1 x + u \quad (\text{y, x, u are random variables } \curvearrowleft, \text{ tier 2})$$

only if $\Delta y = \beta_1 \Delta x, \Delta u = 0$

but we never observe u, so we "restrict" how u and x are related

Assumption #1

$E(u)=0$ (innocuous normalization) that can be imposed without loss of generality.
and f_0 makes the assumption $E(u)=0$ innocuous.

Assumption #2

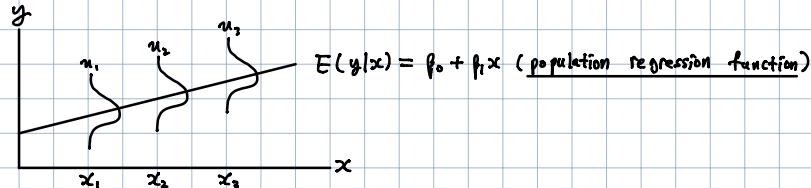
$E(u|x) = E(u)$ (u is mean independent of x) is an "almost certainly violated" assumption.

Assumption #1 + Assumption #2

If $E(u)=0$ and $E(u|x) = E(u)$, then $E(u|x)=0$.



$$\begin{aligned} E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\ &= \beta_0 + \beta_1 x + E(u|x) \\ &= \beta_0 + \beta_1 x \end{aligned}$$



Also,

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

$$\hookrightarrow E(u) = 0$$

$\hookrightarrow E(xu) = 0$ (Law of Iterated Expectation)

$$\text{Cov}(x, u) = E(xu) - E(x)E(u) = 0$$

$\hookrightarrow x, u$ are uncorrelated



$$y = \beta_0 + \beta_1 x + u \quad (\text{population})$$

$$\begin{aligned} E(u) &= E(y - \beta_0 - \beta_1 x) = 0 \\ E(xu) &= E[x(y - \beta_0 - \beta_1 x)] = 0 \end{aligned}$$

→ two equations, two unknowns (β_0, β_1)?
↳ we can't observe all x, y (population).



So we estimate β_0, β_1 from sample through "Method of Moments" instead.

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \hat{u}_i \quad (\text{prediction})$$

$$\begin{aligned} E(\hat{u}_i) &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \hat{u}_i \text{'s are overall close to 0} \\ E(x_i \hat{u}_i) &= \frac{1}{n} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{aligned}$$

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

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)}$$



$\hat{\beta}_0$ and $\hat{\beta}_1$ are called OLS estimates because,

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \text{SSR} \quad (\text{Sum of Squared Residuals} = \text{size of mistake})$$

is minimized under Method of Moments' β_0 and β_1 .



Simple Regression Model

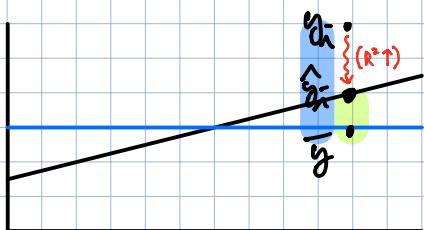
$$y = \beta_0 + \beta_1 x + u, \quad E(u|x) = 0 \quad (\text{Assumption #1, #2 definition of OLS})$$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + \hat{u}_i, \quad E(\hat{u}_i|x_i) = 0$$



"One unit increase in x (increases/decreases) predicted y by $\hat{\beta}_1$."

"Goodness-of-Fit"



$$\left. \begin{array}{l} SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \end{array} \right\} SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$R^2 = \frac{SSE}{SST} = \text{"fraction of total variation in } y_i \text{ (SST) that is explained by } x_i\text{"}$

R^2 is a prediction \rightarrow purely fit!

R^2 does not explain causal relation.

Little R^2 may also have significant causal relation.

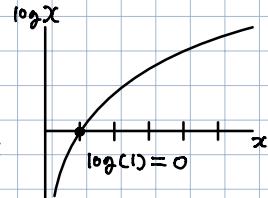
"Units of measurement"

- change in unit of x : $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$
 - change in unit of y : $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$
- R^2 does not change

$100 \times \text{Change in } \log(x)$ is equivalent to $\% \text{ change in } x$

$$100 \Delta \log(x) \approx \% \Delta x \quad \text{for } \frac{\Delta x}{x_0} \approx 0$$

$$\begin{aligned} \text{because, } 100 \cdot [\log(x_1) - \log(x_0)] &\approx 100 \times \frac{\Delta x}{x_0} \\ &= \log(x_1/x_0) \\ &= \log\left(\frac{x_0 + \Delta x - x_0}{x_0}\right) \\ &= \log\left(1 + \frac{\Delta x}{x_0}\right) \approx \frac{\Delta x}{x_0} \quad (\text{for } \frac{\Delta x}{x_0} \approx 0) \end{aligned}$$



$$\text{Thus, } \log(y) = \beta_0 + \beta_1 x + u$$

Ceteris paribus, $\beta_1 = \frac{\Delta \log y}{\Delta x}$ and ($100 \cdot \Delta \log y = \% \Delta y$)

$$\beta_1 = \frac{\% \Delta y}{100 \cdot \Delta x}$$

- $100 \times \beta_1 = \frac{\% \Delta y}{\Delta x} = \% \text{ change in } y \text{ when } x \text{ changes by one unit}$
- $\beta_1 / 100 = \frac{\Delta y}{\% \Delta x} \quad (y, \log(x))$
- $\beta_1 = \frac{\% \Delta y}{\% \Delta x} \quad (\log(y), \log(x))$
- $\beta_1 = \frac{\Delta y}{\Delta x} \quad (y, x)$

This R^2 is not directly comparable to R^2 when just $y \text{ vs } x$

Assumptions of SLR (Simple Linear Regression)

SLR.1 (Linear in parameters)

population model is " $y = \beta_0 + \beta_1 x + u$ " (β_0 and β_1 are unknown parameters) (x, u, y are r.v.)

SLR.2 (Random sampling)

We have a random sample (i.i.d.) of size n , following the population model. ($y_i = \beta_0 + \beta_1 x_i + u_i$)
 Cross-Sectional Data

SLR.3 (Sample variation in x_i)

Sample outcomes on x_i are not all the same value (very mild assumption)

SLR.4 (Zero conditional mean, or Exogeneity)

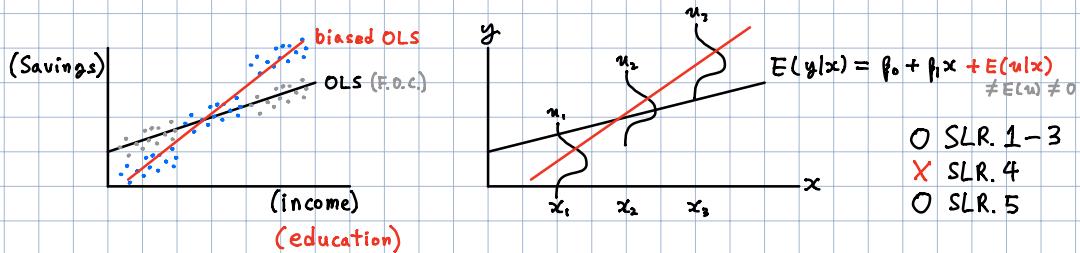
In population, error term u has zero mean given any value of regressor x
 $E(u|x) = 0$ for all $x \Rightarrow (\text{cov}(x, u) = 0)$ & (Key for showing unbiasedness of OLS)

SLR.5 (Constant Variance, or Homoskedasticity)

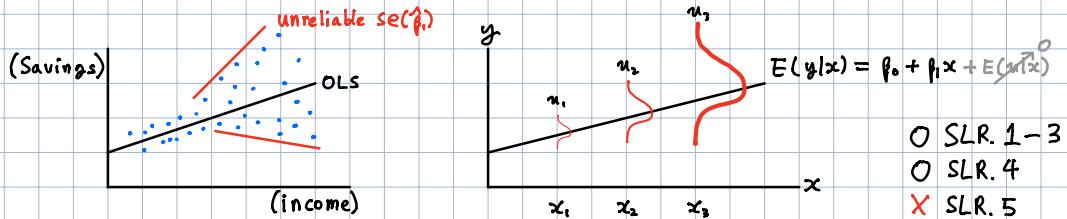
Error term has same variance given any value of regressor x (Very Strong assumption)

$\text{Var}(u|x) = \sigma^2 > 0$ for all x (where σ^2 is unknown)
 $(=c)$
 $(=E(u^2))$

Endogeneity



Heteroskedasticity



Theorem

- Unbiasedness of OLS (estimator, not estimate)

$$E(\hat{\beta}_i) = \beta_i, \text{ conditional on } X \text{ (under SLR.1-4)}$$

(Estimator, or recipe, that is used to get $\hat{\beta}_i$, is unbiased under SLR.1-4)

$$E(\hat{\beta}_i) = \beta_i + \sum_{j=1}^n w_j E(u_j) \text{ under SLR.4 [} E(u|x)=0 \text{] \& SLR.1-3.}$$

$\hat{\beta}_i$'s value was initially obtained with SLR.4 (=the OLS definition) too.

So, SLR.1-4 allows us to calculate unbiased $\hat{\beta}_i$ value (through OLS estimator)

- Sampling Variance of OLS

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{SST_x}, \text{ conditional on } X \text{ (under SLR.1-5)}$$

as σ^2 (error variance) \uparrow , $\text{Var}(\hat{\beta}_i) \uparrow$

"The more noise (u) in the relationship between y and x , the harder it is to learn about β_i "

as SST_x (sample variance of x) \uparrow , $\text{Var}(\hat{\beta}_i) \downarrow$
"More data ($n \uparrow$) shrinks sampling variance of $\hat{\beta}_i$ "

$$\text{at } \frac{1}{n} \text{ rate } (\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{SST_x} \underset{n \uparrow}{\approx} \frac{\sigma^2}{n \bar{x}^2})$$

||

$$se(\hat{\beta}_i) = \frac{\sigma}{\sqrt{SST_x}} = \text{"standard error of } \hat{\beta}_i \text{"}$$

||

But σ^2 ($= E(u^2)$) is unknown, as we never observe u .

However, $E(\hat{\sigma}^2) = \sigma^2$ (unbiased) under SLR.1-5. So,

||

$$se(\hat{\beta}_i) = \frac{\hat{\sigma}}{\sqrt{SST_x}}$$

$(\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \text{root mean squared error} = \text{Root MSE})$

Motivation for Multiple Regression (SLR.4 is violated)

- Multiple Linear Regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u, \quad E(u|x_1, \dots, x_k) = 0 \quad \text{MLR. 4}$$

($K+1$ unknown parameters in total)

β_1 measures change in y with respect to x_1 , while holding everything else (x_2, \dots, x_k, u) constant.
 \Rightarrow partial effect

- OLS regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

(slope coefficients now explicitly have *ceteris paribus* interpretation)
without having to find two different observations that differ in x_1 , but same in x_2 , thanks to OLS.

R^2 never falls when another regressor is added to regression,
because adding another x cannot increase SSR.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{SST} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

If we focus on R^2 , we might include silly variables.

Adjusted R^2 overcomes this problem,
and can be used to compare "Goodness-of-Fit" of different multiple regression models.

$$\bar{R}^2 = 1 - \frac{[SSR / (n - k - 1)]}{[SST / (n - 1)]} \quad \text{as more regressors are added (}\downarrow\text{)}$$

- Compare Simple and Multiple OLS regression lines

$$\begin{aligned} \tilde{y} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_1, & \text{If } \hat{x}_{i2} = \tilde{\delta}_0 + \tilde{\delta}_1 x_{i2} \quad (\text{OLS slope}), \\ \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, & \text{It is always true for any sample that } \tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1. \end{aligned}$$

Case 1.

$$\begin{aligned} \hat{\beta}_2 &> 0, \quad \tilde{\delta}_1 > 0 \quad (\text{positive correlation in } y \text{ vs } x_2) \\ &\quad (\text{negative correlation in } x_2 \text{ vs } x_1) \end{aligned}$$

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 \\ &= \hat{\beta}_1 + (+)(+) \end{aligned}$$

$\tilde{\beta}_1$ is over estimated, as $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$

Case 2.

$$\begin{aligned} \hat{\beta}_2 &> 0, \quad \tilde{\delta}_1 < 0 \quad (\text{positive correlation in } y \text{ vs } x_2) \\ &\quad (\text{negative correlation in } x_2 \text{ vs } x_1) \end{aligned}$$

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 \\ &= \hat{\beta}_1 + (+)(-) \end{aligned}$$

$\tilde{\beta}_1$ is under estimated, as $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2$

Assumptions of MLR (Multiple Linear Regression)

reg y α_1

| Source | SS | df | MS | Number of obs | n |
|----------|----|----|----|---------------|--|
| Model | | | | F(,) | |
| Residual | | | | Prob > F | |
| Total | | | | R-squared | $\frac{SSE}{SST}$ |
| | | | | Adj R-squared | $1 - \frac{ESSR/(n-k-1)}{SST/(n-1)}$ |
| | | | | Root MSE | $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ |

y Coef Std. Err. t p > |t| [95% C.I.]

| | | |
|------------|-----------------|---------------------|
| α_1 | $\hat{\beta}_1$ | $SE(\hat{\beta}_1)$ |
| _cons | $\hat{\beta}_0$ | $SE(\hat{\beta}_0)$ |