Cross-sectional data (multiple units observed 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.
  - ↳ however, if x is independent of other relevant factors (i.e. 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?

$\Delta y = f_1 \Delta x_1$ only if $\Delta u = 0$

$$y = \beta_0 + \beta_1 x + u$$ (y, x, u are random variables ∩ , tier 2)

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.
β_0 makes the assumption $E(u) = 0$ innocuous.

Assumption #2 $E(u|x) = 0$

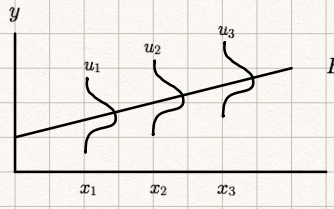
(u is **mean independent** of x), which is an "almost certainly violated" assumption.

Assumption #1 + Assumption #2

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

⇓

$$\begin{aligned} E(y|x) &= E(\beta_0 + \beta_1 x + u|x) \\ &= \beta_0 + \beta_1 x + \cancel{E(u|x)}^{\,0} \\ &= \beta_0 + \beta_1 x \end{aligned}$$



$E(y|x) = \beta_0 + \beta_1 x$ (population regression function)

Also,

$E(u|x) = 0$

↳ $E(u) = 0$

↳ $E(xu) = 0$ (Law of Iterated Expectation)

implies

$cov(x,u) = E(xu) - E(x)E(u) = 0$

↳ x, u are uncorrelated

⇓

$y = \beta_0 + \beta_1 x + u$ (population model)

$E(u) = E(y - \beta_0 - \beta_1 x) = 0$
$E(xu) = E[x(y - \beta_0 - \beta_1 x)] = 0$
→ two equations, two unknowns (β_0, β_1)?
but we can't observe all x, y (population)

⇓

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

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

$E(\hat{u}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$ = $\hat{u}_i$'s are overall close to 0

$E(x_i\hat{u}) = \frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$

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

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

⇓

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

$\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ = SSR (Sum of Squared Residuals = size of mistake)
is minimized under Method of Moments' β_0 and β_1.

⇓

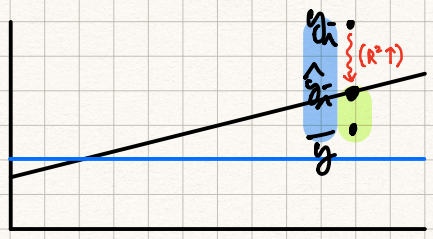
Simple Regression Model

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

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

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

"Goodness-of-Fit"



$$\left.\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ SSE &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \end{aligned}\right\} \quad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (\hat{u}_i^2)$$

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

R^2 is a prediction → 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 × Change in $\log(x)$ is equivalent to % change in x, when change in x is small

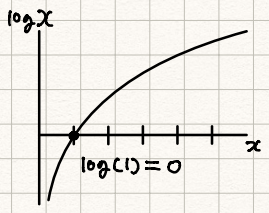
$100\,\Delta \log(x) \approx \%\Delta x$ for $\frac{\Delta x}{x_0} \approx 0$

because, $100 \cdot [\log(x_1) - \log(x_0)] \approx 100 \times \frac{\Delta x}{x_0}$

$= \log(x_1/x_0)$

$= \log\left(\frac{x_0 + x_1 - x_0}{x_0}\right)$

$= \log\left(1 + \frac{\Delta x}{x_0}\right) \approx \frac{\Delta x}{x_0}$ (for $\frac{\Delta x}{x_0} \approx 0$) as



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

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

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

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

This R^2 is not directly comparable to R^2 when just y on x

Assumptions of SLR (Simple Linear Regression)

SLR.1 (Linear in parameters)
population model is "$y = \beta_0 + \beta_1 x + u$" ($\beta_0$ 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

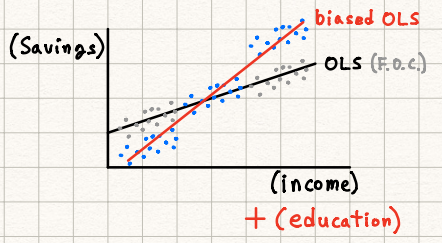
unbiased OLS estimator

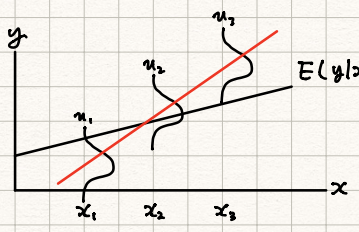
SLR.3 (Sample variation in x_i)
Sample outcomes on x_i are not all the same value (very mild assumption) — Can compute OLS estimates

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" $\Rightarrow$ ($cov(x,u) = 0$) & (Key for showing unbiasedness of OLS)

SLR.5 (Constant Variance, or Homoskedasticity)
Can estimate $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}}$
Error term has same variance given any value of regressor x (very strong assumption)
"$Var(u|x) = \sigma^2 > 0$ for all x" (where σ^2 is unknown)
($= c$) ($= E(u^2)$)

Endogeneity

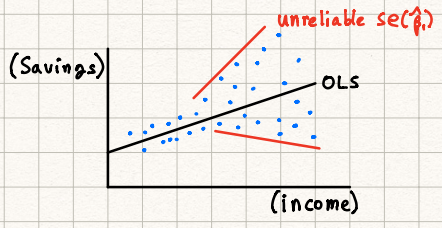


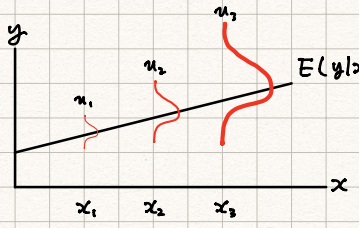


$E(y|x) = \beta_0 + \beta_1 x + E(u|x)$, $\neq E(u) \neq 0$

O SLR. 1–3
X SLR. 4
O SLR. 5

Heteroskedasticity





$E(y|x) = \beta_0 + \beta_1 x + E(u|x)$, $= 0$

O SLR. 1–3
O SLR. 4
X SLR. 5

Theorem

- Unbiasedness of OLS (estimator, not estimate)

  $E(\hat{\beta}_1) = \beta_1$, conditional on $\mathbf{X}$ (under SLR.1–4)

  (Estimator, or recipe, that is used to get $\hat{\beta}_1$ is unbiased under SLR.1–4)

  $E(\hat{\beta}_1) = \beta_1 + \sum_{i=1}^{n} w_i E(u_i)$ ($E(u_i) = 0$) under SLR.4 [$E(u|x) = 0$] & SLR.1–3.

  $\hat{\beta}_1$'s value was initially obtained with SLR.4 (= the OLS definition) too.

  So, SLR.1–4 allows us to calculate unbiased $\hat{\beta}_1$ value (through OLS estimator)

- Sampling variance of OLS

  $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$, conditional on $\mathbf{X}$ (under SLR.1–5)

  as σ^2 (error variance) ↑, $\mathrm{Var}(\hat{\beta}_1)$ ↑

  "The more noise (u) in the relationship between y and x, the harder it is to learn about β_1"

  as SST_x (sample variance of x) ↑, $\mathrm{Var}(\hat{\beta}_1)$ ↓

  "More data (n ↑) shrinks sampling variance of $\hat{\beta}_1$"

  at $\frac{1}{n}$ rate $\left(\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \approx \frac{\sigma^2}{n\sigma_x^2}, \ \sigma_x^2 = c\right)$

  ⇓

  $$se(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}} = \text{"Standard error of } \hat{\beta}_1\text{"}$$

  ⇓

  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,

  ⇓

  $$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}}$$

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

Motivation for Multiple Regression (SLR.4 is violated, now OLS is biased!)

- Multiple Linear Regression model

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

(K+1 unknown parameters in total)

β_1 measures change in y with respect to x_1, while holding everything else $(x_2, \ldots, x_k, u)$ 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 cetris paribus interpretation)
without having to find two different observations that differs 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.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{SST} \qquad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

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.

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

- Compare Simple and Multiple OLS regression lines

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2,$$

If $\hat{x}_{i2} = \tilde{\delta}_0 + \tilde{\delta}_1 x_{i1}$ (OLS slope),
It is always true for any sample that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$

Case 1.

$\hat{\beta}_2 > 0, \ \tilde{\delta}_1 > 0$ (positive correlation in $y \sim x_2$; negative correlation in $x_2 \sim x_1$)

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 = \hat{\beta}_1 + (+)(+)$$

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

Case 2.

$\hat{\beta}_2 > 0, \ \tilde{\delta}_1 < 0$ (positive correlation in $y \sim x_2$; negative correlation in $x_2 \sim x_1$)

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 = \hat{\beta}_1 + (+)(-) \ (-)$$

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

Assumptions of MLS (Multiple Linear Regression)

MLR.1 Linear in "parameters"

In population, it holds $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$

MLR.2 Random sampling from population

We have a **random sample** $\{(y_i, x_{i1}, \ldots, x_{ik}) : i = 1, \ldots, n\}$ of size n from population

MLR.3 No perfect collinearity in sample

None of regressor is **constant**, and there are **no exact linear relationships** among them

Can compute OLS

unbiased OLS estimator

MLR.4 Exogenous regressor (Zero conditional mean)

$E(u|x_1, \ldots, x_k) = E(u) = 0$ for all $(x_1, \ldots, x_k)$

Gauss – Markov Theorem

Under MLR. 1–5, OLS e- is the Best Linear Unbiased (Smallest Variance) for (BLUE)

MLR.5 Homoskedasticity (Constant variance)

Variance of u **does not change** with any of $x_1, \ldots, x_k$

$Var(u|x_1, \ldots, x_k) = Var(u) = \sigma^2$

Correct $Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}$

MLR.6 Normality of u

Error term u is **independent** of $(x_1, \ldots, x_k)$

and is **normally distributed with mean zero and variance** σ^2

$u \sim Normal(0, \sigma^2)$ i.i.d.

Hypothesis testing requires e sampling distribution of $\hat{\beta}_j$

Under MLR

$\hat{\beta}_j \sim$ al $(\beta_j, Var(\hat{\beta}_j))$

$\Rightarrow \frac{\hat{\beta}_j}{s} \sim$ Normal (0, 1) = Z

$\sim \hat{\beta}_j \sim t_{n-k-1} = t_{df}$

MLR.6 implies both **MLR.4** and **MLR.5**

MLR.6 imposes **full independence** between u and $(x_1, \ldots, x_k)$

(not just mean and variance independence)

MLR.6 imposes very **specific distributional assumption** for u

(a bell-shaped curve)

t static (or t ratio):

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \cancel{\beta_j}^{0}}{se\left(\hat{\beta}_j\right)}$$

How far $\hat{\beta}_j$ is from zero relative to its standard error.
*→ how big does t static has to be to conclude Ho is "unlikely"? (use **2** as threshold for significance as rule of thumb)*

Significance level:
It is probability of rejecting Ho when it is in fact true (Type I Error). *(1%, 5%, 10%)*

Critical value:
It is a point on the test distribution that is compared to the test statistic to determine whether to reject the null hypothesis. (2.467, 1.701, 1.313)

p-value:
It is probability of observing the statistic as extreme as we did if Ho is true.
So smaller p-values provide more evidence against null.
→ ***we can conclude that we got very rare sample or that null hypothesis is highly unlikely***
Given the observed value of t statistic, what is the smallest significance level at which we can reject Ho?
Such smallest value is known as p-value.
It uses the observed static as critical value, *and then finds significant level of the test using that critical value.*

Confidence interval (CI, interval estimates):
Loosely, CI is supposed to give "likely" range of values for corresponding population parameter.

Statements like "there is 95% chance that β_j is in interval [$\hat{\beta}_j \pm 2se\left(\hat{\beta}_j\right)$] is incorrect!
What 95% CI means is that ***for 95% of random samples that he draw from population, the interval we compute using the rule $\hat{\beta}_j \pm c \cdot se\left(\hat{\beta}_j\right)$ will include the value β_j.*** *But for a particular sample, we do not know whether β_j is in the interval.*
β_j is some fixed value, and it either ***is*** *or* ***not*** *in the interval.*

reg x

	SS	df	MS
Number of obs			n
Prob > F			
R-squared			$\frac{SSE}{SST}$
Adj R-squared			$1-\frac{[SSR/(n-k-1)]}{[SST/(n-1)]}$
Root MSE			$\hat{\sigma}=\sqrt{\hat{\sigma}^2}$

y	Coef	Std. Err.	t	$P>\|t\|$	[95% C.I]
x_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$			
_cons	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$			