Lecture 2: Making regression make sense

Lingguo Cheng

2025 Fall

Business School, Nanjing University

0. Random Sampling

- Population
- Sample
- Sampling uncertainty: *random sampling* versus *random assignment*

- Population quantity, *parameter (fixed feature of the population, does not reference sample size) vs. sample statistics (differing from sample to sample)*
  - *Expectation $E(Y_i)$ vs. sample mean*

Average or Mean

- Population average, mean, mathematical expectation, expectation

$$E(Y_i) = \int t f_Y(t) dt$$

- Sample statistics, *estimator (a function of sample data used to estimate parameters)*

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

LLN versus unbiasedness

- Two properties of the sample mean (sample statistics):
- Law of Large Numbers (LLN): the sample average gets closer and closer to the population average when N grows larger and larger (N-- > infinity)

$$\bar{Y} \xrightarrow{p} E(Y_i) \qquad \operatorname*{plim}_{N\to\infty} \bar{Y} = E(Y_i)$$

- Unbiasedness of the sample mean (given the sample size) 无偏

$$E(\bar{Y}) = E(Y_i)$$

Variability 方差

- Variance
- Population variance

$$V(Y_i) = E[(Y_i - E[Y_i])^2] \equiv \sigma_Y^2$$

- Sample variance

$$S(Y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Or

$$S(Y_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Standard deviations vs. standard errors

- Standard deviation: the square root of the population variance, σ_{Y} ;
  - The sample counterpart, *$S(Y_i)$, which is the squared root of $S(Y_i)^2$*
  - *Variance or SD is a descriptive fact about the distribution of Y_i*
  - *Also known as descriptive variance.*

- Standard error: the property of the sample statistics, the standard deviation of the sample statistics

<描述性统计> 标准差：对总体的分布离散程度的描述

标准误：样本统计量的偏离程度.

Sampling variance

- The population variance of a sample statistics, such as the sample mean

$$V(\bar{Y}) = E[(\bar{Y} - E[\bar{Y}])^2] = E[(\bar{Y} - E[Y_i])^2]$$

- Sampling variance (vs. descriptive variance): The quantity $V(\bar{Y})$ measures the variability of a sample statistic in repeated samples, as opposed to the dispersion of raw data.

- Sampling variance is related to the descriptive variance, but it is also determined by sample size (when n approaches infinity, the sample variance disappears)

$$\boxed{V(\bar{Y})} = V(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_Y^2 = \boxed{\frac{\sigma_Y^2}{n}}$$

$n \uparrow$, $V(\bar{Y}) \downarrow$

The standard error (SE)

- The standard error: the standard deviation of a statistic like the sample average; the square root of sampling variance.
    - The standard error of the sample mean:

$$\mathrm{SE}(\overline{Y}) = \frac{\sigma_Y}{\sqrt{n}}$$

    - The standard error summarizes the variability in an estimate due to random sampling

$$\hat{SE}(\overline{Y}) = \frac{S(Y_i)}{\sqrt{n}}$$

- The estimated standard error: quantifying the standard error

- In real life, the "estimated " is often omitted (source of chaos), but it is on you to have it in mind.

The t-Statistic and Central Limit Theorem

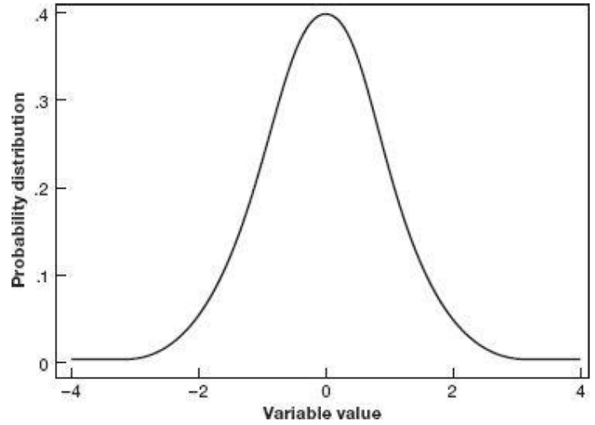
- The t-Statistic

$$t(\mu) = \frac{\overline{Y} - \mu}{\widehat{SE}(\overline{Y})}$$

$\overline{Y} = \frac{1}{N} \Sigma Y.$

Where $H_0 : E(Y_i) = \mu$

A standard normal distribution



- One miraculous statistical fact is that: $n \to \infty$
    - If the null hypothesis is true, then, *as long as the sample is large enough*, the sampling distribution of the t-statistic (*regardless of whether Y_i itself is normally distributed*) is very close to the standard normal distribution.
    - Or, the large sample distribution of a t-statistic is independent of the distribution of the underlying data used to calculate it.

r: 重复次数

$\hat{x} \to x \quad r\to\infty.$

$X \xrightarrow{d.} N(0,1) \quad n\to\infty.$

重复越多, 个体 → 个体真实分布

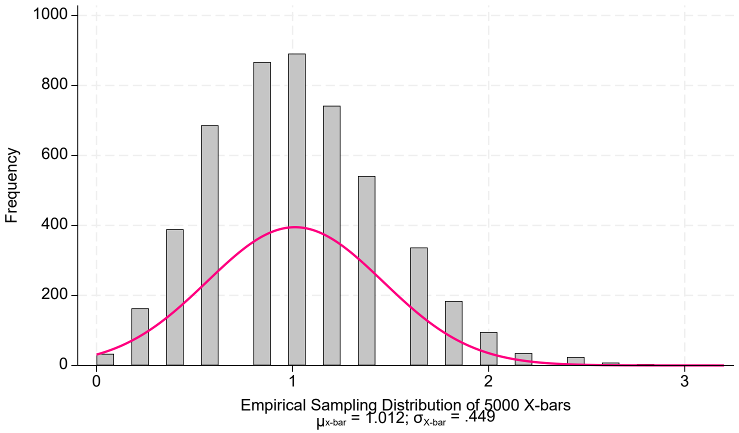
(*N*=5)

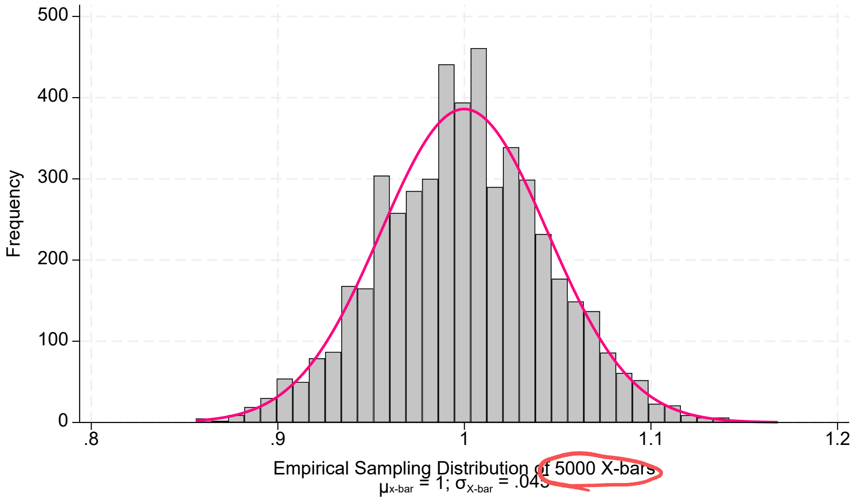
(*N*=50)

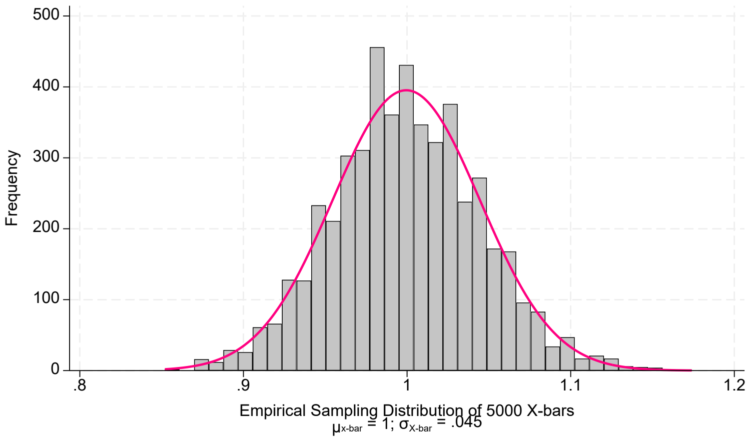
个体真实分布, 样本越多 → 正态分布.

样量 (*N*=500)

X 是 x 的分布.







5000 $\bar{X}$

重复次数 (*R*=5000)

Random Sampling Assumption (RSA) in Modern Econometrics

- A stochastic setting is necessary to give proper treatment to modern econometrics analysis.

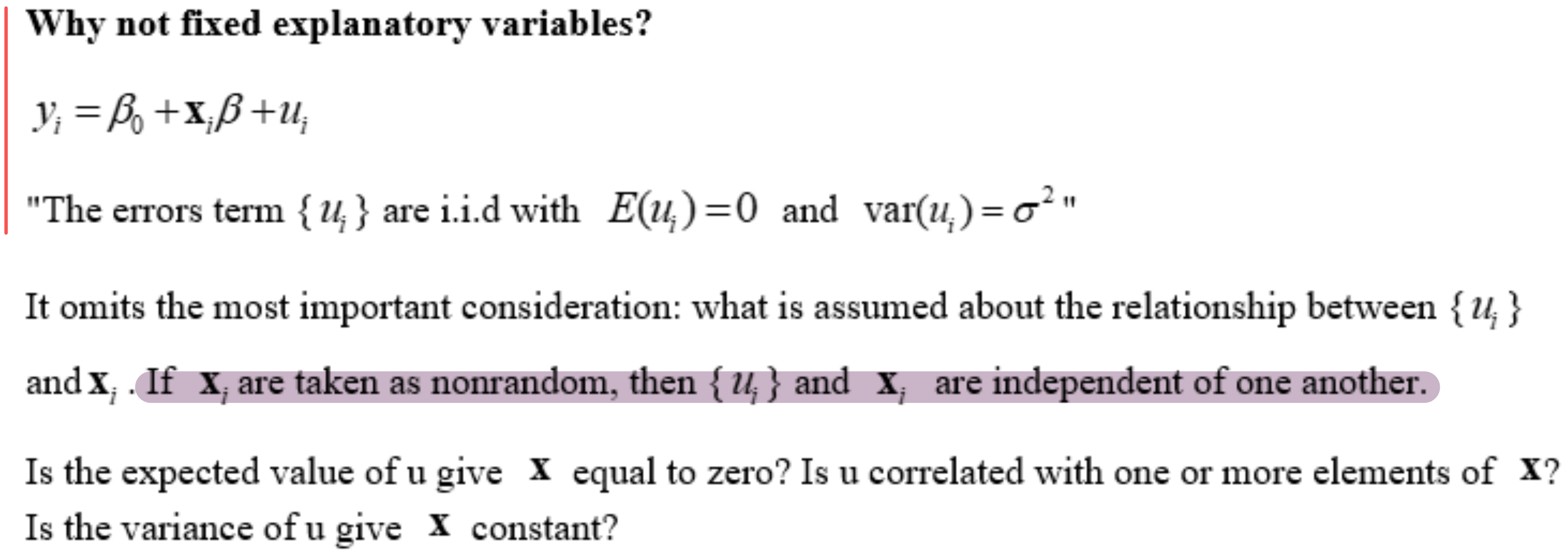
$$\{(Y_i, X_i)\}_{i=1,2,\ldots,N}$$

- Data Generating Process (DGP), random sampling
  - ○ A population model has been specified (requires us first to clearly define the population of interest)
  - ○ An independent, identically distributed sample can be drawn from the population.
- **Comments:**

○The biggest virtue of the RSA is that it allows us to separate the sampling assumptions on the population model.

$$Y_i = f(X_i) + u_i$$

- X_i are random variables drawn from population with Y_i at the same time.



Why not fixed explanatory variables?

$$y_i = \beta_0 + \mathbf{x}_i\beta + u_i$$

"The errors term $\{u_i\}$ are i.i.d with $E(u_i)=0$ and $\mathrm{var}(u_i)=\sigma^2$"

It omits the most important consideration: what is assumed about the relationship between $\{u_i\}$ and $\mathbf{x}_i$. If $\mathbf{x}_i$ are taken as nonrandom, then $\{u_i\}$ and $\mathbf{x}_i$ are independent of one another.

Is the expected value of u give $\mathbf{x}$ equal to zero? Is u correlated with one or more elements of $\mathbf{x}$?
Is the variance of u give $\mathbf{x}$ constant?

- Different from experimental data, which researcher can set values of explanatory variables and then observe values of the response variable. So experiment data fall under fixed explanatory paradigm. a special case of random sampling.

- Provide a benchmark for other non-i.i.d sampling analysis. Non-i.i.d sampling can be deal with in this framework with adding only a little difficulty.

I. Economic relationship and the conditional expectation (CEF) function

- What are we mostly interested in when investigating the economic relationship?
  - For instance, education and earnings
- The **Conditional Expectation Function (CEF)** 条件期望

- The CEF for a dependent variable Y_i, given a $K \times 1$ vector of covariates X_i (with elements x_{ki}), is the expectation, or population average, of Y_i , with X_i held fixed.

- The CEF is written $E[Y_i|X_i]$ and is a function of X_i. Because X_i is random, the CEF is random.

- For a specific value of X_i, say $X_i = x$, we write $E[Y_i|X_i = x]$.
  - $E[Y_i|X_i = 42]$
  - $E[Y_i|D_i = 1]$, $E[Y_i|D_i = 0]$

What is CEF?

- For continuous Y_i with conditional density $f_y(t|X_i = x)$ at Y_i = t, the CEF is

$$E(Y_i \mid X_i = x) = \int t f_{\mathrm{y}}(t \mid X_i = x)dt$$

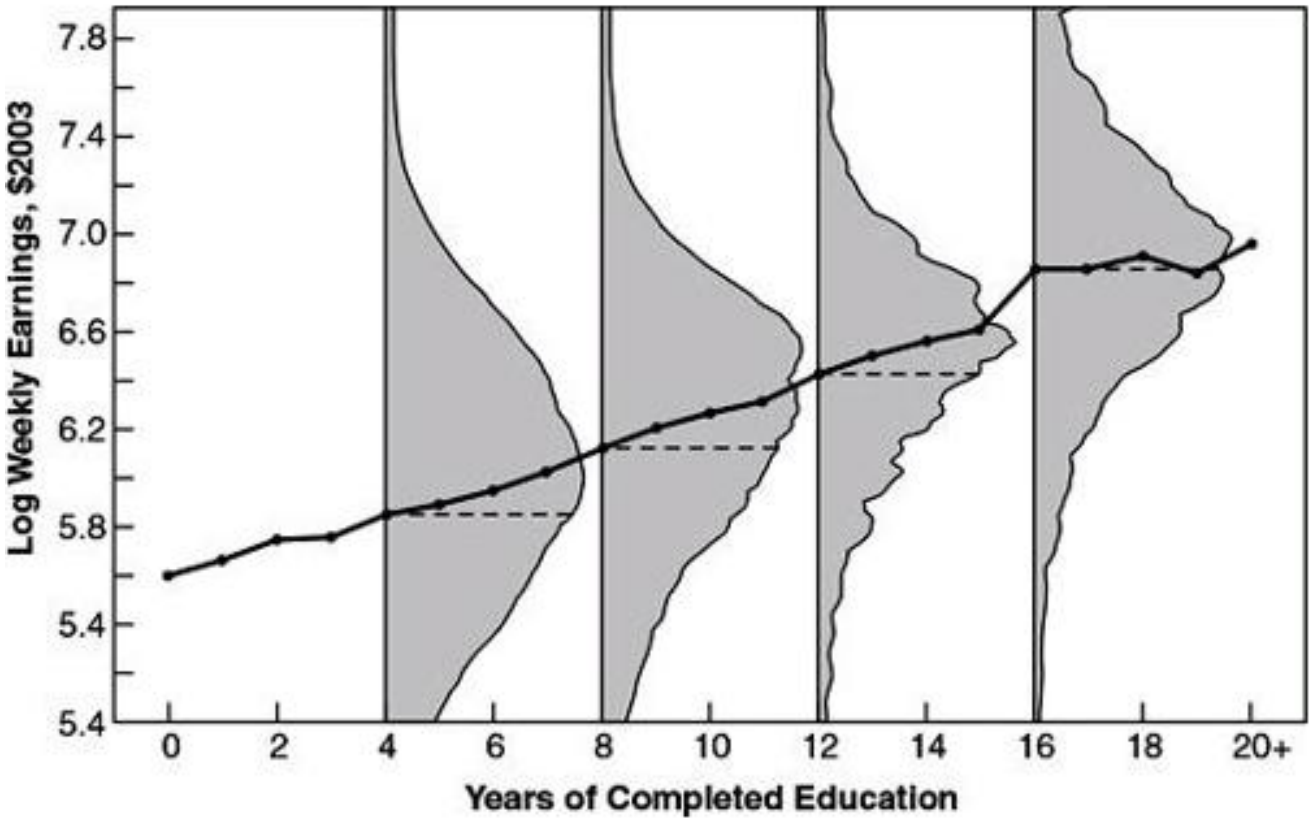
- If Y_i is discrete, $E[Y_i|X_i = x]$ equals

$$E(Y_i \mid X_i = x) = \sum_t tP(Y_i = t \mid X_i = x)$$

where $P(Y_i = t|X_i = x)$ is the conditional probability mass function for Y_i given $X_i = x$.

- Expectation is a population concept. In practice, data usually comes in the form of samples and rarely consists of an entire population. We, therefore, use samples to make inferences about the population....*Statistical inference*

- Our *"population-first" approach* to econometrics is motivated by the fact that we must define the objects of interest before we can use data to study them.



$E(Y_i|X_i)$

Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

- Notwithstanding the enormous variation individual circumstances, people with more schooling generally earn more.
- The average earnings gain associated with a year of schooling is typically about 10 percent.

Property of CEF

P1: Law of Iterated Expectations (LIE)

$$E[Y_i] = E\{E[Y_i|X_i]\}.$$

Proof:

$$\begin{aligned}
E\{E[Y_i|X_i]\} &= \int E[Y_i|X_i = u] g_x(u)du \\
&= \int \left[\int tf_y(t|X_i = u)dt\right] g_x(u)du \\
&= \int\int tf_y(t|X_i = u)g_x(u)dudt \\
&= \int t\left[\int f_y(t|X_i = u)g_x(u)du\right]dt \\
&= \int t\left[\int f_{xy}(u,t)du\right]dt \\
&= \int tg_y(t)dt = E[Y_i].
\end{aligned}$$

交换积分次序

LIE version 2:

$$E(y \mid \underbrace{\mathbf{x}}_{\text{小信息集占主导}}) = E[E(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}] = E[E(y \mid x) \mid x, z]$$

Proof:

$$\begin{aligned}
E[E(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}] &= \int E(y \mid \mathbf{x}, \mathbf{z}) f(\mathbf{z} \mid \mathbf{x}) dz \\
&= \int [\int y f(y \mid \mathbf{x}, \mathbf{z}) dy] f(\mathbf{z} \mid \mathbf{x}) dz \\
&= \int y dy \int f(y \mid \mathbf{x}, \mathbf{z}) f(\mathbf{z} \mid \mathbf{x}) dz \\
&= \int y f(y \mid \mathbf{x}) dy = E(y \mid \mathbf{x})
\end{aligned}$$

The smaller information set always dominates

LIE version 3:

w has more information than x.

x=f(w), the most general form of LIE can be stated as

$$E(y\,|\,\mathbf{x}) = E[E(y\,|\,\underline{\mathbf{w}})\,|\,\mathbf{x}]$$

Or

$$E(y\,|\,\mathbf{x}) = E[E(y\,|\,\mathbf{x})\,|\,\underline{\mathbf{w}}]$$

Suppose for some vector function $\mathbf{f}(\mathbf{x})$ and a real-valued function g(.), $\underline{E(y\,|\,\mathbf{x}) = g[\mathbf{f}(\mathbf{x})]}$.

Then

$$\underline{E(y\,|\,\mathbf{f}(\mathbf{x})) = E(y\,|\,\mathbf{x})} \overset{\text{代入}}{=} g[\mathbf{f}(\mathbf{x})]$$

Proof:

$$\underline{E[y\,|\,\mathbf{f}(\mathbf{x})]} = E\{E[y\,|\,\mathbf{x}]\,|\,\mathbf{f}(\mathbf{x})\} = E\{g[\mathbf{f}(\mathbf{x})]\,|\,\mathbf{f}(\mathbf{x})\} = g[\mathbf{f}(\mathbf{x})] = \underline{E(y\,|\,\mathbf{x})}$$

e.g. If wage function is

$$E[wage\,|\,educ, \text{exper}] = \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 educ \cdot \text{exper}$$

then

$$E[wage\,|\,educ, \text{exper}, \text{exper}^2, educ \cdot \text{exper}]$$
$$= \beta_0 + \beta_1 educ + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 educ \cdot \text{exper}$$

A more general form

$$E(y\,|\,\mathbf{x}) = \beta_0 + \beta_1 g_1(\mathbf{x}) + \beta_2 g_2(\mathbf{x}) + \ldots + \beta_M g_M(\mathbf{x})$$

then

$$E(y\,|\,z_1, z_2, \ldots, z_M) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_M z_M, \quad z_j = g_j(\mathbf{x})$$

(分解性质.

P2: The CEF Decomposition Property

$$Y_i = E[Y_i|X_i] + \varepsilon_i,$$

error term $\varepsilon_i = Y_i - E(Y_i | X_i)$

where (i) ε_i is mean independent of X_i, that is, $E[\varepsilon_i|X_i] = 0$, and therefore (ii) ε_i is uncorrelated with any function of X_i.

Proof:

(1) $E(\overset{\varepsilon_i}{\cancel{u}} \mid \mathbf{x}) = E(y - E(y \mid \mathbf{x}) \mid \mathbf{x}) = E(y \mid \mathbf{x}) - E(y \mid \mathbf{x}) = 0$

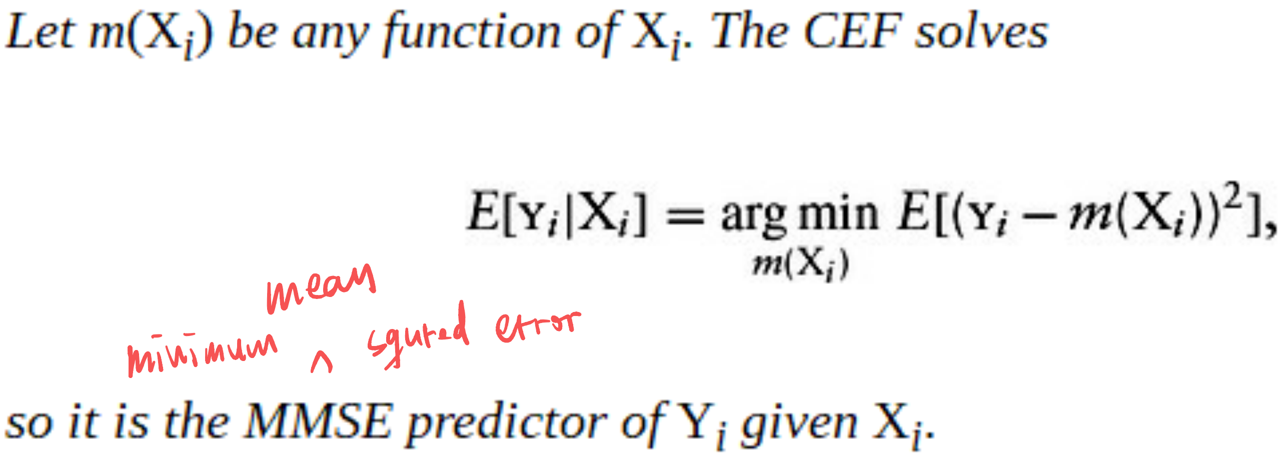
(2) Let h ($\mathbf{x}$) be any function of $\mathbf{x}$. By the law of iterated expectations,

$$E(h(\mathbf{x})u) = E[E(h(\mathbf{x})u \mid \mathbf{x})] = E[h(\mathbf{x})E(u \mid \mathbf{x})] = E[h(\mathbf{x}) * 0] = 0$$

$\mathrm{cov}(\varepsilon_i, h(u)) = 0.$

$= [\underset{0}{\varepsilon_i - E(\varepsilon_i)}][h(u) - E(h(u))] \underset{=}{=} \varepsilon_i h(u) - E(h(u)\varepsilon_i)$

Property 3: The CEF Prediction Property



Let $m(X_i)$ be any function of X_i. The CEF solves

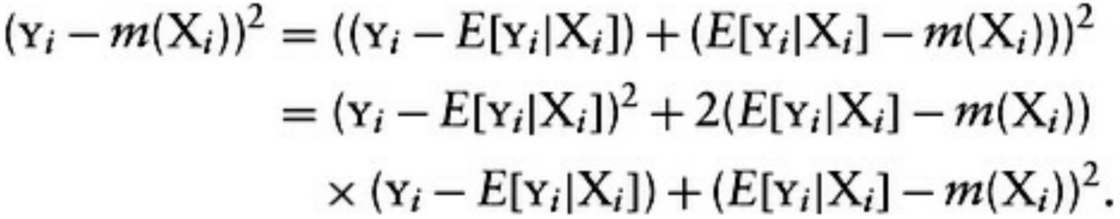
$$E[Y_i|X_i] = \underset{m(X_i)}{\arg\min}\; E[(Y_i - m(X_i))^2],$$

mean

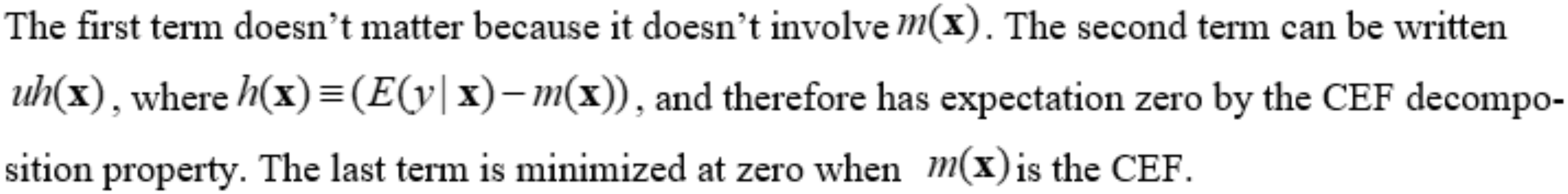
minimum ^ squred error

so it is the MMSE predictor of Y_i given X_i.

Proof: Write



$$\begin{aligned}(Y_i - m(X_i))^2 &= ((Y_i - E[Y_i|X_i]) + (E[Y_i|X_i] - m(X_i)))^2 \\ &= (Y_i - E[Y_i|X_i])^2 + 2(E[Y_i|X_i] - m(X_i)) \\ &\quad \times (Y_i - E[Y_i|X_i]) + (E[Y_i|X_i] - m(X_i))^2.\end{aligned}$$



The first term doesn't matter because it doesn't involve $m(\mathbf{x})$. The second term can be written $uh(\mathbf{x})$, where $h(\mathbf{x}) \equiv (E(y\,|\,\mathbf{x}) - m(\mathbf{x}))$, and therefore has expectation zero by the CEF decomposition property. The last term is minimized at zero when $m(\mathbf{x})$ is the CEF.

Property 4: The ANOVA Theorem

$$Var(Y_i) = Var(E[Y_i|X_i]) + E[Var(Y_i|X_i)].$$

$$V(Y_i) = V(E[Y_i|X_i]) + E[V(Y_i|X_i)],$$

where $V(\cdot)$ *denotes variance and* $V(Y_i|X_i)$ *is the conditional variance of* Y_i *given* X_i.

Proof:

$$E[\varepsilon_i^2] = E[E[\varepsilon_i^2|X_i]] = E[V[Y_i|X_i]],$$

where $E[\varepsilon_i^2|X_i] = V[Y_i|X_i]$ because $\varepsilon_i \equiv Y_i - E[Y_i|X_i]$.

$$Y_i = E(Y_i|X_i) + \varepsilon_i$$

$$cov(\cdots, \cdots) = 0.$$

$$Var(Y_i) = Var[E(Y_i|X_i) + \varepsilon_i]$$

$$= \underline{Var[E(Y_i|X_i)]} + Var[\varepsilon_i]$$

$$[\varepsilon_i - \underbrace{E(\varepsilon_i)}_{0}]^2$$

$$Var(Y_i) = Var[E(Y_i|X_i)] + E(\varepsilon_i^2)$$

II. Linear Regression

- Regression estimates provide a valuable baseline for almost all empirical research because regression is tightly linked to the CEF, and the CEF provides a natural summary of empirical relationships.

- Let the *K*×*1* regression coefficient vector β be defined by solving

$$\beta = \arg\min_{b} E[(Y_i - \underset{\sim}{\overset{1\times K}{X_i'}}\,\underset{\sim}{\overset{K\times 1}{b}})^2].$$

$$x_i = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$$

$$x_i' b = (x_{1i}b_1 + x_{2i}b_2 + \cdots + x_{ni}b_n)$$

$$\frac{\partial x_i' b}{\partial b} = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \cdots \\ x_{ni} \end{pmatrix} x_i' b = x_i x_i' b.$$

- Using the first-order condition,

$$\text{FOC:} \quad 2E[X_i(Y_i - X_i'b)] = 0.$$

⇨

$$E[X_i(Y_i - X_i'b)] = 0,$$

$$E[X_iY_i] = E[X_iX_i']\beta.$$

$$\beta = E[X_iX_i']^{-1}\ E[X_iY_i].$$

$$\beta = E[X_iX_i']^{-1}\ E[X_iY_i].$$

CEF and Linear projection

Let y, $x_1, \ldots, x_K$ be random variables representing some population such that $E(y^2) < \infty$, $E(x_j^2) < \infty$, j=1,2,...,K.

Definition 1

Define $\mathbf{x} = (x_1, \ldots, x_K)$ as a 1*K vector, and assume that the K*K variance matrix of x is nonsingular (positive definite). Then the linear projection of y on 1, $x_1, \ldots, x_K$ always exists and is unique:

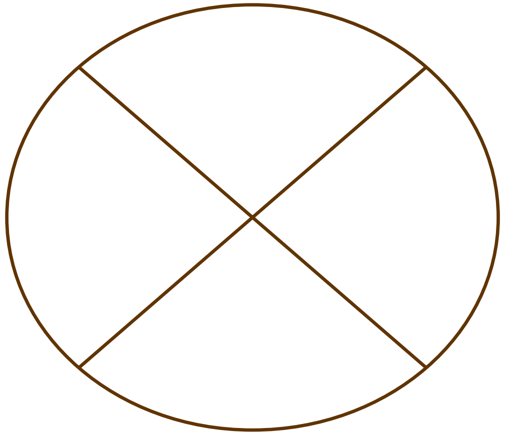
$$\begin{aligned} L(y \mid 1, \mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K \\ &= \beta_0 + \mathbf{x}\boldsymbol{\beta} \end{aligned}$$

麻烦

Where, by definition,

$$\beta \equiv [\mathrm{var}(\mathbf{x})]^{-1} \mathrm{cov}(\mathbf{x}, y)$$

$$\beta_0 \equiv E(y) - E(\mathbf{x})\boldsymbol{\beta}$$



Definition 2

If we include unity as an element of **x, the linear projection can be written as**

$$
\begin{aligned}
L(y \mid \mathbf{x}) &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_K x_K \\
&= \mathbf{x}\boldsymbol{\beta}
\end{aligned}
$$

Where, by definition,

$$\beta \equiv [E(\mathbf{x}'\mathbf{x})]^{-1} E(\mathbf{x}'y)$$

$x = (1, x_2, x_3, \ldots, x_K)$ $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_K \end{pmatrix}$

**************Appendix 1: about notation **************

x, (1*1); , **x**, vector, K*1; X, matrix, N*K
y, (1*1); **y**, vector, K*1; Y, matrix, K*1

- About the vector of the independent variables on the right hand of regression equation:

(1) Regarded as a row vector or column vector
(2) Include constant or not

$$y_i = \mathbf{x}_i \beta + \varepsilon_i \quad (i = 1, 2, 3, \ldots, N)$$

$$\Downarrow$$

$$y_1 = \mathbf{x}_1 \beta + \varepsilon_1$$

$$y_2 = \mathbf{x}_2 \beta + \varepsilon_2$$

$$\vdots$$

$$y_N = \mathbf{x}_N \beta + \varepsilon_N$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\Downarrow$$

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

When x is a row vector

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

$$\mathbf{x}'_i\, y_i = \mathbf{x}'_i\, \mathbf{x}_i \boldsymbol{\beta} + \mathbf{x}'_i\, \varepsilon_i$$

$$\mathrm{E}(\mathbf{x}'_i\, y_i) = \mathrm{E}(\mathbf{x}'_i\, \mathbf{x}_i)\boldsymbol{\beta} + \mathrm{E}(\mathbf{x}'_i\, \varepsilon_i)$$

$$\boldsymbol{\beta} = \mathrm{E}(\mathbf{x}'_i\, \mathbf{x}_i)^{-1}\, \mathrm{E}(\mathbf{x}'_i\, y_i)$$

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$= \left(\begin{bmatrix} \mathbf{x}'_1 & \mathbf{x}'_2 & \cdots & \mathbf{x}'_N \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{x}'_1 & \mathbf{x}'_2 & \cdots & \mathbf{x}'_N \end{bmatrix} \mathbf{y}$$

$$= \left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}'_{\mathrm{i}} \mathbf{x}_{\mathrm{i}}\right)^{-1} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbf{x}'_{\mathrm{i}} \mathrm{y}_{\mathrm{i}}$$

↓转置一下.

$$y_i = \mathbf{x}'_i \beta + \varepsilon_i \quad (i = 1,2,3,\ldots,N)$$

$$\Downarrow$$

$$y_1 = \mathbf{x}'_1 \beta + \varepsilon_1$$

$$y_2 = \mathbf{x}'_2 \beta + \varepsilon_2$$

$$\vdots$$

$$y_N = \mathbf{x}'_N \beta + \varepsilon_N$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{x}'_1} \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_N \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

$$\Downarrow$$

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

$$\Rightarrow$$

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$$

$$\mathbf{x}_i y_i = \mathbf{x}_i \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{x}_i \varepsilon_i$$

利用E消除 ε_i.

$$\mathrm{E}(\mathbf{x}_i y_i) = \mathrm{E}(\mathbf{x}_i \mathbf{x}'_i)\boldsymbol{\beta} + \mathrm{E}(\mathbf{x}_i \varepsilon_i)$$

$$\boldsymbol{\beta} = \mathrm{E}(\mathbf{x}_i \mathbf{x}'_i)^{-1} \mathrm{E}(\mathbf{x}_i y_i)$$

$$\Downarrow$$

When x is a column vector

$$\Rightarrow$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

$$= \left(\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix} \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_N \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix} \mathbf{y}$$

$$= \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}'_i \right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$$

向量层面和矩阵层面

相似

- Note that by construction, $E[X_i(Y_i - X_i'\beta)] = 0$. In other words, the population residual, which we define as $Y_i - X_i'\beta = e_i$, is uncorrelated with the regressors, X_i.
- In the simple bivariate case where the regression vector includes only the single regressor, x_i, and a constant, the slope coefficient is

$$\beta_1 = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$

and the intercept is $\alpha = E[Y_i] - \beta_1 E[X_i]$

(Can you prove it by yourself?)

aggrate

Proof: here, I will give you a matrix derivation

$$\boldsymbol{\beta} = E(X_i X_{\mathrm{i}}^{'})^{-1} E(X_i Y_i)$$

$$= E(\begin{bmatrix} 1 \\ \mathrm{x}_{1\mathrm{i}} \end{bmatrix} \begin{bmatrix} 1 & \mathrm{x}_{1\mathrm{i}} \end{bmatrix})^{-1} \mathrm{E}(\begin{bmatrix} 1 \\ \mathrm{x}_{1\mathrm{i}} \end{bmatrix} Y_i)$$

$$= E(\begin{bmatrix} 1 & \mathrm{x}_{1\mathrm{i}} \\ \mathrm{x}_{1\mathrm{i}} & \mathrm{x}_{1\mathrm{i}}\mathrm{x}_{1\mathrm{i}} \end{bmatrix})^{-1} \mathrm{E}(\begin{bmatrix} Y_i \\ \mathrm{x}_{1\mathrm{i}} Y_i \end{bmatrix})$$

$$= \begin{bmatrix} 1 & E(\mathrm{x}_{1\mathrm{i}}) \\ E(\mathrm{x}_{1\mathrm{i}}) & E(\mathrm{x}_{1\mathrm{i}}\mathrm{x}_{1\mathrm{i}}) \end{bmatrix})^{-1} \begin{bmatrix} \mathrm{E}(Y_i) \\ \mathrm{E}(\mathrm{x}_{1\mathrm{i}} Y_i) \end{bmatrix}$$

$$= \frac{1}{E(\mathrm{x}_{1\mathrm{i}}\mathrm{x}_{1\mathrm{i}}) - E(\mathrm{x}_{1\mathrm{i}})\mathrm{E}(\mathrm{x}_{1\mathrm{i}})} \begin{bmatrix} E(\mathrm{x}_{1\mathrm{i}}\mathrm{x}_{1\mathrm{i}}) & -E(\mathrm{x}_{1\mathrm{i}}) \\ -E(\mathrm{x}_{1\mathrm{i}}) & 1 \end{bmatrix} \bullet \begin{bmatrix} \mathrm{E}(Y_i) \\ \mathrm{E}(\mathrm{x}_{1\mathrm{i}} Y_i) \end{bmatrix}$$

$$\beta_1 = \frac{\mathrm{E}(\mathrm{x}_{1\mathrm{i}} Y_i) - \mathrm{E}(Y_i) E(\mathrm{x}_{1\mathrm{i}})}{E(\mathrm{x}_{1\mathrm{i}}\mathrm{x}_{1\mathrm{i}}) - E(\mathrm{x}_{1\mathrm{i}})\mathrm{E}(\mathrm{x}_{1\mathrm{i}})} = \frac{\mathrm{cov}(Y_i, x_i)}{\mathrm{var}(x_i)}$$

- Regression Anatomy
- In the multivariate case, with more than one non-constant regressor, the slope coefficient for the kth regressor is given below:

$$\boxed{\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})},}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \frac{cov(Y_i, \tilde{x}_{1i})}{var(\tilde{x}_{1i})} \\ \cdots \\ \frac{cov(\cdots)}{var(\cdots)} \end{pmatrix}$$

Where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates.

- In other words, $E[X_iX_i']^{-1}E[X_iY_i]$ is the $K \times 1$ vector with kth element $\frac{\mathrm{cov}(Y_i, \tilde{x}_{ki})}{\mathrm{var}(\tilde{x}_{ki})}$.
- This important formula is said to describe the anatomy of a multivariate regression coefficient because it reveals much more than the matrix formula $\beta = E[X_iX_i']^{-1}E[X_iY_i]$. It shows us that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialling out all the other covariates.

Proof:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_K x_{Ki} + e_i$$

$$x_{ki} = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + \ldots + \gamma_{k\text{-}1} x_{k\text{-}1i} + \gamma_{k+1} x_{k+1i} + \ldots + \gamma_K x_{Ki} + f_i$$

$$\beta_k = \frac{\operatorname{cov}(Y_i, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})}$$

$$= \frac{\operatorname{cov}(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})}$$

$$= \frac{\operatorname{cov}(\beta_k x_{ki}, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})} = \frac{\beta_k \operatorname{cov}(x_{ki}, \tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})} = \frac{\beta_k \operatorname{var}(\tilde{x}_{ki})}{\operatorname{var}(\tilde{x}_{ki})}$$

- The regression anatomy formula is usually attributed to Frisch and Waugh (1933). You can also do regression anatomy this way:

$$\beta_k = \frac{Cov(\tilde{Y}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})},$$

$$\beta_k = \frac{cov(Y_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})}$$

reg Y_{ki} on $x_1, x_2, \cdots, x_{k-1}, x_{k+1}, \cdots, x_K$,

Where $\tilde{Y}_{ki}$ is the residual from a regression of Y_{ki} on every covariate except for x_{ki}.

- Proof: the fitted value removed from Y_{ki} are uncorrelated $\tilde{x}_{ki}$.
- Comments: often it is very useful to plot $\tilde{Y}_{ki}$ against $\tilde{x}_{ki}$. The slope of the least squares fit in this scatterplot is the multivariate β_k , even though the plot is two-dimensional.
- Note that:

$$\beta_k \neq \frac{\text{cov}(\tilde{Y}_i, x_k)}{\text{var}(x_k)} = \frac{\text{cov}(\tilde{Y}_i, \tilde{x}_k)}{\text{var}(\tilde{x}_{ki})} \frac{\text{var}(\tilde{x}_{ki})}{\text{var}(x_k)}$$

Unless x_{ki} is uncorrelated with other covariates.

$$\beta_k = \frac{cov(\tilde{Y}_{ki}, \tilde{x}_{ki})}{var(\tilde{x}_{ki})}$$

$$= \frac{cov(Y_{ki} - L(Y_{ki} | x_{-k}), \tilde{x}_{ki})}{var(\tilde{x}_{ki})}$$

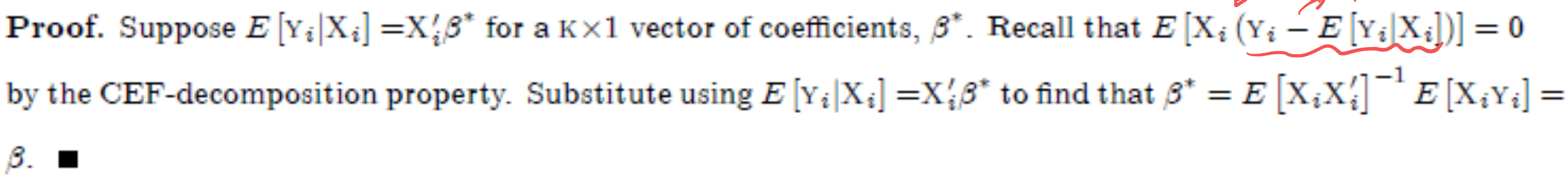
$$= \frac{cov(Y_{ki}, \tilde{x}_{ki})}{var(\tilde{X}_{ki})}$$

III. Linear Regression vs. CEF

- **The Linear CEF Theorem (Regression Justification I):**
  - Suppose the CEF is linear. Then the population regression function is it.

Proof:

$Y_i = E(Y_i|X_i) + \varepsilon_i$

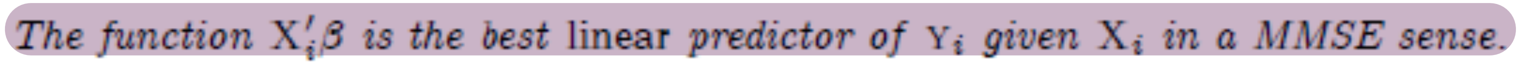


Proof. Suppose $E[Y_i|X_i] = X_i'\beta^*$ for a $K \times 1$ vector of coefficients, β^*. Recall that $E[X_i(\overbrace{Y_i - E[Y_i|X_i]}^{\varepsilon_i})] = 0$ by the CEF-decomposition property. Substitute using $E[Y_i|X_i] = X_i'\beta^*$ to find that $\beta^* = E[X_iX_i']^{-1}E[X_iY_i] = \beta$. ■

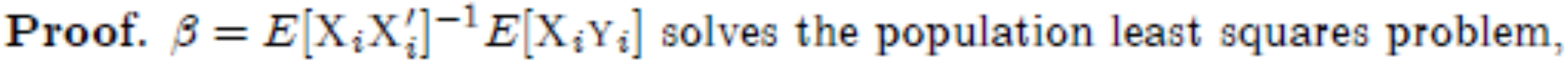
$E[X_i(Y_i - X_i'\beta^*)] = 0,$

$\beta^* = E[X_iX_i']^{-1}E[X_iY_i] = \beta.$

The Best Linear Predictor Theorem (Regression Justification II).



The function $X_i'\beta$ *is the best* linear *predictor of* Y_i *given* X_i *in a MMSE sense.*



Proof. $\beta = E[X_iX_i']^{-1}E[X_iY_i]$ solves the population least squares problem,

- $E[Y_i|X_i]$, is the best (i.e., MMSE) predictor of Y_i given X_i in the class of all functions of X_i , the population regression function is the best we can do in the class of linear functions.

The Regression CEF Theorem (Regression Justification III)

- The function *X'$_i$β* provides the MMSE linear approximation to *E[Y$_i$|X$_i$]*, that is,

$$\beta = \arg\min_{b} E\{(E[Y_i|X_i] - X_i'b)^2\}.$$

○ Proof:

$$\begin{aligned}(Y_i - X_i'b)^2 &= \{(Y_i - E[Y_i|X_i]) + (E[Y_i|X_i] - X_i'b)\}^2 \\ &= (Y_i - E[Y_i|X_i])^2 + (E[Y_i|X_i] - X_i'b)^2 \\ &\quad + 2(Y_i - E[Y_i|X_i])(E[Y_i|X_i] - X_i'b).\end{aligned}$$

- It is the favorite way to motivate regression.
- The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships without necessarily trying to pin them down exactly.

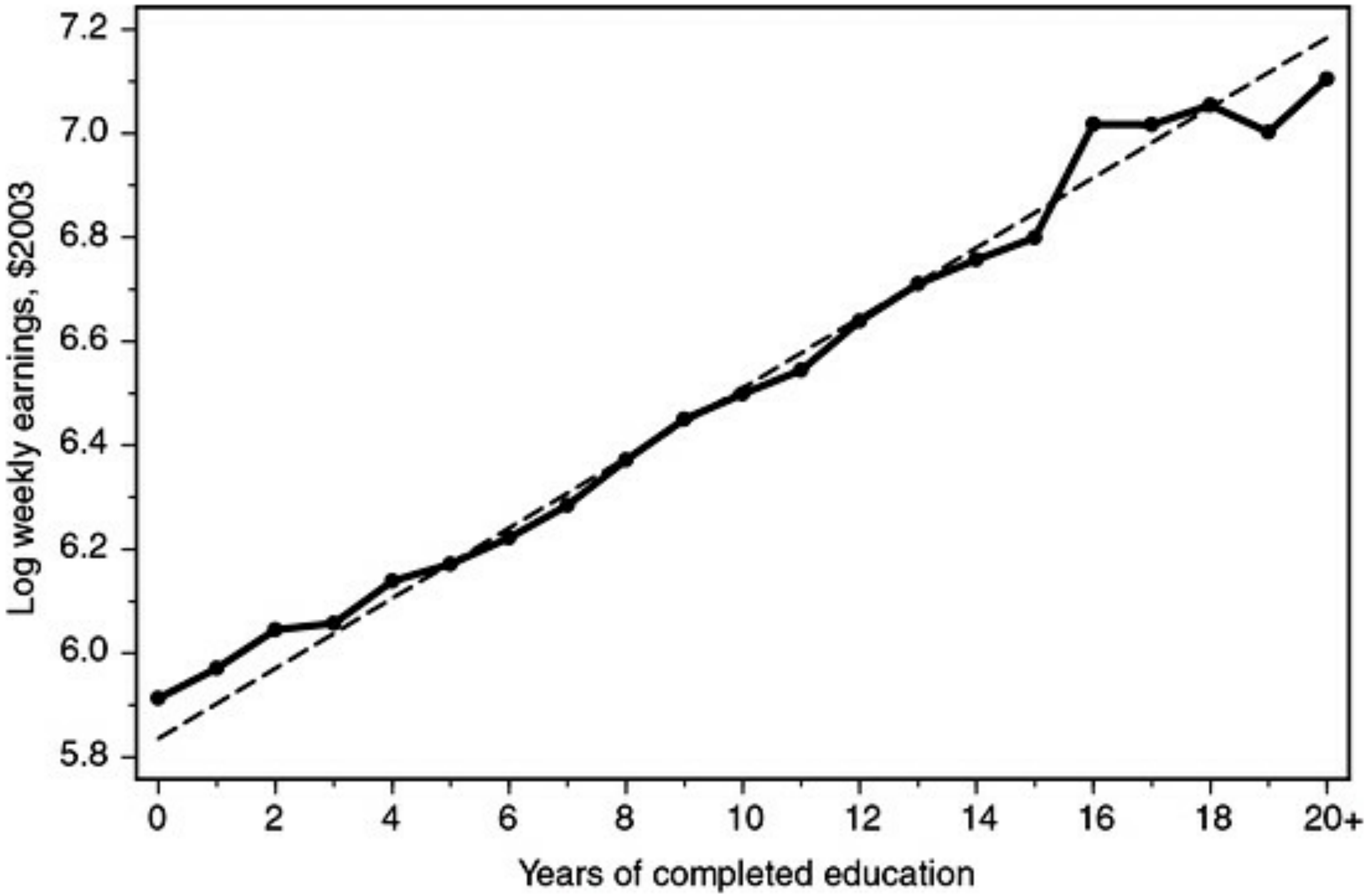
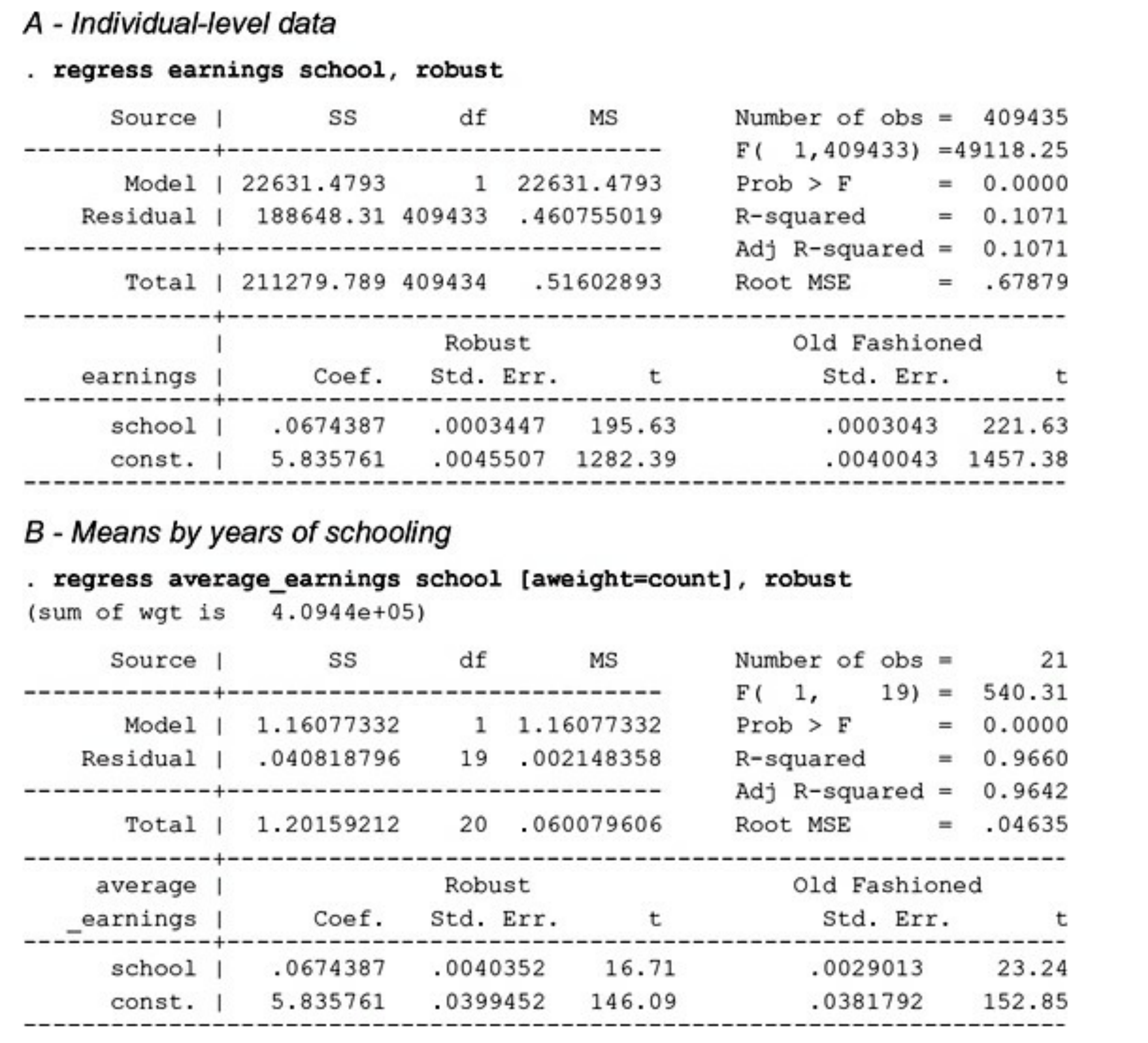


Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

- An implication of the regression CEF theorem is that regression coefficients can be obtained by using E[Y_i|X_i] as a dependent variable instead of Y_i itself.
- An even simpler way to see this is to iterate expectations in the formula for β

$$\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i] = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_iE(\mathrm{Y}_i|\mathrm{X}_i)].$$

- The CEF or grouped data version of the regression formula is of practical use when working on a project that precludes the analysis of microdata.
- Angrist (1998) used grouped data to study the effect of voluntary military service on earnings later in life.
- The earnings data come from the U.S. Social Security system, but Social Security earnings records cannot be released to the public. Instead of individual earnings, Angrist worked with average earnings conditional on race, sex, test scores, education, and veteran status.



A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS              Number of obs =  409435
-------------+------------------------------           F(  1,409433) =49118.25
       Model |  22631.4793     1  22631.4793           Prob > F      =  0.0000
    Residual |   188648.31 409433  .460755019           R-squared     =  0.1071
-------------+------------------------------           Adj R-squared =  0.1071
       Total |  211279.789 409434   .51602893           Root MSE      =  .67879

-------------+----------------------------------------------------------------
             |               Robust                       Old Fashioned
    earnings |      Coef.   Std. Err.      t               Std. Err.      t
-------------+----------------------------------------------------------------
      school |    .0674387   .0003447    195.63             .0003043    221.63
      const. |    5.835761   .0045507   1282.39             .0040043   1457.38
------------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,     19) =  540.31
       Model |  1.16077332     1  1.16077332           Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358           R-squared     =  0.9660
-------------+------------------------------           Adj R-squared =  0.9642
       Total |  1.20159212    20  .060079606           Root MSE      =  .04635

-------------+----------------------------------------------------------------
     average |               Robust                       Old Fashioned
   _earnings |      Coef.   Std. Err.      t               Std. Err.      t
-------------+----------------------------------------------------------------
      school |    .0674387   .0040352     16.71            .0029013      23.24
      const. |    5.835761   .0399452    146.09            .0381792     152.85
------------------------------------------------------------------------------
```

IV. The asymptotic OLS statistical Inference

- In practice, we don’t usually know what the CEF or the population regression vector is. We therefore draw statistical inferences about these quantities using samples.

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i]$$

- The method of moments estimator of β

$$\hat{\beta} = \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i Y_i .$$

Method of moment estimator

Suppose the vector $W_i \equiv (Y_i, X_i')'$ is independently and identically distributed in a sample of size N.

A natural estimator of the first-order population moment $E(W_i)$ is

$$\frac{1}{N}\sum_{i=1}^{N} W_i$$

Similarly, higher-order moments of the elements of W_i

$$E(W_i W_i')$$

The method of moments estimator will be

$$\frac{1}{N}\sum_{i=1}^{N} W_i W_i'$$

By the law of large numbers

$$\frac{1}{N}\sum_{i=1}^{N} W_i \xrightarrow{p} E(W_i)$$

$$\frac{1}{N}\sum_{i=1}^{N} W_i W_i' \xrightarrow{p} E(W_i W_i')$$

- We have known the (population) regression coefficient

$$\beta = E(X_i X_i')^{-1} E(X_i Y_i)$$

- The method of moment estimator of β

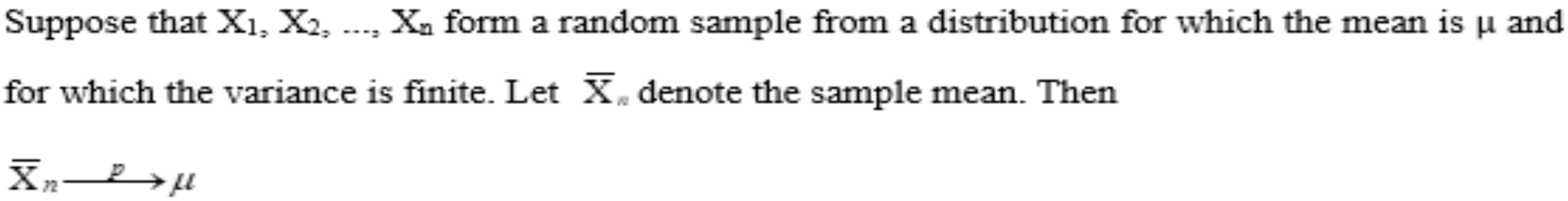
$$\begin{aligned}\widehat{\beta} &= [\frac{1}{N}\sum_{i=1}^{N}(X_i X_i')]^{-1}\frac{1}{N}\sum_{i=1}^{N}(X_i Y_i) \\ &= [\sum_{i=1}^{N}(X_i X_i')]^{-1}\sum_{i=1}^{N}(X_i Y_i) \\ &= (X'X)^{-1}X'Y\end{aligned}$$

- *What is the asymptotic sampling distribution of* $\hat{\beta}$?

Basic Asymptotic Theory

- **The law of large numbers**

○ Sample moments converge in probability to the corresponding population moments.



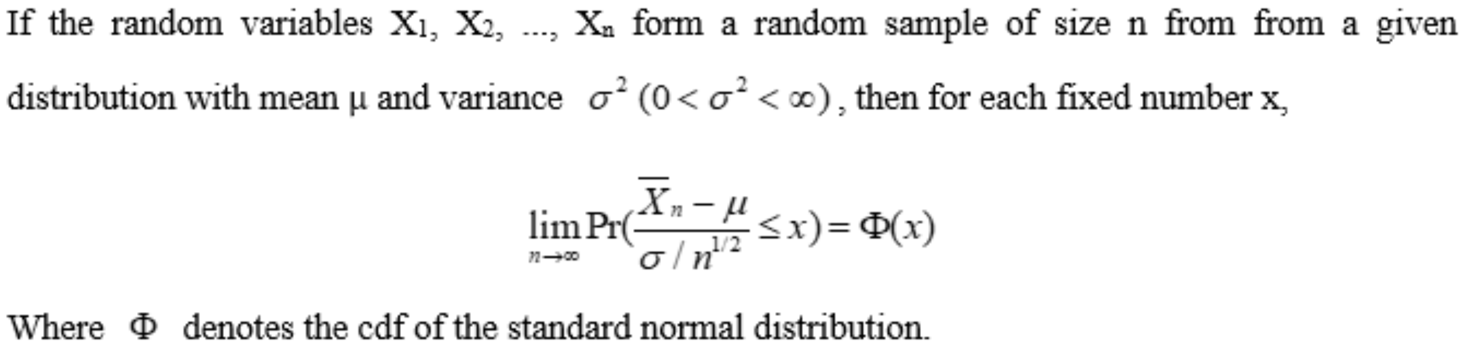
Suppose that X_1, X_2, ..., X_n form a random sample from a distribution for which the mean is μ and for which the variance is finite. Let $\bar{X}_n$ denote the sample mean. Then

$$\bar{X}_n \xrightarrow{p} \mu$$

Basic Asymptotic Theory

- **The central limit theorem**

○Sample moments are asymptotically normally distributed (after subtracting the corresponding population moment and multiplying by the square root of the sample size).

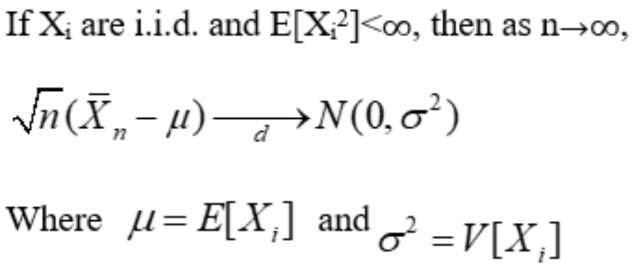


If the random variables X_1, X_2, ..., X_n form a random sample of size n from from a given distribution with mean μ and variance σ^2 $(0 < \sigma^2 < \infty)$, then for each fixed number x,

$$\lim_{n\to\infty} \Pr\left(\frac{\overline{X}_n - \mu}{\sigma / n^{1/2}} \le x\right) = \Phi(x)$$

Where Φ denotes the cdf of the standard normal distribution.

○Or,



If X_i are i.i.d. and $E[X_i^2]<\infty$, then as $n\to\infty$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[d]{} N(0,\sigma^2)$$

Where $\mu = E[X_i]$ and $\sigma^2 = V[X_i]$

- **Slutsky's theorem**
  - Let a_N be a statistic with an asymptotic distribution and let b_N be a statistic with probability limit b. Then $a_N + b_N$ and $a_N + b$ have the same asymptotic distribution.
  - Let a_N be a statistic with an asymptotic distribution and let b_N be a statistic with probability limit b. Then $a_N b_N$ and $a_N b$ have the same asymptotic distribution.

- **The Continuous Mapping Theorem**
  - The probability limit of $h(b_N)$ is $h(b)$, where $plim b_N = b$ and $h(\cdot)$ is continuous at b.

- **The Delta Method**
  - Consider a vector-valued random variable that is asymptotically normally distributed. Continuously differentiable scalar functions of this random variable are also asymptotically normally distributed, with the covariance matrix given by a quadratic form with the covariance matrix of the random variable on the inside and the gradient of the function evaluated at the probability limit of the random variable on the outside.
  - The asymptotic distribution of $h(b_N)$ is normal with covariance matrix $\nabla h(b)' \Omega \nabla h(b)$, where plim $b_N = b$, h(·) is continuously differentiable at b with gradient $\nabla h(b)$, and b_N is normally distributed and has asymptotic covariance matrix Ω.

$$\mathrm{Y_i} = \mathrm{X_i'}\beta + (\mathrm{Y_i} - \mathrm{X_i'}\beta)$$

$$= \mathrm{X_i'}\beta + \mathrm{e_i}$$

$$\mathrm{e_i} \equiv \mathrm{Y_i} - \mathrm{X_i'}\beta \ \ \text{(residual term)}$$

$$\mathrm{E(X_i'e_i)} = 0 \ \ \text{(F.O.C)}$$

$$\hat{\beta} = [\sum_{i=1}^{N}(X_i X_i')]^{-1}\sum_{i=1}^{N}(X_i Y_i)$$

$$= [\sum_{i=1}^{N}(X_i X_i')]^{-1}\sum_{i=1}^{N}[X_i(\mathrm{X_i'}\beta + \mathrm{e_i})]$$

$$= \beta + [\sum_{i=1}^{N}(X_i X_i')]^{-1}\sum_{i=1}^{N}X_i \mathrm{e_i}$$

- Therefore, $\hat{\beta}$ has an asymptotic normal distribution with probability limit β and covariance matrix

$$\Omega = E(X_i X_i')^{-1}E(X_i X_i' e_i^2)E(X_i X_i')^{-1}$$

- Proof:

$$\sqrt{N}(\hat{\beta} - \beta) = \sqrt{N}[\sum_{i=1}^{N}(X_i X_i')]^{-1}\sum_{i=1}^{N}X_i e_i$$

$$= \sqrt{N}[\frac{1}{N}\sum_{i=1}^{N}(X_i X_i')]^{-1}\sum_{i=1}^{N}\frac{1}{N}(X_i e_i)\xrightarrow{d}N(0,\Omega)$$

$$[\frac{1}{N}\sum_{i=1}^{N}(X_i X_i')]^{-1}\xrightarrow{\mathrm{p}}\mathrm{E}(X_i X_i')^{-1}$$

$$\sum_{i=1}^{N}\frac{1}{N}(X_i e_i)\xrightarrow{d}N(0,\Omega_0)$$

- In practice, these standard errors are estimated by substituting sums for expectations and using the estimated residuals to form the empirical fourth-moment matrix:

$$\widehat{\Omega} = [\sum_{i=1}^{N}(X_i X_i')/N]^{-1}[\sum_{i=1}^{N}(X_i X_i' \hat{e}_i^2)/N][\sum_{i=1}^{N}(X_i X_i')/N]^{-1}$$

- Asymptotic standard errors computed in this way are known as heteroskedasticity-consistent standard errors, White (1980a) standard errors, or Eicker-White standard errors, in recognition of Eicker’s (1967) derivation. They are also known as “robust” standard errors (e.g., in Stata).

- Why robust?

- In large enough samples, they provide accurate hypothesis tests and confidence intervals given minimal assumptions about the data and model. In particular, our derivation of the limiting distribution makes no assumptions other than those needed to ensure that basic statistical results like CLT

- Given that (Homoskedasticity Assumption),

$$E(e_i^2 \mid X_i) = \sigma^2$$

- By iterating expectations, we have

$$E(X_i X_i ' e_i^2) = E(X_i X_i '\mathrm{E}(e_i^2 \mid X_i)) = \sigma^2 E(X_i X_i ')$$

- The asymptotic covariance matrix of $\hat{\beta}$ then simplifies to (by default in Stata)

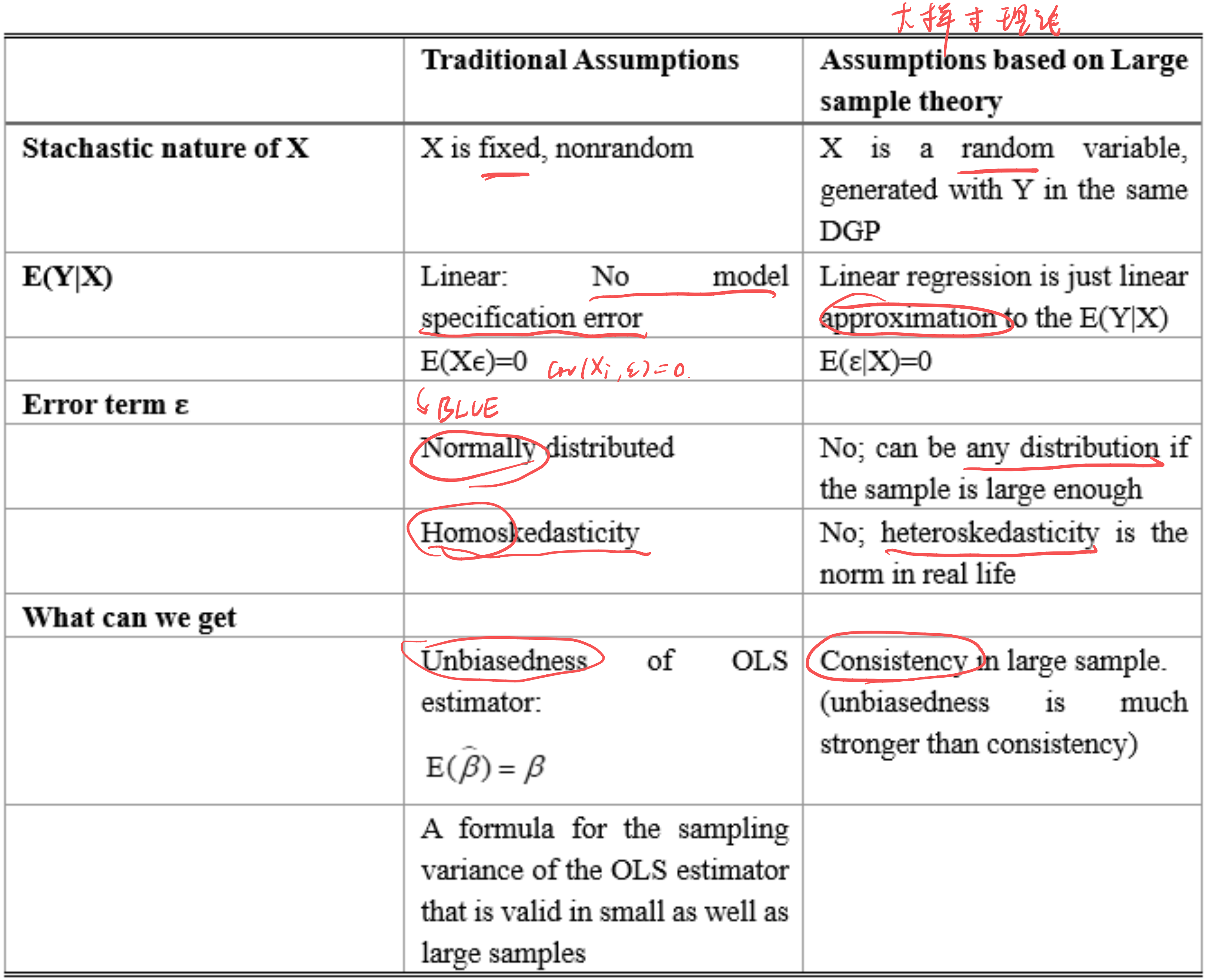
$$\begin{aligned}
&E(X_i X_i ')^{-1} E(X_i X_i ' e_i^2) E(X_i X_i ')^{-1} \\
&= E(X_i X_i ')^{-1} \sigma^2 E(X_i X_i ') E(X_i X_i ')^{-1} \\
&= \sigma^2 E(X_i X_i ')^{-1}
\end{aligned}$$

- Our view of regression as an approximation to the CEF makes heteroskedasticity seem natural. If the CEF is nonlinear and you use a linear model to approximate it, then the quality of fit between the regression line and the CEF will vary with X_i. Hence, the residuals will be larger, on average, at values of X_i where the fit is poorer.

$$
\begin{aligned}
&E[(Y_i - X_i^{'}\beta)^2 \mid X_i] \\
&\quad = \mathrm{E}\{[Y_i - E(Y_i \mid X_i)] + [E(Y_i \mid X_i) - X_i\beta] \mid X_i\}^2 \\
&\quad = \mathrm{E}\{[Y_i - E(Y_i \mid X_i)]^2 + [E(Y_i \mid X_i) - X_i\beta]^2 \\
&\qquad\qquad + 2[Y_i - E(Y_i \mid X_i)][E(Y_i \mid X_i) - X_i\beta] \mid X_i\} \\
&\quad = \mathrm{E}\{[Y_i - E(Y_i \mid X_i)]^2 + [E(Y_i \mid X_i) - X_i\beta]^2 \mid X_i\} \\
&\quad = \mathrm{V}(Y_i \mid X_i) + [E(Y_i \mid X_i) - X_i\beta]^2
\end{aligned}
$$

- Therefore, even if *$V[Y_i|X_i]$* is constant, the residual variance increases with the square of the gap between the regression line and the CEF, a fact noted in White (1980b)

大样本理论



	Traditional Assumptions	Assumptions based on Large sample theory
Stachastic nature of X	X is fixed, nonrandom	X is a random variable, generated with Y in the same DGP
E(Y|X)	Linear: No model specification error	Linear regression is just linear approximation to the E(Y|X)
	E(Xϵ)=0 $Cov(X_i, \varepsilon)=0.$	E(ε|X)=0
Error term ε	↳ BLUE	
	Normally distributed	No; can be any distribution if the sample is large enough
	Homoskedasticity	No; heteroskedasticity is the norm in real life
What can we get		
	Unbiasedness of OLS estimator: $E(\hat{\beta}) = \beta$	Consistency in large sample. (unbiasedness is much stronger than consistency)
	A formula for the sampling variance of the OLS estimator that is valid in small as well as large samples	

$V(\hat{\beta}).$

Saturated Model

- Saturated regression models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables.
- Case 1: $Y_i = a + bD_i + u_i$

- Case 2: A saturated regression model for s_i , s_i = 0, 1, 2, . . . , τ, is

$$Y_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + \cdots + \beta_\tau d_{\tau i} + \varepsilon_i,$$

where $d_{ji} = 1[s_i = j]$ is a dummy variable indicating schooling level j, and β_j is said to be the jth-level schooling effect. Note that

$$\beta_j = E[Y_i|s_i = j] - E[Y_i|s_i = 0],$$

while $\alpha = E[Y_i|s_i = 0]$. In practice, you can pick any value of s_i for the reference group; a regression model is saturated as long as it has one parameter for every possible j in $E[Y_i|s_i = j]$.

- Saturated regression models fit the CEF perfectly because the CEF is a linear function of the dummy regressors used to saturate. This is an important special case of the linear CEF theorem.

- Finally, it’s important to recognize that a saturated model fits the CEF perfectly regardless of the distribution of Yi. For example, this is true for linear probability models and other limited dependent variable models (e.g., non-negative Y_i).

If there are two discrete variables

- Let x_{1i} indicate college graduates and x_{2i} indicate women. The CEF given x_{1i} and x_{2i} takes on four values:

$$E[Y_i|x_{1i}=0, x_{2i}=0],$$
$$E[Y_i|x_{1i}=1, x_{2i}=0],$$
$$E[Y_i|x_{1i}=0, x_{2i}=1],$$
$$E[Y_i|x_{1i}=1, x_{2i}=1].$$

- We can label these using the following scheme

$$E[Y_i|x_{1i}=0, x_{2i}=0]=\alpha$$
$$E[Y_i|x_{1i}=1, x_{2i}=0]=\alpha+\beta_1$$
$$E[Y_i|x_{1i}=0, x_{2i}=1]=\alpha+\gamma$$
$$E[Y_i|x_{1i}=1, x_{2i}=1]=\alpha+\beta_1+\gamma+\delta_1.$$

- It can be written in terms of Greek letters as

$$E[Y_i|x_{1i}, x_{2i}]=\alpha+\beta_1 x_{1i}+\gamma x_{2i}+\delta_1(x_{1i}x_{2i}),$$

- The saturated regression equation becomes

$$Y_i=\alpha+\beta_1 x_{1i}+\gamma x_{2i}+\delta_1(x_{1i}x_{2i})+\varepsilon_i.$$

X_{1i} X_{2i}

$$\begin{cases} X_{10}=1, \text{ if } X_{1i}=1, X_{2i}=0 \\ X_{01}=1, \text{ if } X_{1i}=0, X_{2i}=1 \\ X_{11}=1, \text{ if } X_{1i}=1, X_{2i}=1 \end{cases}$$

$$Y_i=\tilde{\alpha}+\tilde{\beta}_1 X_{10i}+\tilde{\gamma} X_{01i}+\tilde{\delta} X_{11i}+\varepsilon_i$$

$$\tilde{\alpha}=\alpha \quad \tilde{\beta}_1=\beta_1 \quad \tilde{\gamma}=\gamma$$

$$\tilde{\delta}=\beta_1+\gamma+\delta.$$

- We can combine the multivalued schooling variable with sex to produce a saturated model that has τ main effects for schooling, one main effect for sex, and τ sex-schooling interactions

$$Y_i = \alpha + \sum_{j=1}^{\tau} \beta_j d_{ji} + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i.$$

- The coefficients on the interaction terms, δ_j, tell us *how each of the schooling effects differ by sex*. The CEF in this case takes on 2(τ + 1) values, while the regression has this many parameters.

- It would be strange to estimate a model that included interaction terms but omitted the corresponding main effects.

$$Y_i = \alpha + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji} x_{2i}) + \varepsilon_i.$$

- This model allows schooling to shift wages only for women, something very far from the truth. Consequently, the results are likely to be hard to interpret.

V. Regression and Causality: Conditional Independence Assumption (CIA)

- *When regression has a causal interpretation?*
- When can we think of a regression coefficient as approximating the causal effect that might be revealed in an experiment?
- *A regression is causal when the CEF it approximates is causal. (A regression inherits its legitimacy from a CEF)*
- Causal relationships in terms of the potential outcomes notation describe what would happen to a given individual in a hypothetical comparison of alternative hospitalization scenarios. Differences in these potential outcomes were said to be the causal effect of some treatment event.
- *The CEF is causal when it describes differences in average potential outcomes for a fixed reference population.*

Schooling and Earnings

- let’s stick with the schooling example. The causal connection between schooling and earnings can be defined as the functional relationship that describes what a given individual would earn if he or she obtained different levels of education.
- In particular, we might think of schooling decisions as being made in a series of episodes where the decision maker can realistically go one way or another, even if certain choices are more likely than others.
- In empirical work, the causal relationship between schooling and earnings tells us what people would earn, on average, if we could either change their schooling in a perfectly controlled environment or change their schooling randomly so that those with different levels of schooling would be otherwise comparable.
- Experiments ensure that the causal variable of interest is independent of potential outcomes so that the groups being compared are truly comparable. Here, we would like to generalize this notion to causal variables that take on more than two values, and to more complicated situations where we must hold a variety of control variables fixed for causal inferences to be valid.
- Generalize the notion to causal variables that take on more than two variables, and to more complicated situations where we must hold a variety of control variables fixed for causal inferences to be valid.

- ***Conditional independence assumption (CIA),*** a core assumption that provides the (sometimes implicit) justification for the causal interpretation of regression estimates.
- This assumption is also called ***selection on observables*** because the covariates to be held fixed are assumed to be known and observed (e.g., in Goldberger, 1972; Barnow, Cain, and Goldberger, 1981).

(The big question, therefore, is what these control variables are, or should be.)

- X_i, a vector including education level, ability, and family background

- To address this question, we imagine two potential earnings variables:

$$\textit{Potential outcome} = \begin{cases} Y_{1i} & \text{if } C_i = 1 \\ Y_{0i} & \text{if } C_i = 0 \end{cases}.$$

- The observed outcome, Y_i, can be written in terms of potential outcomes as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})C_i.$$

- We therefore hope to measure the average of Y_{1i} - Y_{0i}, or the average for some groups, such as those who went to college. This is $E[Y_{1i}-Y_{0i}|C_i = 1]$.

- In general, comparisons of those who do and don't go to college are likely to be a poor measure of the causal effect of college attendance.

$$\underbrace{E[Y_i|C_i = 1] - E[Y_i|C_i = 0]}_{\text{Observed difference in earnings}} = \underbrace{E[Y_{1i} - Y_{0i}|C_i = 1]}_{\text{Average treatment effect on the treated}}$$
$$+ \underbrace{E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]}_{\text{Selection bias}}.$$

A simple case of CIA

- The CIA asserts that conditional on observed characteristics, X_i, selection bias disappears. Formally, this means

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp C_i | X_i,$$

- Given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,

$$E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i].$$

Generalization to continuous or multi-value variables

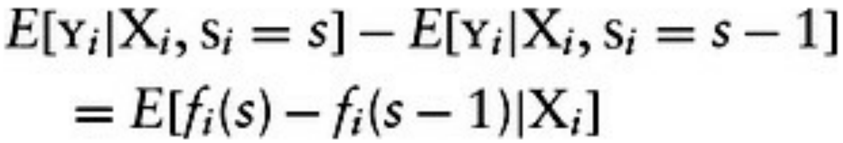
- Now, we'd like to expand the conditional independence assumption to causal relations that involve variables that can take on more than two values, such as years of schooling, s_i.
- The causal relationship between schooling and earnings is likely to be different for each person. We therefore use the individual-specific functional notation,

$$\mathrm{Y}_{si} \equiv f_i(s),$$

- To denote the potential earnings that person *i* would receive after obtaining *s* years of education.
- The CIA in this more general setup becomes

$$\mathrm{Y}_{si} \perp\!\!\!\perp \mathrm{s}_i | \mathrm{X}_i, \text{ for all } s.$$

- In many randomized experiments, the CIA crops up because s_i is randomly assigned conditional on X_i (in the Tennessee STAR experiment, for example, small classes were randomly assigned within schools).
- In an observational study, the CIA means that s_i can be said to be “as good as randomly assigned,” conditional on X_i.
- The data reveals only $Y_i = f_i(s_i)$, that is, $f_i(s)$ for $s = s_i$.
- However, given the CIA, conditional-on-X_i comparisons of average earnings across schooling levels have a causal interpretation. In other words,



$$\begin{aligned} &E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = s] - E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = s-1] \\ &\quad = E[f_i(s) - f_i(s-1)|\mathrm{X}_i] \end{aligned}$$

for any value of s.

- For example, the average causal effect of high school graduation

$$\begin{aligned}&E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 12] - E[\mathrm{Y}_i|\mathrm{X}_i, \mathrm{s}_i = 11]\\&\quad = E[f_i(12)|\mathrm{X}_i, \mathrm{s}_i = 12] - E[f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 11].\end{aligned}$$

- This comparison has a causal interpretation because, given the CIA

$$\begin{aligned}&E[f_i(12)|\mathrm{X}_i, \mathrm{s}_i = 12] - E[f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 11]\\&\quad = E[f_i(12) - f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 12].\end{aligned}$$

- Note also that in this case, the causal effect of graduating from high school on high school graduates is equal to the average high school graduation effect at X_i:

$$E[f_i(12) - f_i(11)|\mathrm{X}_i, \mathrm{s}_i = 12] = E[f_i(12) - f_i(11)|\mathrm{X}_i].$$

- *How many treatment effects??*

- So far, we have constructed separate causal effects for each value taken on by the conditioning variables. *This leads to as many causal effects as there are values of X_i, an embarrassment of riches.*

- Empiricists almost always find it useful to boil a set of estimates down to a single summary measure, such as the unconditional or overall average causal effect. By the law of iterated expectations, *the unconditional average causal effect of high school graduation* is

$$\begin{aligned} &E\{E[\text{Y}_i|\text{X}_i, \text{s}_i = 12] - E[\text{Y}_i|\text{X}_i, \text{s}_i = 11]\} \\ &= E\{E[f_i(12) - f_i(11)|\text{X}_i]\} \\ &= E[f_i(12) - f_i(11)]. \end{aligned}$$

- In the same spirit, we might be interested in the average causal effect of high school graduation on high school graduates:

$$\begin{aligned} &E\{E[\text{Y}_i|\text{X}_i, \text{s}_i = 12] - E[\text{Y}_i|\text{X}_i, \text{s}_i = 11]|\text{s}_i = 12\} \\ &= E\{E[f_i(12) - f_i(11)|\text{X}_i]|\text{s}_i = 12\} \\ &= E[f_i(12) - f_i(11)|\text{s}_i = 12]. \end{aligned}$$

- This parameter tells us how much high school graduates gained by virtue of having graduated

- The unconditional average effect, can be computed by averaging all the *X*-specific effects weighted by the marginal distribution of X_i, while the average effect on high school or college graduates averages the *X*-specific effects weighted by the distribution of X_i in these groups.

X_i

- In both cases, the empirical counterpart is a ***matching estimator***: (1) We make comparisons across schooling groups for individuals with the same covariate values, compute the difference in their average earnings, and then (2) average these differences in some way.

- Drawbacks of the matching approach:
  - It is not automatic; rather, it requires two steps, matching and averaging.
  - Estimating the standard errors of the resulting estimates may not be straightforward, either.
  - Since s_i takes on many values, there are separate average causal effects for each possible increment in s_i, which also must be summarized in some way.
- These considerations lead us back to regression.

CIA and Regression

- Regression provides an easy-to-use empirical strategy that automatically turns the CIA into causal effects.

- Two routes can be traced from the CIA to regression:

  - One assumes that $f_i(s)$ is both linear in s and the same for everyone except for an additive error term, in which case linear regression is a natural tool to estimate the features of $f_i(s)$.

  - A more general but somewhat longer route recognizes that $f_i(s)$ almost certainly differs for different people, and moreover need not be linear in s. Even so, allowing for random variation in $f_i(s)$ across people and for nonlinearity for a given person, regression can be thought of as a strategy for the *estimation of a weighted average of the individual-specific difference, $f_i(s) - f_i(s - 1)$.*

- In fact, regression can be seen as a particular sort of matching estimator, capturing an average causal effect

- We therefore start with the first route, a linear constant effects causal model. Suppose that

$$f_i(s) = \alpha + \rho s + \eta_i.$$

- Substituting the observed value s_i for s in equation (3.2.7), we have

$$Y_i = \alpha + \rho S_i + \eta_i.$$

- Suppose now that the CIA holds given a vector of observed covariates, X_i. In addition to the functional form assumption for potential outcomes embodied in (3.2.8), we decompose the random part of potential earnings, η_i, into a linear function of observable characteristics, X_i, and an error term, v_i:

$$\eta_i = X_i'\gamma + v_i,$$

where γ is a vector of population regression coefficients that is assumed to satisfy $E[\eta_i|X_i] = \gamma$. Since γ is defined by the regression of η_i on X_i , the residual v_i and X_i are uncorrelated by construction. Moreover, by virtue of the CIA, we have

$$\begin{aligned} E[f_i(s)|X_i, s_i] = E[f_i(s)|X_i] &= \alpha + \rho s + E[\eta_i|X] \\ &= \alpha + \rho s + X_i'\gamma. \end{aligned}$$

- The residual in the linear causal model

$$Y_i = \alpha + \rho s_i + X_i'\gamma + v_i$$

is therefore uncorrelated with the regressors, s_i and X_i, and the regression coefficient ρ is the causal effect of interest.

⇒

- $E(v_i | s_i, X_i)=0$

At this point, it’s worth considering when the CIA is most likely to give a plausible basis for empirical work.

The best-case scenario is a random assignment of s_i, conditional on X_i, in some sort of (possibly natural) experiment. Black et al. (2003).

A second scenario that favors the CIA leans on detailed institutional knowledge regarding the process that determines s_i. An example is the Angrist (1998) study of the effect of voluntary military service on the later earnings of soldiers.

VI. Omitted Variable Bias (OVB)

- The omitted variables bias (OVB) formula describes the relationship between regression estimates in models with different sets of control variables.
- This important formula is often motivated by the notion that a longer regression—one with controls—has a causal interpretation, while a shorter regression does not. The coefficients on the variables included in the shorter regression are, therefore, said to be biased.
- In fact, the OVB formula is a mechanical link between coefficient vectors that applies to short and long regressions whether or not the longer regression is causal.

- The regression of wages on schooling, s_i , controlling for ability can be written as

$$Y_i = \alpha + \rho S_i + A_i'\gamma + e_i,$$

where α, ρ, and γ are population regression coefficients and e_i is a regression residual that is uncorrelated with all regressors by definition. If the CIA applies given A_i, then ρ can be equated with the coefficient in the linear causal model, while the residual e_i is the random part of potential earnings that is left over after controlling for A_i.

- In practice, ability is hard to measure, resulting in the “short regression”

$$Y_i = \alpha + \rho S_i + \eta_i.$$

- What are the consequences of leaving ability out of regression?

- ***Omitted variables bias formula***

$$\frac{Cov(Y_i, s_i)}{V(s_i)} = \rho + \gamma' \delta_{As},$$

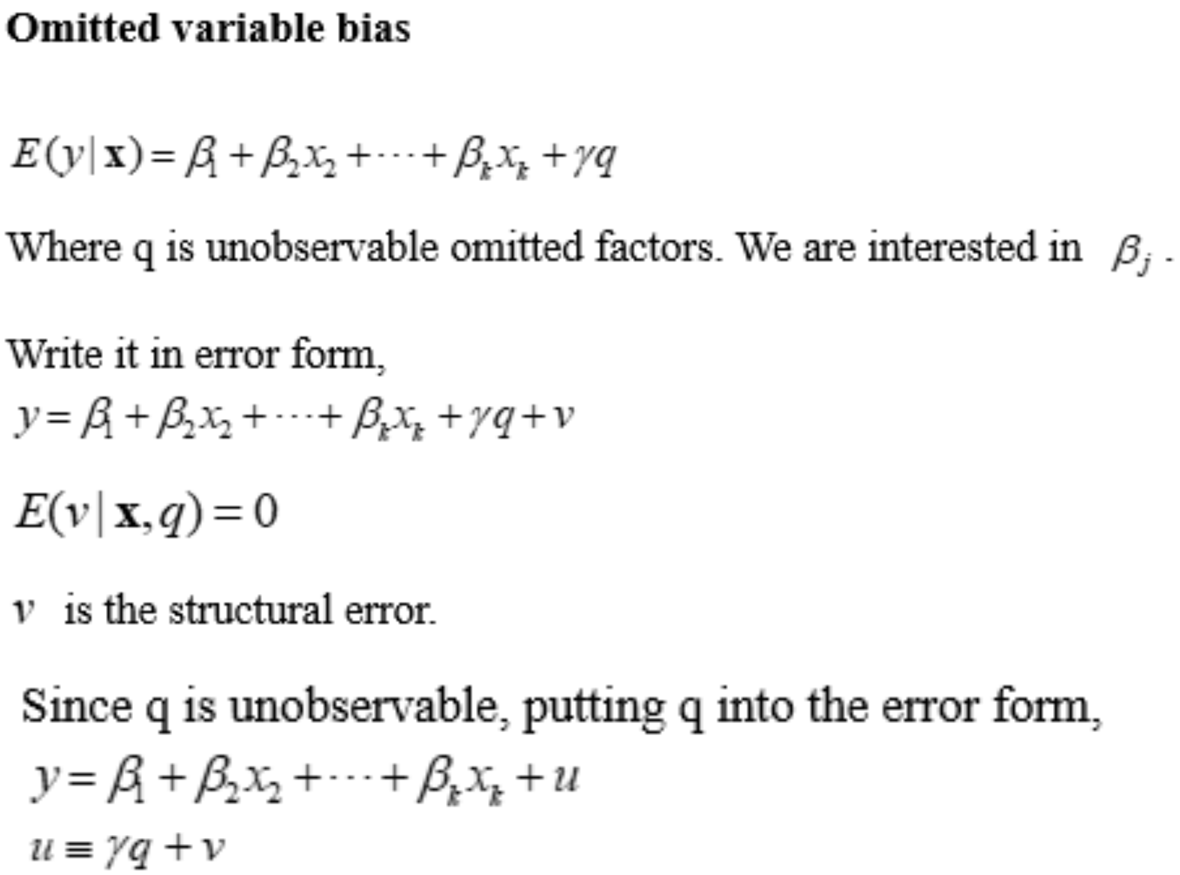
where δ_{As} is the vector of coefficients from regressions of the elements of A_i on s_i.

- Proof:

- $$\frac{Cov(Y_i, s_i)}{V(s_i)}$$
$$= \frac{Cov(\alpha + \rho s_i + A_i^{'}\gamma + v_i, s_i)}{V(s_i)}$$
$$= \frac{\rho V(s_i) + Cov(A_i^{'}\gamma, s_i)}{V(s_i)}$$
$$= \rho + \gamma ' \delta_{\mathrm{AS}}$$

- We can use the OVB formula to get a sense of the likely consequences of omitting ability for schooling coefficients.
  - Ability variables have positive effects on wages, and these variables are also likely to be positively correlated with schooling.
  - The short regression coefficient may, therefore, be “too big” relative to what we want.

More Generally



Omitted variable bias

$$E(y|\mathbf{x}) = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma q$$

Where q is unobservable omitted factors. We are interested in β_j .

Write it in error form,

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma q + v$$

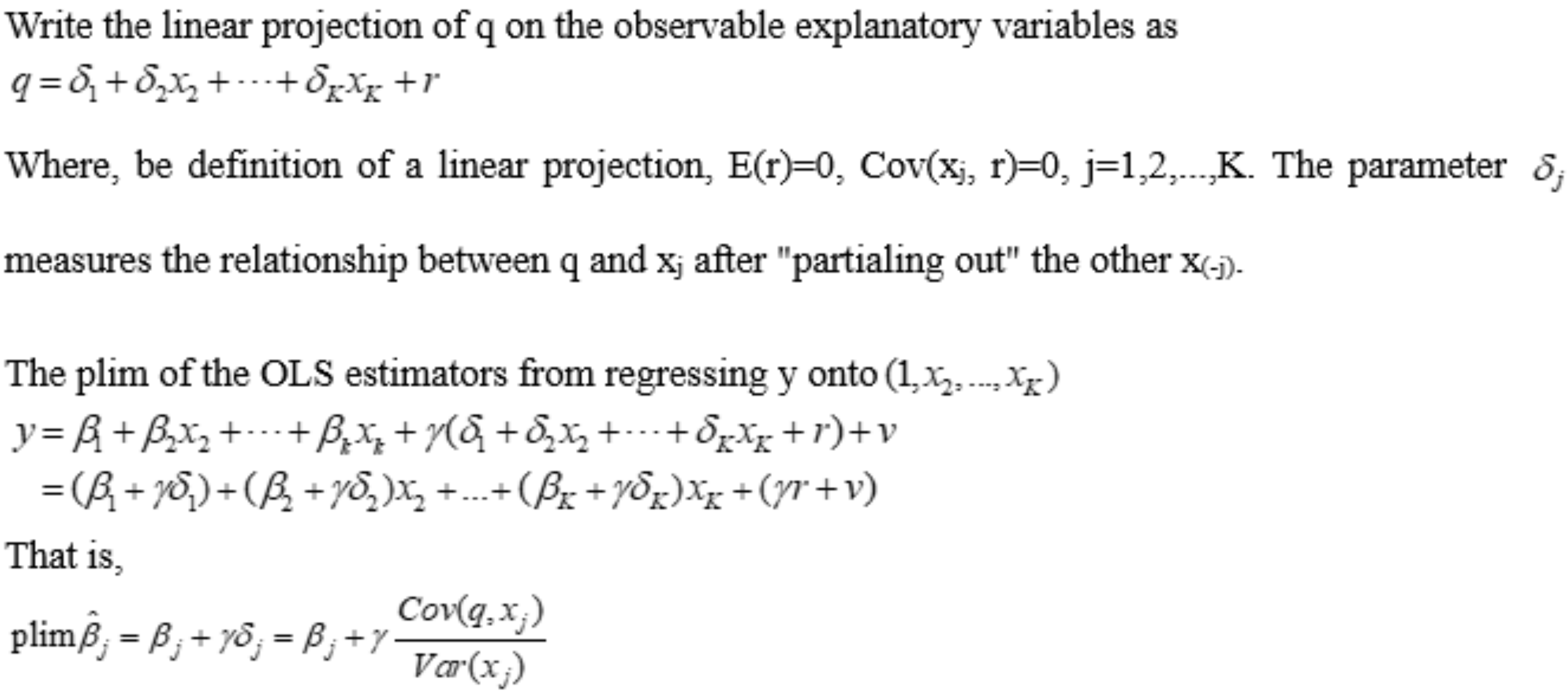
$$E(v \mid \mathbf{x}, q) = 0$$

v is the structural error.

Since q is unobservable, putting q into the error form,

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

$$u \equiv \gamma q + v$$



Write the linear projection of q on the observable explanatory variables as

$$q = \delta_1 + \delta_2 x_2 + \cdots + \delta_K x_K + r$$

Where, be definition of a linear projection, E(r)=0, Cov(x_j, r)=0, j=1,2,...,K. The parameter δ_j measures the relationship between q and x_j after "partialing out" the other $x_{(-j)}$.

The plim of the OLS estimators from regressing y onto $(1, x_2, \ldots, x_K)$

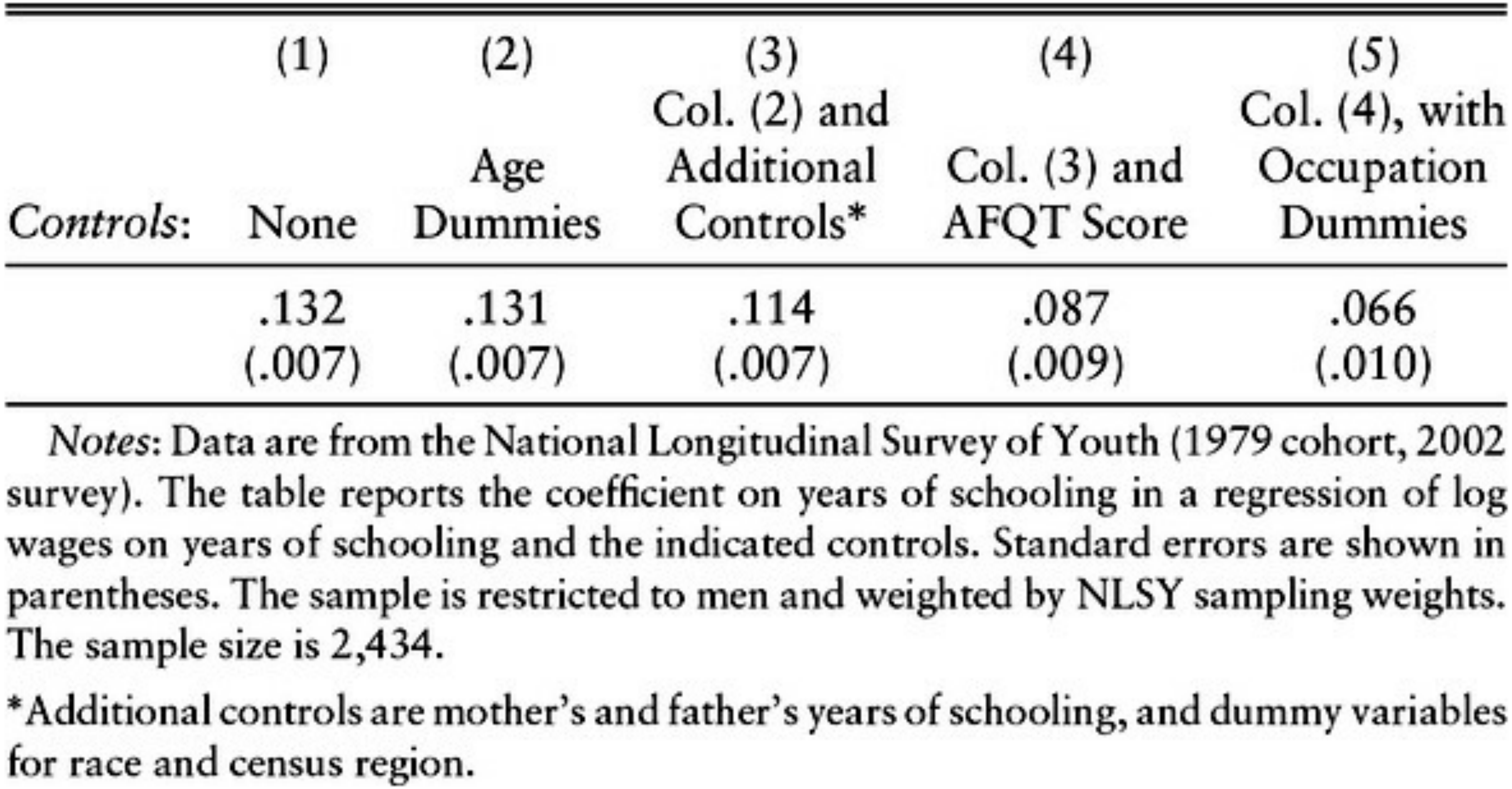
$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \gamma(\delta_1 + \delta_2 x_2 + \cdots + \delta_K x_K + r) + v$$
$$= (\beta_1 + \gamma\delta_1) + (\beta_2 + \gamma\delta_2)x_2 + \ldots + (\beta_K + \gamma\delta_K)x_K + (\gamma r + v)$$

That is,

$$\text{plim}\,\hat{\beta}_j = \beta_j + \gamma\delta_j = \beta_j + \gamma\frac{Cov(q, x_j)}{Var(x_j)}$$

What’s the difference? Sometimes one omitted variable can contaminate the estimation of all coefficients, which might be ignored unconsciously. Must be very cautious.

Regression sensitive analysis: one of the OVB applications



	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age Dummies	Col. (2) and Additional Controls*	Col. (3) and AFQT Score	Col. (4), with Occupation Dummies
	.132	.131	.114	.087	.066
	(.007)	(.007)	(.007)	(.009)	(.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2,434.

*Additional controls are mother's and father's years of schooling, and dummy variables for race and census region.

VII. Bad Controls

- We’ve made the point that control for covariates can increase the likelihood that regression estimates have a causal interpretation. But more control is not always better.
- Some variables are bad controls and should not be included in a regression model even when their inclusion might be expected to change the short regression coefficients.
- ***Bad controls*** are variables that are themselves outcome variables in the notional experiment at hand. That is, bad controls might just as well be dependent variables too.
- ***Good controls*** are variables that we can think of as having been fixed at the time the regressor of interest was determined.

- The essence of the bad control problem is a version of selection bias, albeit somewhat more subtle.
- To illustrate, suppose we are interested in the effects of a college degree on earnings and that people can work in one of two occupations, white collar, and blue collar. A college degree clearly opens the door to higher-paying white-collar jobs.
- Should occupation, therefore, be seen as an omitted variable in a regression of wages on schooling?
- After all, occupation is highly correlated with both education and pay. Perhaps it’s best to look at the effect of college on wages for those within an occupation, say white collar only.
- The problem with this argument is that once we acknowledge the fact that college affects occupation, comparisons of wages by college degree status within an occupation are no longer apples-to-apples comparisons, even if college degree completion is randomly assigned.

- Formally, let w_i be a dummy variable that denotes white collar workers and let Y_i denote earnings.
- The realization of these variables is determined by college graduation status and potential outcomes that are indexed against c_i. We have

$$\begin{aligned} Y_i &= C_i Y_{1i} + (1 - C_i) Y_{0i} \\ W_i &= C_i W_{1i} + (1 - C_i) W_{0i}, \end{aligned}$$

where c_i = 1 for college graduates and is zero otherwise, $\{Y_{1i}, Y_{0i}\}$ denotes potential earnings, and $\{w_{1i}, w_{0i}\}$ denotes potential white collar status.

- At this point, we assume that c_i is randomly assigned, so it is independent of all potential outcomes. We have no trouble estimating the causal effect of c_i on either Y_i or w_i since independence gives us

$$\begin{aligned} E[Y_i|C_i = 1] - E[Y_i|C_i = 0] &= E[Y_{1i} - Y_{0i}], \\ E[W_i|C_i = 1] - E[W_i|C_i = 0] &= E[W_{1i} - W_{0i}]. \end{aligned}$$

- In practice, we can estimate these average treatment effects by regressing Y_i and w_i on c_i .

- Bad control means that a comparison of earnings conditional on w_i does not have a causal interpretation.

- Consider the difference in mean earnings between college graduates and others, conditional on working at a white-collar job. We can compute this in a regression model that includes w_i or by regressing Y_i on c_i in the sample where $w_i = 1$.

- The estimand in the latter case is the difference in means with c_i switched off and on, conditional on $w_i = 1$:

$$\begin{aligned} &E[\mathrm{Y}_i|\mathrm{W}_i = 1, \mathrm{C}_i = 1] - E[\mathrm{Y}_i|\mathrm{W}_i = 1, \mathrm{C}_i = 0] \\ &\quad = E[\mathrm{Y}_{1i}|\mathrm{W}_{1i} = 1, \mathrm{C}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{W}_{0i} = 1, \mathrm{C}_i = 0]. \end{aligned}$$

$$\begin{aligned} &E[\mathrm{Y}_{1i}|\mathrm{W}_{1i} = 1, \mathrm{C}_i = 1] - E[\mathrm{Y}_{0i}|\mathrm{W}_{0i} = 1, \mathrm{C}_i = 0] \\ &\quad = E[\mathrm{Y}_{1i}|\mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i}|\mathrm{W}_{0i} = 1]. \end{aligned}$$

- By the joint independence of $\{Y_{1i}, w_{1i}, Y_{0i}, w_{0i}\}$ and c_i, we have

$$\begin{aligned} &E[\mathrm{Y}_{1i}|\mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i}|\mathrm{W}_{0i} = 1] \\ &\quad = \underbrace{E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i}|\mathrm{W}_{1i} = 1]}_{\text{Causal effect}} + \underbrace{\{E[\mathrm{Y}_{0i}|\mathrm{W}_{1i} = 1] - E[\mathrm{Y}_{0i}|\mathrm{W}_{0i} = 1]\}}_{\text{Selection bias}}. \end{aligned}$$

- This expression illustrates the apples-to-oranges nature of the bad control problem

$$
\begin{aligned}
&E[\mathrm{Y}_{1i}|\mathrm{W}_{1i}=1]-E[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1]\\
&\quad=\underbrace{E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1]}_{\text{Causal effect}}+\underbrace{\{E[\mathrm{Y}_{0i}|\mathrm{W}_{1i}=1]-E[\mathrm{Y}_{0i}|\mathrm{W}_{0i}=1]\}}_{\text{Selection bias}}.
\end{aligned}
$$

- In other words, the difference in wages between those with and those without a college degree conditional on working in a white-collar job equals the causal effect of college on those with w_{1i} = 1 (people who work at a white-collar job when they have a college degree) and a selection bias term that reflects the fact that college changes the composition of the pool of white-collar workers.
- It is also incorrect to say that the conditional comparison captures the part of the effect of college that is “not explained by occupation.”
- In fact, the conditional comparison does not tell us much that is useful without a more elaborate model of the links between college, occupation, and earnings.

Bad control scenario II: Proxy control

- A second version of the bad control scenario involves proxy control, that is, the inclusion of variables that might partially control for omitted factors but are themselves affected by the variable of interest.

- A simple version of the proxy control story goes like this:

$$Y_i = \alpha + \rho s_i + \gamma a_i + e_i,$$

Where a_i is a scalar ability measure. Think of this as an *IQ* score that measures innate ability in eighth grade, before any relevant schooling choices are made (assuming everyone completes eighth grade).

- The error term in this equation satisfies E[$s_i e_i$] = E[$a_i e_i$] = 0 by definition.
- Since a_i is measured before s_i is determined, it is a good control.

- Unfortunately, data on a_i are unavailable. However, you have a second ability measure collected later, after schooling is completed (say, the score on a test used to screen job applicants). Call this variable *late ability*, a_{li}. In general, schooling increases late ability relative to innate ability. To be specific, suppose

$$a_{li} = \pi_0 + \pi_1 S_i + \pi_2 a_i.$$

- You're worried about OVB in the regression of Y_i on s_i alone, so you propose to regress y_i on s_i and late ability, a_{li}, since the desired control, a_i, is unavailable. Use a_{li} to substitute for a_i in the original regression function, we get the regression on s_i and a_{li}

$$Y_i = \left(\alpha - \gamma \frac{\pi_0}{\pi_2}\right) + \left(\rho - \gamma \frac{\pi_1}{\pi_2}\right) S_i + \frac{\gamma}{\pi_2} a_{li} + e_i.$$

In this scenario, γ , π_1, and π_2 are all positive, so ρ- $\gamma(\pi_1/\pi_2)$ is too small unless π_1 turns out to be zero.

- In the proxy control scenario, however, your intentions are good. And while proxy control does not generate the regression coefficient of interest, it may be an improvement on no control at all.

- In terms of the parameters in this model, the OVB formula tells us that a regression on s_i with no controls generates a coefficient of $\rho + \gamma\delta_{as}$, where δ_{as} is the slope coefficient from a regression of a_i on s_i. The schooling coefficient, $\rho - \gamma(\pi_1/\pi_2)$, might be closer to ρ than the coefficient you estimate with no control at all.

- Moreover, assuming δ_{as} is positive, you can safely say that the causal effect of interest lies between these two. *(Typically, underestimate is better than overestimate from the perspective of telling a story)*

- One moral of both the bad control and the proxy control stories is that when thinking about controls, timing matters. Variables measured before the variable of interest was determined are generally good controls.

- Often, however, the timing is uncertain or unknown. In such cases, clear reasoning about causal channels requires explicit assumptions about what happened first or the assertion that none of the control variables are themselves caused by the regressor of interest.

Measurement errors

- In the measurement error case, the variable that we do not observe has a well-defined, quantitative meaning, but our measures of it may contain error.
- *Measurement error in the dependent variable*

Let y^* denote the variable that we would like to explain

$$y^* = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + v$$

对y测量误差，方差变大.

Let y represent the observable measure of y^* where $y^* \neq y$.

The population measurement error $e_0 \equiv y - y^*$

Write $y^* = y - e_0$

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + v + e_0$$

Which is an estimable model.

Usually, it is assumed that

$$Cov(e_0, \mathbf{x}) = 0$$

y测量误差与x和v无关

and $Cov(e_0, v) = 0$, then $Var(v + e_0) = \sigma_v^2 + \sigma_e^2 > \sigma_v^2$

That is, measurement error in dependent variable results in a larger error variance than when the dependent variable is not measured with error. It in turn translates into larger asymptotic variances for the OLS estimators than if we could observe y^*.

Measurement error in an explanatory variable

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_K^* + v$$

$$E(y \mid 1, x_2, \ldots, x_{K-1}, x_K^*) = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_K^*$$

We have a measure of x_K^*, x_K, a maintained assumption [1] is v is uncorrelated with x_K.

The measurement error in the population is

$$e_K = x_K - x_K^*$$

Assume that

$$\mathrm{E}(e_K) = 0$$ 均值为0

Since v is assumed to be uncorrelated with x_K and x_K^*, then v is also uncorrelated with e_K.

A maintained assumption [2]:

$$\mathrm{E}(\mathrm{x}_j e_K) = 0\text{, j=1,2,...,K-1}$$ 与其它自变量无关.

与真值无关

- **Case 1:** Assume $Cov(e_K, x_K) = 0$
- We have:
- $$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k (x_K - e_K) + v$$
  $$= \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_K + (v - \beta_k e_K)$$

Then :

- OLS estimation with x_K in place of x_K* produces consistent estimators of all the betas
- The variance of the error is $\mathrm{Var}(v - \beta_k e_K) = \sigma_{\mathrm{v}}^2 + {\beta_k}^2 \sigma_{\mathrm{e_K}}^2$
- Therefore, the only consequence of such a measurement error is the increase in the variance of the error. However, the assumption is not very reasonable. To see this,

$$Cov(e_K, \mathrm{e}_K + x_{\mathrm{K}}^*) = \sigma_{\mathrm{e}_K}^2 + Cov(\mathrm{e}_K, x_{\mathrm{K}}^*)$$

$$Cov(\mathrm{e}_K, x_{\mathrm{K}}^*) = -\sigma_{\mathrm{e}_K}^2$$

Case 2

The classical errors-in-variables (CEV)

$Cov(e_K, x_K^*) = 0$ 与测量值无关，与真值有关.

Which implies that

$$Cov(e_K, x_K) = E(e_K x_K) = E(e_K x_K^*) + E(e_K^{*2}) = \sigma_{e_K}^2 \neq 0$$

Therefore, $$Cov(x_K, \text{v} - \beta_K e_K) = -Cov(x_K, -\beta_K e_K)$$
$$= -\beta_K Cov(x_K, e_K) = -\beta_K \sigma_{e_K}^2$$

The OLS regression generally gives inconsistent estimators of all of the betas and the inconsistency is difficult to characterize.

If x_K^* is uncorrelated with x_j, all $j \neq K$, then so is x_K, and it follows that $\text{plim}\hat{\beta}_j = \beta_j, all\ j \neq K$

$$\text{plim}\hat{\beta}_K = \beta_K \frac{\sigma^2_{r_K^*}}{\sigma^2_{r_K^*} + \sigma^2_{e_K}}$$

Where $x_K^* = \delta_1 + \delta_2 x_2 + \cdots + \delta_{K-1} x_{K-1} + r_K^*$

Proof: [see Wooldridge(2010) Problem 4.10, p85]

Hence $|\text{plim}\hat{\beta}_K| < |\beta_K|$, this is called attenuation bias.

$x_i^* = x - e_k.$

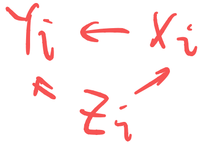
真：$y_i = \alpha + \beta x_i^* + \varepsilon_i$ 实际 $y_i = \alpha + \beta x_i + (\varepsilon_i - \beta e_k)$. 衰减误差.

$$\hat{\beta} = \frac{cov(y_i, x_i)}{var(x_i)} = \frac{cov(\alpha + \beta x_i^* + \varepsilon_i,\ x_i^* + e_k)}{var(x_i^* + e_k)} = \frac{\beta\, var(x_i^*)}{var(x_i^*) + var(e_k)} = \frac{\beta\, \sigma^2_{x_i^*}}{\sigma^2_{x_i^*} + \sigma^2_{e_k}}.$$

$$\begin{aligned} plim(\hat{\beta}_1) &= \frac{Cov(Y_i, X_{1i}^*)}{Var(X_{1i}^*)} \\ &= \frac{Cov[\beta_0 + \beta_1 X_{1i} + u_i, (X_{1i} + w_i)]}{Var(X_{1i} + w_i)} \\ &= \frac{\beta_1 Cov(X_{1i}, X_{1i})}{Var(X_{1i} + w_i)} \\ &= \beta_1 \left(\frac{Var(X_{1i})}{Var(X_{1i}) + Var(w_i)} \right) \\ &= \beta_1 \frac{\sigma^2_{X_{1i}}}{\sigma^2_{X_{1i}} + \sigma^2_w} \end{aligned}$$

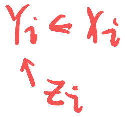
What should we control?

- Theoretically, the selection of controls should be dominated by economic theory
- Without explicit theoretical reasons, we can follow some rules of thumb statistically as follows:
  - ○Case 1: if a variable, Z_i, impacts the variable of interest (X_i) and the dependent variable (Y_i) simultaneously -- OVB Must

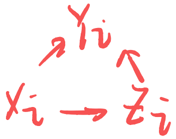


← OVB

  - ○Case 2: Z_i, impacts Y_i but not (X_i) better



  - × ○Case 3: Z_i, impacts neither Y_i nor (X_i) better not



但渠道分析？

  - ○Case 4: Z_i impacts Y_i and impacted by X_i -- Bad Control Must not
  - ○Case 5: Z_i per se is suspiciously an endogenous variable. better not to avoid complexity.

内生性

Occam's Razor: less is more.

Thanks for your Attention!

Any questions or comments please write to:

chenglingguo@nju.edu.cn