The Best Linear Predictor

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's readings

- **Blackwell** Ch. 5: Linear regression (pp. 99–118)
- **Aronow & Miller** §2.2.4: BLP definition (pp. 80–88)
- **Angrist & Pischke** §3.1.1–3.1.2: Regression and the CEF

Everything from estimation and asymptotics comes together in what follows

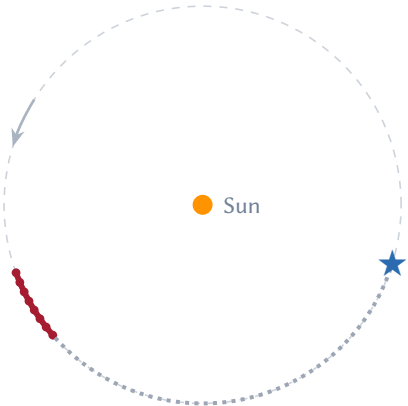
New Year's Day, 1801 — a faint light appears above Palermo, then vanishes

1 January 1801 · *Palermo, Sicily* · *Giuseppe Piazzi*

- 24 observations over 41 nights: position, position, position — meticulous records of a moving light
- 11 February: Ceres passes behind the sun and disappears
- 17 months of silence — no telescope could follow it there

17 months, 41 data points, one prediction — from Gauss, age 24

Gauss had 41 data points to predict where Ceres would reappear



• Piazzi's 24 observations
41 nights, Jan 1 – Feb 11, 1801

· · · Lost behind the sun
17 months unobserved

★ Gauss's prediction
Dec 7, 1801 — accurate to < 0.5

Minimize the sum of squared deviations between observed positions and predicted orbit

Legendre published the method first — Gauss said he had it at 18

1795 Gauss, age 18: develops the method — privately

1801 Gauss applies it to predict Ceres

1805 Legendre: first to publish in *Nouvelles méthodes…*

1809 Gauss publishes *Theoria Motus* — claims use since 1795

Legendre was furious — Gauss did not back down

Priority disputes are as old as the methods themselves

Gauss read the data in a journal — his priority claim was one sentence

- **How did Piazzi fit in?**
  Piazzi's 24 observations were published in von Zach's *Monatliche Correspondenz* (Sept. 1801). Gauss read them there. He never touched a telescope — purely a desk calculation.

- **What did Gauss say he used it for at 18?**
  One sentence in *Theoria Motus* (1809): "Our principle, which we have made use of since 1795, has lately been published by Legendre." No document from 1795 survives. He did not say what he applied it to.

- **Did he make a big deal of it?**
  No — the claim was casual, almost offhand. Legendre wrote a pained private letter. Then in 1820 published a scathing attack, accusing Gauss of "appropriating the discoveries of others."

Gauss likely had it first; Legendre published it first — the math doesn't care either way

- **On Gauss**: Stigler (1981, *Ann. Statist.*) examined the notebooks carefully. Verdict: Gauss probably told the truth.
  His pattern across a career — non-Euclidean geometry, elliptic functions, the FFT — is hoarding discoveries, not fabricating them. *Pauca sed matura*: few but ripe.

- **On Legendre**: He did what Gauss would not — wrote it down clearly and published it.
  Without that, the method stays in a desk drawer. Publication is how science becomes science.

- **On the dispute**: It changed nothing about the mathematics. Least squares is true regardless of who had it first.
  The goal is always facts about the world. Ego is the tax we pay on getting there.

You already know this as OLS — today we name all three things it is

- **Estimand** $\alpha + \beta X$ the BLP — what we want from the population
- **Estimator** OLS procedure the algorithm that recovers it from data
- **Estimate** $\hat{\alpha}$, $\hat{\beta}$ the specific numbers from your sample

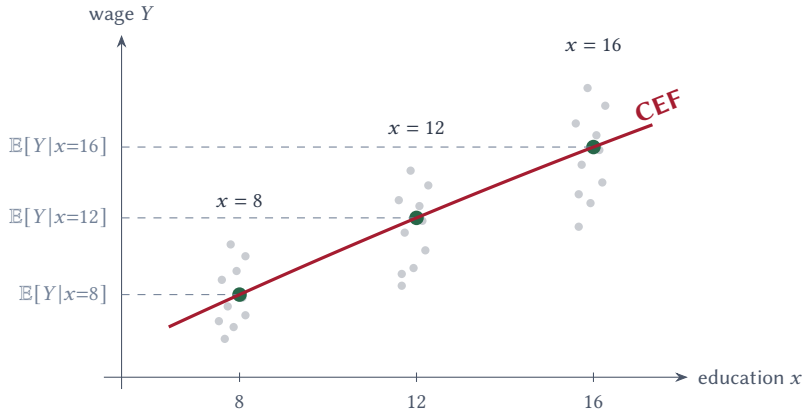
Most researchers identify OLS with only the estimate — today we firm up all three

Behind every question about group averages is a conditional expectation

Field	Research question
Political science	Among voters contacted by a GOTV campaign, what is average turnout?
Economics	Among workers with 16 years of education, what is average hourly wage?
Sociology	Among children of non-graduates, what is average years of schooling?
Industry	Among users who clicked an ad, what is the average purchase amount?

Each is asking: what is $\mathbb{E}[Y \mid X = x]$ for some outcome Y and conditioning value x?

For each x the CEF returns one number; across all x it traces a curve



Each green dot is one value of the CEF at one x — **the curve connecting them is the CEF itself**

The CEF is the curve; the curve is the estimand

Scalar estimands (before)

- Object: one number — μ, σ^2, β
- What moves: estimates $\hat{\theta}$ across samples
- Why: different samples → different units
- Population truth: one fixed number

The CEF (new)

- Object: a function — $x \mapsto \mathbb{E}[Y \mid X = x]$
- What moves: output as x changes
- Why: different x → different subpopulation
- Population truth: fixed at every x

What varies is not the sample — it is which population slice we condition on

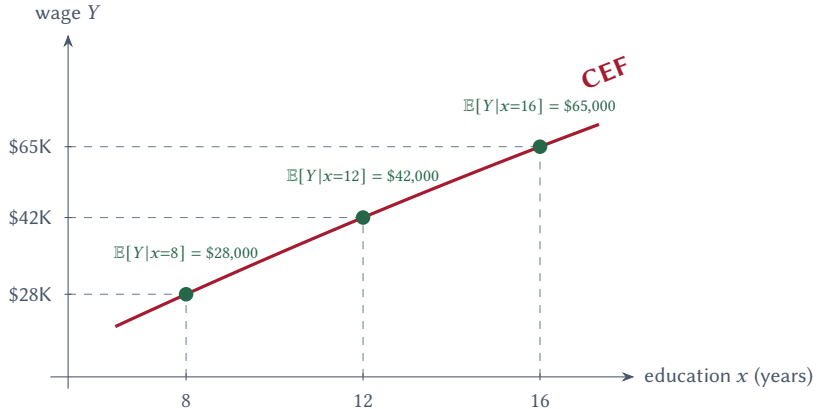
The CEF is a population function: one fixed number out for every x in

- **Input:** a specific value x (e.g., 12 years of education)
- **Output:** $\mathbb{E}[Y \mid X = x]$ — the mean of Y among all population units where $X = x$
- **Fixed:** each output is a deterministic population quantity, not a random variable
- **Function:** different inputs return different outputs — that is why it traces a curve

$$\mathbb{E}[Y \mid X = x] = \int y \, f_{Y|X}(y \mid x) \, dy$$

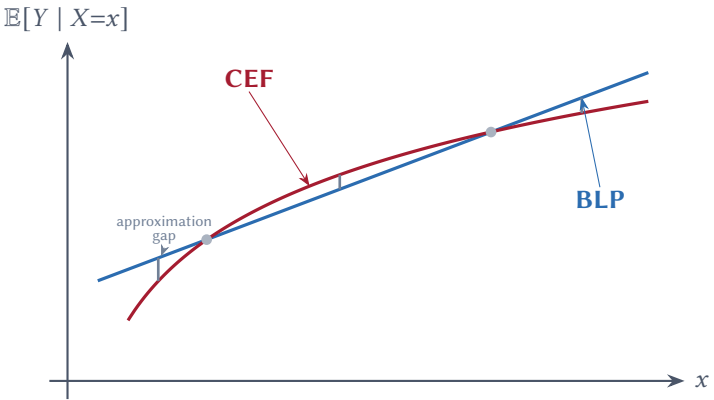
A weighted average over the population distribution at x — no sample, no randomness

Every point on the curve is a population number



Data means sample (a draw of units); *population* means everyone — these values come from the population, so no data yet

The BLP is the line closest to the CEF — the gap is unavoidable



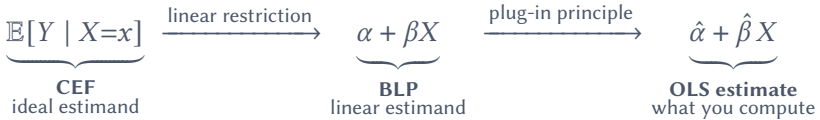
Vertical solid lines: approximation error — the BLP minimizes the expected squared gap from the CEF over the distribution of X

The CEF and the BLP are both population objects — neither is in your data

- **CEF** $\mathbb{E}[Y \mid X = x]$
  Population function — one fixed number for every x in the support of X
  The same kind of object as μ or σ^2: defined in the population, unknown without it

- **BLP** $\alpha + \beta X$
  Population regression — the OLS you would run on the entire population if you had it
  Also unknown for the same reason: you don't have the population
  Only difference from CEF: linear, so just two numbers (α, β) instead of a whole function

OLS is the sample analog of the BLP — it does in your data what BLP does in the population



CEF

Population object
Unknown without
the full population

BLP

Population object
Unknown without
all population units

OLS

Sample procedure
Recovers the BLP
from your data

Two approximations: the BLP linearizes the CEF; OLS estimates the BLP — regression is always an approximation at both steps

Two families of estimators — one targets the BLP, one targets the CEF

Linear estimators

target: BLP

- **OLS** — minimize $\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$
  BLUE under homoskedasticity (Gauss–Markov)
- **WLS / GLS** — minimize a *weighted* sum
  BLUE under heteroskedasticity or autocorrelation

Both recover (α, β) — just two numbers

BLUE = Best Linear Unbiased Estimator

Nonparametric estimators

target: CEF

- **LOWESS / local polynomial** — fit local lines near each x
  weighted regression in a moving window
- **Kernel regression** — weighted average of nearby Y_i
  weights decline with distance from x

Recover $\mathbb{E}[Y \mid X{=}x]$ at every x — a whole function

Parametric means two numbers; nonparametric means the data decides the shape — today: CEF → BLP → OLS

Parametric

- Finite number of parameters — here, just (α, β)
- Form fixed in advance: linear
- Fast: variance scales with $1/n$ — double your sample, variance is cut in half

Nonparametric

- No fixed form — shape determined by the data
- Consistent, but convergence slows as X gets richer
- Curse of dimensionality hits the *estimator*

Today: $\underbrace{\mathbb{E}[Y \mid X{=}x]}_{\textbf{CEF}} \longrightarrow \underbrace{\alpha + \beta X}_{\textbf{BLP}} \longrightarrow \underbrace{\hat{\alpha} + \hat{\beta} X}_{\textbf{OLS}}$

"Regression" names both the estimand and the estimator — they are not the same thing

Usage	Object	Level
"The regression" (pop.)	BLP: $\alpha + \beta X$	Estimand — lives in the distribution
"Running a regression"	OLS: $\hat{\alpha} + \hat{\beta} X$	Estimator — applied to data
"My regression coefficient"	$\hat{\beta}$	Estimate — one number from one sample

CEF → BLP → OLS

ideal estimand → linear estimand → estimator

Angrist & Pischke call the BLP "the population regression" — MHE Ch. 3

OLS will be the most-used model in your empirical career

- After simple averages, OLS is the most common procedure in empirical social science
- And the simple difference-in-means *is* OLS — regress Y on a dummy for treatment
- Today we separate the target from the procedure: CEF and BLP are population objects
- OLS is the estimator — the sample procedure that recovers them

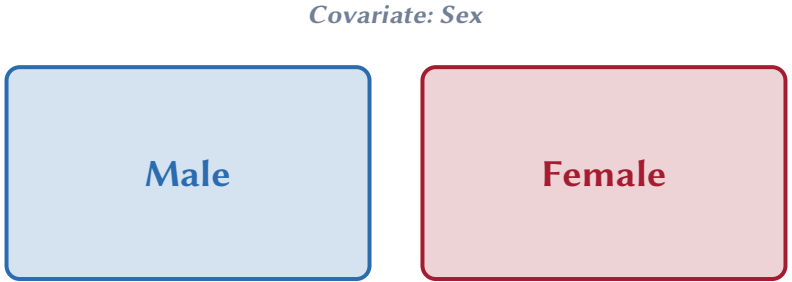
CEF and BLP live in the population; OLS lives in your data

The CEF is a function-valued estimand — the BLP collapses it to two numbers

- **Discrete** X**:** one population mean per value of X — as many quantities as X has values
- **Continuous** X**:** one mean for every point in the support of X — a function, not a number
- **Multiple** X**:** one mean for every combination — a surface over the joint support of X

The BLP restricts to linear functions: α and β replace the entire surface

One covariate partitions the population into 2 cells



2 **cells** $= 2^1$

CEF assigns one $\mathbb{E}[Y \mid \text{Sex}{=}x]$ to each cell — two cells, two population means

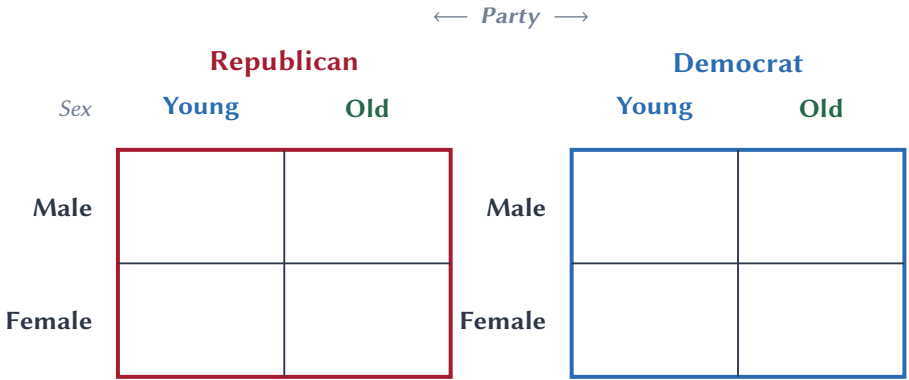
A second covariate doubles the cells — 2 becomes 4

Sex \ *Age*	**Young**	**Old**
Male	**Male × Young**	**Male × Old**
Female	**Female × Young**	**Female × Old**

4 cells $= 2^2$

2 covariates → 4 cells — each new covariate multiplies the cell count by 2

A third covariate doubles again — 4 becomes 8



8 cells = 2^3

k binary covariates → 2^k cells — one population mean per cell

The CEF is defined in 2^k cells — but finite samples cannot fill them all

Population: CEF has 2^k values

Covariates (k)	CEF values (2^k)
1	2
2	4
5	32
10	1,024
20	1,048,576

Each $\mathbb{E}[Y \mid X{=}x]$ exists in the population — not missing, just unknown without all pop. units

Sample ($n = 1{,}000$): avg. obs. per cell

Covariates (k)	Obs. per cell
1	500
2	250
5	31
10	<1
20	≈ 0

Population means exist — the sample just has no observations there

BLP: same two parameters (α, β) regardless of how many covariates

The BLP minimizes mean squared prediction error over all linear functions

$$(\alpha, \beta) = \arg\min_{a,b} \mathbb{E}\left[(Y - a - bX)^2\right]$$

CEF

- Best predictor, *any* function of X
- Minimizes MSE globally

BLP

- Best predictor, *linear* functions only
- Minimizes MSE within $\{\alpha + \beta X\}$

Residuals use $(\hat{\alpha}, \hat{\beta})$ from your data — projection errors use the true (α, β)

Residual sample

What it is: gap between Y_i and the OLS fitted line

$$\hat{e}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$$

- $\hat{\alpha}, \hat{\beta}$ estimated from your n observations
- **Computable:** one number per data point
- Changes if you draw a different sample

Projection error population

What it is: gap between Y and the BLP at the true (α, β)

$$U = Y - (\alpha + \beta X)$$

- α, β are population constants — not estimated
- **Not computable:** you never know the true (α, β)
- Same for every sample drawn from the population

$\hat{e}_i$ estimates U_i — as $n \to \infty$, OLS residuals converge to projection errors

The FOC for α places the line through the population means

From the FOC:

$$\frac{\partial}{\partial \alpha} \mathbb{E}\left[(Y - \alpha - \beta X)^2\right] = -2\ \mathbb{E}[Y - \alpha - \beta X] = 0$$

Solving:

$$\mathbb{E}[Y] - \alpha - \beta\ \mathbb{E}[X] = 0$$

$$\alpha = \mathbb{E}[Y] - \beta\ \mathbb{E}[X]$$

The BLP always passes through $\big(\mathbb{E}[X],\ \mathbb{E}[Y]\big)$

The FOC for β yields the covariance-to-variance ratio

From the FOC:

$$\mathbb{E}[X(Y - \alpha - \beta X)] = 0$$

Substitute $\alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[X]$:

$$\mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \beta\big(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\big)$$

$$\text{Cov}(X, Y) = \beta\ \text{Var}(X)$$

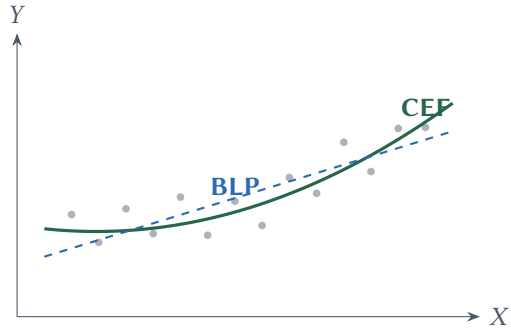
$$\beta = \frac{\text{Cov}(X,\ Y)}{\text{Var}(X)}$$

The BLP has a closed-form solution in population moments

$$\beta = \frac{\text{Cov}(X,\ Y)}{\text{Var}(X)} \qquad \alpha = \mathbb{E}[Y] - \beta\ \mathbb{E}[X]$$

- $\beta > 0$ iff X and Y move together
- $\beta = 0$ iff X and Y are uncorrelated
- Only four moments needed: $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\text{Var}(X)$, $\text{Cov}(X, Y)$

The BLP is the best linear summary of a possibly nonlinear CEF



BLP minimizes $\mathbb{E}[(\mathbb{E}[Y|X] - \alpha - \beta X)^2]$ — closest line to the CEF in MSE

Three theorems tell us when and why regression makes sense

3.1.1 Linear CEF: $\mathbb{E}[Y \mid X]$ linear in $X \implies$ BLP = CEF exactly

3.1.2 Best predictor: BLP minimizes $\mathbb{E}[(Y - a - bX)^2]$ among all (a, b) — always

3.1.3 Best approximation: BLP minimizes $\mathbb{E}[(\mathbb{E}[Y \mid X] - a - bX)^2]$ — always

3.1.3 is the key: the world need not be linear for regression to make sense

Theorem 3.1.1 — a linear CEF means regression is exact, not approximate

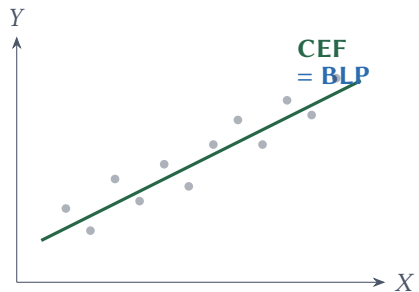
Statement: If $\mathbb{E}[Y \mid X] = \alpha + \beta X$, then BLP = CEF.

Proof: Plug $\mathbb{E}[Y \mid X] = \alpha + \beta X$ into the BLP formula:

$$\beta_{\text{BLP}} = \frac{\text{Cov}(X,\ \alpha + \beta X)}{\text{Var}(X)} = \frac{\beta\ \text{Var}(X)}{\text{Var}(X)} = \beta$$

$$\alpha_{\text{BLP}} = \underbrace{(\alpha + \beta\,\mathbb{E}[X])}_{\mathbb{E}[Y]} - \beta_{\text{BLP}}\ \mathbb{E}[X] = \alpha$$

Same (α, β) — OLS recovers the CEF exactly, not as an approximation



Theorem 3.1.2 — BLP minimizes mean-squared prediction error, always

Problem: $\min_{a,\,b} \mathbb{E}\big[(Y - a - bX)^2\big]$

First-order conditions:

$$\frac{\partial}{\partial a} = 0 \;\Rightarrow\; \mathbb{E}[Y - a - bX] = 0 \;\Rightarrow\; a = \mathbb{E}[Y] - b\,\mathbb{E}[X]$$

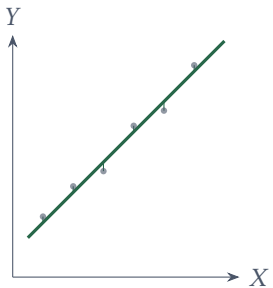
$$\frac{\partial}{\partial b} = 0 \;\Rightarrow\; \mathbb{E}\big[X(Y - a - bX)\big] = 0 \;\Rightarrow\; \mathrm{Cov}(X, Y) = b\,\mathrm{Var}(X)$$

$$\beta = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \qquad \alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[X]$$

No assumption on the shape of $\mathbb{E}[Y \mid X]$ — holds for any joint distribution of (X, Y)

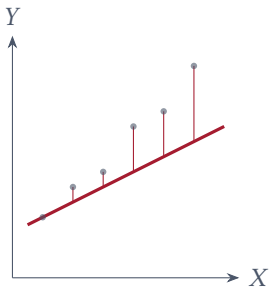
No other line fits as well — the BLP uniquely minimizes squared prediction error

BLP: $\hat{a} = \alpha,\ \hat{b} = \beta$



MSE = 0.06

Any other line: $a \neq \alpha$ **or** $b \neq \beta$



MSE = 1.88

Same six points — the BLP is the unique minimizer of $\mathbb{E}[(Y - a - bX)^2]$

Theorem 3.1.3 — BLP is the closest line to the CEF in mean-squared distance

Problem: $\min_{a,b} \mathbb{E}\left[(\mathbb{E}[Y \mid X] - a - bX)^2\right]$

Same FOC as 3.1.2 with $\mathbb{E}[Y|X]$ replacing Y:

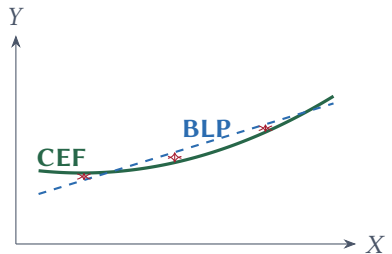
$$\beta = \frac{\text{Cov}(X, \mathbb{E}[Y \mid X])}{\text{Var}(X)}$$

Key step: $\text{Cov}(X, \mathbb{E}[Y \mid X]) = \text{Cov}(X, Y)$

LIE: $\mathbb{E}\left[X\ \mathbb{E}[Y|X]\right] = \mathbb{E}[XY]$

Same (α, β) as Theorem 3.1.2

The world need not be linear for OLS to make sense



BLP minimizes the squared red gaps

BLP residuals are uncorrelated with X by construction

Define the BLP residual: $u = Y - \alpha - \beta X$

From the FOC for β:

$$\mathbb{E}[Xu] = 0 \quad \Rightarrow \quad \text{Cov}(u, X) = \mathbb{E}[Xu] - \mathbb{E}[u]\,\mathbb{E}[X] = 0$$

BLP residual u

- $\text{Cov}(u, X) = 0$
- Uncorrelated with X

CEF residual ε

- $\mathbb{E}[\varepsilon \mid X] = 0$
- Mean-independent of X

Mean independence $\Rightarrow$ uncorrelatedness — not vice versa

OLS is the sample analog of the BLP

Population BLP

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\alpha = \mathbb{E}[Y] - \beta \mathbb{E}[X]$$

Sample OLS

$$\hat{\beta} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

Replace population moments with sample moments — that is all

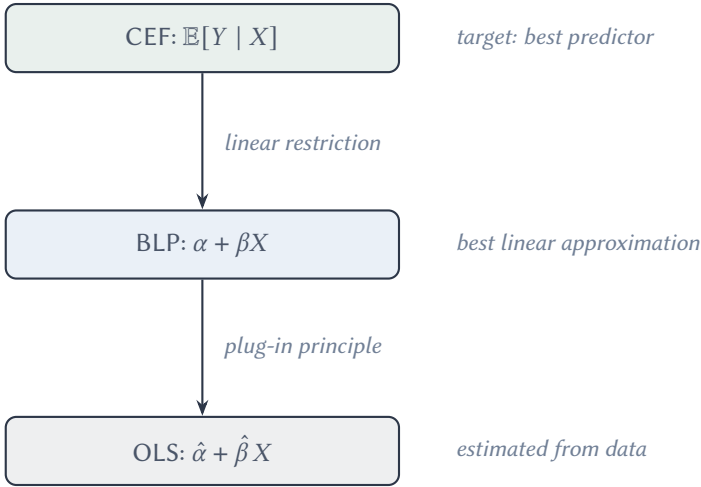
By the LLN, sample moments converge to population moments $\Rightarrow \hat{\beta} \xrightarrow{p} \beta$

The OLS slope has a closed form in terms of sample deviations

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

- Numerator: sample covariation of X and Y
- Denominator: sample variation of X
- $\hat{\beta} = 0$ iff X and Y are sample-uncorrelated

CEF, BLP, and OLS form a complete chain from target to estimator



BLP and OLS give identical answers

```r
library(wooldridge)
data(wage1)

# BLP slope by hand: Cov(X,Y) / Var(X)
beta_blp <- cov(wage1$educ, wage1$wage) / var(wage1$educ)
alpha_blp <- mean(wage1$wage) - beta_blp * mean(wage1$educ)

# OLS via lm()
ols <- lm(wage ~ educ, data = wage1)

# Compare
c(alpha_blp, beta_blp) # -0.905  0.541
coef(ols)              # -0.905  0.541
```

They match — OLS is the sample BLP, nothing more and nothing less

Regression makes sense because it is always approximating something real

1. $\mathbb{E}[Y \mid X]$ is the target — best predictor of Y given X
2. BLP = $\arg\min_{a,b} \mathbb{E}[(Y - a - bX)^2]$ — two moments, closed form
3. $\beta = \text{Cov}(X, Y)/\text{Var}(X)$ — ratio of covariance to variance
4. BLP = best linear approximation to CEF, even when CEF is nonlinear (MHE 3.1.3)
5. BLP residual: $\text{Cov}(u, X) = 0$ by construction
6. OLS = sample BLP — consistent by LLN

Wednesday: from the BLP formula to the SSR criterion

- SSR minimization gives the same $\hat{\beta}$ and $\hat{\alpha}$ — why?
- Fitted values $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ and residuals $\hat{u}_i = Y_i - \hat{Y}_i$
- Key sample properties of OLS residuals
- Variance decomposition: TSS = ESS + RSS
- R^2: fraction of variation in Y explained by the line

Reading: Blackwell Ch. 5 (continue); A&M §2.2.4