EL COLEGIO DE MÉXICO

CEE 60 AÑOS 1964-2024 CENTRO DE ESTUDIOS ECONÓMICOS

CENTRO DE ESTUDIOS ECONÓMICOS

Maestría en Economía 2024–2026

Microeconometrics for Evaluation

4 Regression with Controls I

Disclaimer: I AM NOT the original intellectual author of the material presented in these notes. The content is STRONGLY based on a combination of lecture notes (from Aurora Ramirez), textbook references, and personal annotations for learning purposes. Any errors or omissions are entirely my own responsibility.

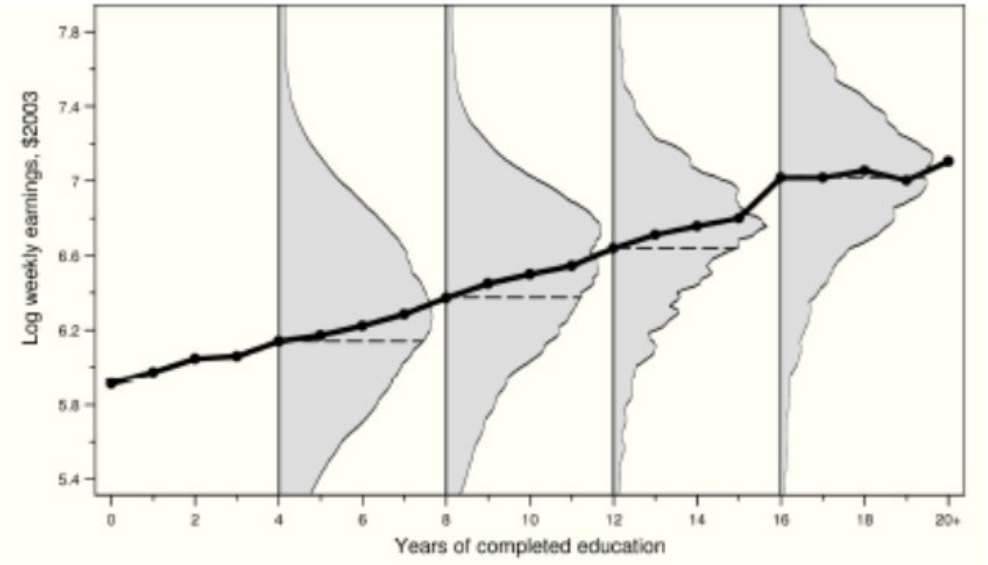
El Colegio de México
Jose Daniel Fuentes García
Microeconometrics for Evaluation
Github : danifuentesga

Índice

The Conditional Expectation Function (CEF)

- The **Conditional Expectation Function** for a dependent outcome Y_i, given a $K \times 1$ vector of covariates X_i (with entries x_{ik}), is written $E[Y_i|X_i]$ and is a function of X_i.
- Since X_i is random, the CEF is also random. For a dummy D_i, the CEF has two values: $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$.
- For a specific X_i, say $X_i = 42$, we denote it as $E[Y_i|X_i = 42]$.
- For continuous Y_i with density $f_Y(\cdot|X_i = x)$, the CEF is $E[Y_i|X_i = x] = \int t f_Y(t|X_i = x)dt$.
- If Y_i is discrete, $E[Y_i|X_i = x]$ equals $\sum_t t f_Y(t|X_i = x)$.
- The CEF residual is uncorrelated with any function of X_i. Let $\epsilon_i = Y_i - E[Y_i|X_i]$. For any function $h(x)$, we have $E[\epsilon_i h(X_i)] = E[(Y_i - E[Y_i|X_i])h(X_i)] = 0$.
- **Intuition:** The CEF is the "best guess" of Y given X. It averages possible Y values conditional on X, whether continuous or discrete, and ensures that prediction errors are unrelated to X itself.

CEF of Wages and Education



Raw data and the **CEF** of mean log weekly earnings by schooling. The sample consists of white males aged 40–49 in the 1990 IPUMS 5 % file.

- The plot shows the link between **years of education** and average *log weekly wages*.
- Each dot traces the **conditional expectation** of earnings for a given education level.

- Shaded areas represent the *distribution* of wages within each schooling group.
- The figure highlights that wages rise steadily with schooling, with **plateaus** around high school and college completion.
- **Intuition:** More education generally leads to higher wages, but gains are not uniform—big jumps often occur when completing key degrees (e.g., finishing high school or college).

Population Regression

Define population regression (from here onwards regression) as the solution to the population least squares problem. Specifically, the K × 1 regression coefficient vector beta is defined by solving

$$\beta = \arg\min_b \mathbb{E}\Big[(Y_i - X_i'b)^2\Big] \quad \textbf{Definición del problema de mínimos cuadrados} \tag{1}$$

$$\begin{aligned} L(b) &= \mathbb{E}\Big[(Y_i - X_i'b)^2\Big] && \textbf{Función de pérdida} && (2)\\ &= \mathbb{E}[Y_i^2 - 2Y_iX_i'b + (X_i'b)^2] && \textbf{Expansión del cuadrado} && (3)\end{aligned}$$

$$\begin{aligned} \frac{\partial L(b)}{\partial b} &= \frac{\partial}{\partial b}\Big(\mathbb{E}[Y_i^2] - 2\mathbb{E}[Y_iX_i'b] + \mathbb{E}[(X_i'b)^2]\Big) && \textbf{Derivamos con respecto a } b && (4)\\ &= -2\mathbb{E}[X_iY_i] + 2\mathbb{E}[X_iX_i']b && \textbf{Aplicamos reglas de derivación matricial} && (5)\end{aligned}$$

$$\text{FOC:} \quad -2\mathbb{E}[X_iY_i] + 2\mathbb{E}[X_iX_i']b = 0 \quad \textbf{Condición de primer orden} \tag{6}$$

$$\mathbb{E}[X_iX_i']b = \mathbb{E}[X_iY_i] \quad \textbf{Reordenamos} \tag{7}$$

$$\begin{aligned} b &= \Big(\mathbb{E}[X_iX_i']\Big)^{-1}\mathbb{E}[X_iY_i] && \textbf{Solución explícita para } b && (8)\\ \beta &= b && \textbf{Identificación del estimador poblacional} && (9)\end{aligned}$$

$$\begin{aligned} \epsilon_i &= Y_i - X_i'\beta && \textbf{Definimos residual poblacional} && (10)\\ \mathbb{E}[X_i\epsilon_i] &= \mathbb{E}[X_i(Y_i - X_i'\beta)] = 0 && \textbf{Propiedad de ortogonalidad} && (11)\end{aligned}$$

- **Intuition:** Population regression chooses β that minimizes the expected squared error between actual Y and its linear prediction $X\beta$.
- The solution comes from the first-order condition: errors must be *uncorrelated* with regressors.
- In simple words: regression picks coefficients so that on average, X carries no leftover information about residuals.

Why Regression? Three Reasons

- Regression solves the population least squares problem and is therefore the **Best Linear Predictor (BLP)** of Y_i given X_i.
- If the Conditional Expectation Function (CEF) is linear, then regression coincides with it.
- Regression provides the **best linear approximation** to the CEF in general.

The first statement is by definition. The second follows from the orthogonality of the CEF. The third point is formalized in the following theorem:

The Regression-CEF Theorem (MHE 3.1.6)

Regression-CEF Theorem (MHE 3.1.6) The population regression function $X_i'\beta$ delivers the *Minimum Mean Squared Error (MMSE)* linear approximation to $E[Y_i|X_i]$, that is:

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

- **Intuition:** Regression works because it either **exactly equals the CEF** when the relation is linear, or otherwise it finds the **closest linear fit** to the true conditional expectation in terms of mean squared error. In plain words: regression is the best linear "shortcut" to describe how Y depends on X.

Proof

$$\beta = \arg\min_b \mathbb{E}\Big[(Y_i - X_i'b)^2\Big] \qquad \textbf{Starting point} \tag{12}$$

$$Y_i - X_i'b = \Big(Y_i - \mathbb{E}[Y_i|X_i]\Big) + \Big(\mathbb{E}[Y_i|X_i] - X_i'b\Big) \qquad \textbf{Add and subtract } E[Y_i|X_i] \tag{13}$$

$$(Y_i - X_i'b)^2 = \Big(\big(Y_i - \mathbb{E}[Y_i|X_i]\big) + \big(\mathbb{E}[Y_i|X_i] - X_i'b\big)\Big)^2 \qquad \textbf{Square both sides} \tag{14}$$

$$= (Y_i - \mathbb{E}[Y_i|X_i])^2 + 2(Y_i - \mathbb{E}[Y_i|X_i])(\mathbb{E}[Y_i|X_i] - X_i'b) \tag{15}$$
$$+ (\mathbb{E}[Y_i|X_i] - X_i'b)^2 \qquad \textbf{Expansion of square} \tag{16}$$

$$\mathbb{E}[(Y_i - X_i'b)^2] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])^2] + 2\,\mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])(\mathbb{E}[Y_i|X_i] - X_i'b)] \tag{17}$$
$$+ \mathbb{E}[(\mathbb{E}[Y_i|X_i] - X_i'b)^2] \qquad \textbf{Take expectation} \tag{18}$$

$$\mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])^2] = \mathrm{Var}(Y_i|X_i) \qquad \textbf{Does not depend on } b \tag{19}$$

$$\mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])(\mathbb{E}[Y_i|X_i] - X_i'b)] = \mathbb{E}\Big[\mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])(\mathbb{E}[Y_i|X_i] - X_i'b) \mid X_i]\Big] \tag{20}$$

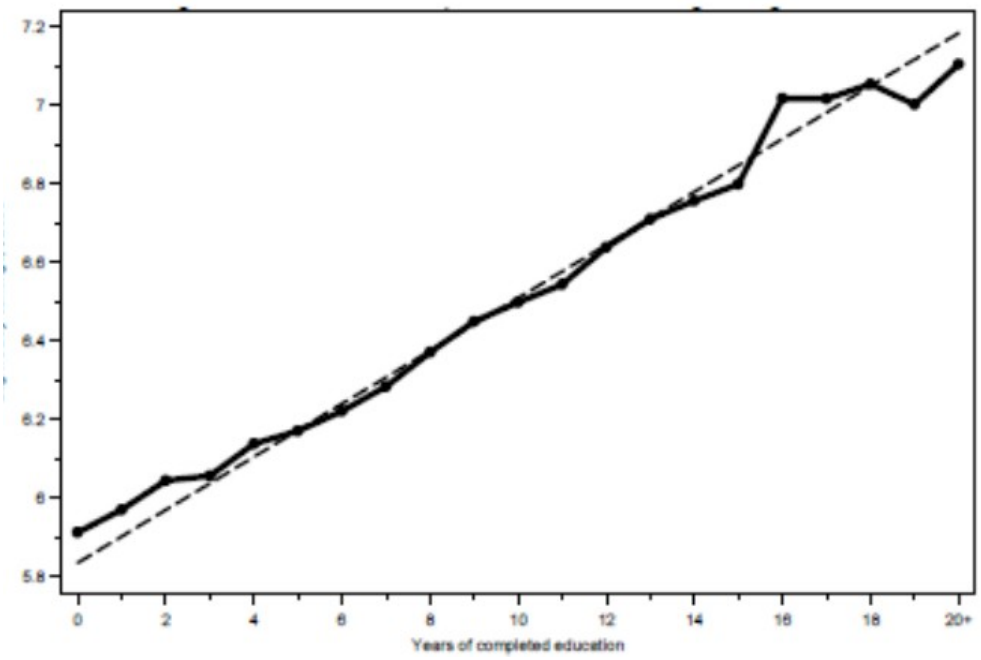
$$= \mathbb{E}\Big[(\mathbb{E}[Y_i|X_i] - X_i'b)\,\mathbb{E}[Y_i - \mathbb{E}[Y_i|X_i] \mid X_i]\Big] \qquad \textbf{Law of iterated expectations} \tag{21}$$

$$= \mathbb{E}\Big[(\mathbb{E}[Y_i|X_i] - X_i'b)\cdot 0\Big] = 0 \qquad \textbf{Property: residual has mean zero} \tag{22}$$

$$\mathbb{E}[(Y_i - X_i'b)^2] = \underbrace{\mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])^2]}_{\text{irrelevant for } b} + \mathbb{E}[(\mathbb{E}[Y_i|X_i] - X_i'b)^2] \tag{23}$$

$$\beta = \arg\min_b \mathbb{E}[(\mathbb{E}[Y_i|X_i] - X_i'b)^2] \qquad \textbf{Final result} \tag{24}$$

- **Intuition:** By adding and subtracting $E[Y|X]$, the error splits into two parts: (i) the variance around the conditional mean (independent of b), and (ii) the squared distance between the conditional mean and the linear predictor Xb. Minimizing the full expression is thus the same as making Xb the **closest linear approximation** to $E[Y|X]$ in mean squared error.



Regression line (dashed) compared with the **CEF** of mean log weekly earnings by years of schooling. Sample restricted to white men aged 40–49, from the 1980 Census IPUMS 5 % file.

- The solid curve shows the **empirical CEF** of wages by education level.
- The dashed line is the **linear regression fit** to that CEF.
- The regression line smooths out local fluctuations and provides a simple linear summary of the relationship.
- Although the true CEF is not perfectly linear, the regression still captures the main upward trend in wages with education.
- **Intuition:** The regression acts like a "straight thread" through the curved conditional expectation. It cannot capture every bend, but it summarizes the overall slope: more schooling is strongly associated with higher wages on average.

The CEF is All You Need

- The Regression–CEF Theorem shows that we can use $\mathbb{E}[Y_i|X_i]$ directly as the dependent variable instead of Y_i (though the weighting is important).
- Another way to express this result is:

$$\beta = \left(\mathbb{E}[X_i X_i']\right)^{-1}\mathbb{E}[X_i Y_i] \qquad \textbf{Population regression formula} \qquad (25)$$

$$= \left(\mathbb{E}[X_i X_i']\right)^{-1}\mathbb{E}[X_i\, \mathbb{E}[Y_i|X_i]] \qquad \textbf{Replace } Y_i \textbf{ with its CEF} \qquad (26)$$

- This equivalence implies that grouped–data regression (using averages of Y_i within X_i categories) is valid when micro–data cannot be analyzed.
- For example, the schooling coefficient in a wage equation can be estimated using conditional means of log earnings across 21 education levels.
- In practice, if we weight groups by the number of individuals at each level, the regression based on grouped data yields the same coefficients as regression on micro–data.
- **Intuition:** The CEF contains all the relevant predictive information about Y given X. Whether you regress Y directly or its conditional expectation, you recover the same β — as long as weights are respected. In plain words: you can "compress" the data into averages without losing information about the slope.

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

Comparison of regression results:
(A) Individual-level data with robust errors,
(B) Grouped means by schooling with weights.

- Panel A shows the regression of **individual earnings** on years of schooling using the full micro–sample ($N \approx 409{,}000$).
- The estimated coefficient on schooling is about 0,067, with extremely high statistical significance.
- Panel B repeats the regression using **grouped averages** of earnings by schooling level ($N = 21$ groups), weighting by group size.
- The estimated schooling coefficient is essentially the same (0,067), despite working with far fewer observations.
- This confirms the earlier theorem: regression using grouped CEF data produces coefficients identical to those from the underlying micro–data, provided correct weighting is applied.
- **Intuition:** Whether we use millions of individuals or just a handful of averages, the slope is the same because the CEF already condenses all the information we need. Grouped data is like a "compressed file" — smaller in size but containing the same essential signal about the schooling–wage relationship.

Saturated Models and Main Effects

- A **saturated regression model** is one where the explanatory variable is discrete, and the regression includes a separate parameter for each possible value of that variable.
- In other words, the model perfectly matches group means by allowing one coefficient per category.

Example. Suppose schooling $s_i \in \{0, 1, 2, \ldots, \tau\}$. A saturated regression model is:

$$Y_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + \cdots + \beta_\tau d_{\tau i} + \epsilon_i \tag{27}$$

where $d_{ji} = 1[s_i = j]$ is a dummy for being in schooling level j. Here, β_j represents the effect of schooling level j, and α is the reference category mean.

$$\beta_j = \mathbb{E}[Y_i|s_i = j] - \mathbb{E}[Y_i|s_i = 0] \quad \textbf{Difference in group means} \tag{28}$$

$$\alpha = \mathbb{E}[Y_i|s_i = 0] \quad \textbf{Reference group mean} \tag{29}$$

- **Intuition:** A saturated regression is basically the same as comparing group averages. Each coefficient β_j tells us how much higher or lower group j is compared to the baseline group. Nothing is approximated — the model exactly fits the sample means for each category.

Saturated Models II: Two Binary Regressors

- Let x_{1i} indicate whether an individual is a **college graduate**.
- Let x_{2i} indicate whether an individual is **female**.
- The CEF conditional on (x_{1i}, x_{2i}) has four possible values:

$$\mathbb{E}[Y_i \mid x_{1i} = 0, x_{2i} = 0] = \alpha \quad \text{baseline: no college, male} \tag{30}$$

$$\mathbb{E}[Y_i \mid x_{1i} = 1, x_{2i} = 0] = \alpha + \beta_1 \quad \text{effect of college, men only} \tag{31}$$

$$\mathbb{E}[Y_i \mid x_{1i} = 0, x_{2i} = 1] = \alpha + \gamma \quad \text{effect of being female, no college} \tag{32}$$

$$\mathbb{E}[Y_i \mid x_{1i} = 1, x_{2i} = 1] = \alpha + \beta_1 + \gamma + \delta_1 \quad \text{college + female plus interaction} \tag{33}$$

These four cases can be summarized with a saturated regression:

$$\mathbb{E}[Y_i \mid x_{1i}, x_{2i}] = \alpha + \beta_1 x_{1i} + \gamma x_{2i} + \delta_1 (x_{1i} \cdot x_{2i}) \tag{34}$$

- The model contains two **main effects** (β_1 for college, γ for female) and one **interaction** (δ_1 for "female college graduate").
- It is generally odd to include the interaction term without also including the main effects, since the meaning of the interaction depends on them.
- **Intuition:** This regression just assigns an average outcome to each of the four groups (male/non-college, male/college, female/non-college, female/college). The interaction term δ_1 captures whether the combined effect of being both female and college-educated is more (or less) than the sum of the individual effects.

Regression Anatomy: Coefficients as Partial Slopes

Step 1: Bivariate regression recap

$$\beta_1 = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)} \quad \text{Slope coefficient in simple regression} \tag{35}$$

$$\alpha = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i] \quad \text{Intercept from mean restriction} \tag{36}$$

Step 2: Multivariate regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \tag{37}$$

We want the general formula for the coefficient β_k.

Step 3: Residualizing X_{ki}

$$\tilde{X}_{ki} = X_{ki} - \hat{X}_{ki} \qquad \text{Residual from regression of } X_{ki} \text{ on all other } X_{-k} \tag{38}$$

By construction, $\tilde{X}_{ki}$ is orthogonal to all other regressors X_{-k}.

Step 4: Regression anatomy claim

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{X}_{ki})}{\text{Var}(\tilde{X}_{ki})} \qquad \text{Formula for partial slope} \tag{39}$$

Step 5: Verification algebra

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i \qquad \text{Population regression} \tag{40}$$

$$\text{Cov}(Y_i, \tilde{X}_{ki}) = \text{Cov}\Big(\beta_0 + \sum_{j=1}^{K} \beta_j X_{ji} + \epsilon_i,\ \tilde{X}_{ki}\Big) \qquad \text{Take covariance with } \tilde{X}_{ki} \tag{41}$$

$$= \beta_0 \cdot \text{Cov}(1, \tilde{X}_{ki}) + \sum_{j=1}^{K} \beta_j \text{Cov}(X_{ji}, \tilde{X}_{ki}) + \text{Cov}(\epsilon_i, \tilde{X}_{ki}) \tag{42}$$

$$= \beta_k \,\text{Cov}(X_{ki}, \tilde{X}_{ki}) \qquad \text{Only } X_{ki} \text{ survives: others orthogonal by construction, residual uncorrelated with regressors} \tag{43}$$

$$\text{Cov}(X_{ki}, \tilde{X}_{ki}) = \text{Var}(\tilde{X}_{ki}) \qquad \text{Key property of residualization} \tag{44}$$

$$\text{Cov}(Y_i, \tilde{X}_{ki}) = \beta_k \,\text{Var}(\tilde{X}_{ki}) \tag{45}$$

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{X}_{ki})}{\text{Var}(\tilde{X}_{ki})} \quad \blacksquare \tag{46}$$

- **Intuition:** Each regression coefficient β_k is like a simple slope — but only after "partialling out" the influence of other regressors. By residualizing X_{ki} against all other covariates, we isolate its unique variation. Then the coefficient is just the simple regression of Y on this purified residual. In other words: multivariate regression is just bivariate regression, done on the leftover part of each variable.

Omitted Variables Bias (OVB)

The OVB formula shows how regression coefficients differ when some controls are omitted.

Step 1: Long regression (with ability A_i included)

$$Y_i = \alpha + \rho s_i + A_i'\gamma + \epsilon_i \qquad \text{Wages on schooling } s_i \text{ controlling for ability } A_i \tag{47}$$

Step 2: Short regression (omit A_i)

$$Y_i = \alpha^* + \pi s_i + u_i \qquad \text{Coefficient on schooling changes to } \pi \tag{48}$$

Step 3: Expression for short regression slope

$$\pi = \frac{\text{Cov}(Y_i, s_i)}{\text{Var}(s_i)} \qquad \text{Definition of slope in simple regression} \tag{49}$$

Step 4: Substitute Y_i from long regression

$$\begin{aligned} \text{Cov}(Y_i, s_i) &= \text{Cov}(\alpha + \rho s_i + A_i'\gamma + \epsilon_i,\ s_i) && (50)\\ &= \rho\,\text{Cov}(s_i, s_i) + \text{Cov}(A_i'\gamma, s_i) + \text{Cov}(\epsilon_i, s_i) && (51)\\ &= \rho\,\text{Var}(s_i) + \gamma'\,\text{Cov}(A_i, s_i) \qquad \text{Assume } \epsilon_i \text{ uncorrelated with } s_i && (52) \end{aligned}$$

Step 5: Plug back into slope formula

$$\begin{aligned} \pi &= \frac{\rho\,\text{Var}(s_i) + \gamma'\,\text{Cov}(A_i, s_i)}{\text{Var}(s_i)} && (53)\\ &= \rho + \gamma'\frac{\text{Cov}(A_i, s_i)}{\text{Var}(s_i)} && (54) \end{aligned}$$

Step 6: Define δ_{As} as regression of A_i on s_i

$$\delta_{As} = \frac{\text{Cov}(A_i, s_i)}{\text{Var}(s_i)} \qquad \begin{array}{l}\text{Slope vector from regressing} \\ A_i \text{ on } s_i\end{array} \tag{55}$$

Final Result (OVB formula)

$$\pi = \rho + \gamma' \delta_{As} \qquad \begin{array}{l}\text{Short regression = true effect} \\ \text{+ bias from omitted variables}\end{array} \tag{56}$$

- **Intuition:** The coefficient on schooling changes when ability is omitted because part of the schooling effect captures the correlation between schooling and ability.
- If s_i and A_i are uncorrelated, then $\delta_{As} = 0$ and short regression equals the long regression.
- Otherwise, the bias is the product of: (i) the effect of the omitted variable on Y (γ), and (ii) the correlation between the omitted and included regressor (δ_{As}).
- In plain words: omitted variables bias = **effect of the omitted** $\times$ **correlation with the included**.

Table 3.2.1: Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupation dummies
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

Table 3.2.1: Estimates of the returns to education for men in the NLSY (1979 cohort, 2002 survey). The table shows coefficients on years of schooling from regressions of log wages on schooling and various sets of controls. Standard errors are in parentheses. Data are weighted using NLSY sampling weights. Sample size: 2,434. *Additional controls include mother's and father's schooling, plus dummies for race and Census region.

- Column (1): No controls, estimated return $\approx$ 13,2 %.
- Column (2): Adding age dummies barely changes the estimate.
- Column (3): Adding parental background and demographics reduces the estimate to 11,4 %.
- Column (4): Adding AFQT (ability) lowers it further to 8,7 %.
- Column (5): Adding occupation dummies yields about 6,6 %.

- This progression shows how including controls changes the coefficient, illustrating omitted variables bias in action.
- **Intuition:** The "raw" estimate overstates the return to schooling because schooling is correlated with ability, family background, and occupation. As controls are added, the coefficient drops, showing how much of the apparent return to schooling was actually due to omitted factors.

Regression and Causality

- Casual regressions are often exploratory or descriptive.
- Causal regressions are more ambitious: they describe counterfactual states of the world and are key for policy analysis.
- Example: Do private colleges increase earnings? Let C_i indicate private attendance ($C_i = 1$ if private, $C_i = 0$ if not).

Potential outcomes framework:

$$Y_{1i} = \text{Earnings if } C_i = 1 \text{ (private college)} \tag{57}$$
$$Y_{0i} = \text{Earnings if } C_i = 0 \text{ (non-private college)} \tag{58}$$

Observed outcome:

$$Y_i = \begin{cases} Y_{1i}, & \text{if } C_i = 1 \\ Y_{0i}, & \text{if } C_i = 0 \end{cases} \tag{59}$$

This can be rewritten algebraically:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})C_i \qquad \text{Add and subtract } Y_{0i} \tag{60}$$

Define the **individual treatment effect:**

$$\tau_i = Y_{1i} - Y_{0i} \qquad \text{Causal effect for individual } i \tag{61}$$

Then the observed outcome becomes:

$$Y_i = Y_{0i} + \tau_i C_i \qquad \text{Observed outcome = baseline + effect if treated} \tag{62}$$

- We only observe one of $\{Y_{1i}, Y_{0i}\}$ for each individual.
- The parameter of interest is usually an **average treatment effect**.
- Example: the average effect of private college among those who attended:

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid C_i = 1] \quad \text{(Treatment on the Treated, TOT).}$$

- **Intuition:** Causal regressions are about comparing "what happened" with "what would have happened otherwise." We only observe one reality per person, so we rely on assumptions and design to recover average causal effects.

Regression and Causality II: Selection Bias

Step 1: Observed mean difference

$$\Delta = \mathbb{E}[Y_i \mid C_i = 1] - \mathbb{E}[Y_i \mid C_i = 0] \qquad \text{Naive comparison of treated vs untreated} \tag{63}$$

Step 2: Substitute potential outcomes

$$\mathbb{E}[Y_i \mid C_i = 1] = \mathbb{E}[Y_{1i} \mid C_i = 1] \qquad \text{If treated, we see } Y_{1i} \tag{64}$$

$$\mathbb{E}[Y_i \mid C_i = 0] = \mathbb{E}[Y_{0i} \mid C_i = 0] \qquad \text{If untreated, we see } Y_{0i} \tag{65}$$

$$\Delta = \mathbb{E}[Y_{1i} \mid C_i = 1] - \mathbb{E}[Y_{0i} \mid C_i = 0] \tag{66}$$

Step 3: Add and subtract $\mathbb{E}[Y_{0i} \mid C_i = 1]$

$$\Delta = \Big(\mathbb{E}[Y_{1i} \mid C_i = 1] - \mathbb{E}[Y_{0i} \mid C_i = 1]\Big) + \Big(\mathbb{E}[Y_{0i} \mid C_i = 1] - \mathbb{E}[Y_{0i} \mid C_i = 0]\Big) \tag{67}$$

Step 4: Identify components

$$\Delta = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid C_i = 1]}_{\text{Treatment on the Treated (TOT)}} + \underbrace{\Big(\mathbb{E}[Y_{0i} \mid C_i = 1] - \mathbb{E}[Y_{0i} \mid C_i = 0]\Big)}_{\text{Selection Bias}} \tag{68}$$

Final decomposition:

$$\mathbb{E}[Y_i \mid C_i = 1] - \mathbb{E}[Y_i \mid C_i = 0] = \text{TOT} + \text{Selection Bias} \quad \blacksquare \tag{69}$$

- The naive comparison mixes causal effect with differences in baseline potential outcomes.
- If those who choose private college would have earned more anyway, the selection bias term is positive.
- In causal models, **selection bias is the analogue of OVB**.
- The Conditional Independence Assumption (CIA) states that after conditioning on X_i, treatment C_i is independent of potential outcomes:

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp C_i \mid X_i.$$

Under CIA, the selection bias term disappears.

- **Intuition:** Comparing outcomes between treated and untreated is misleading because the groups differ in ways beyond treatment. By adjusting for observables (CIA), we "balance" groups so that the only remaining difference is due to treatment itself.

Using the CIA

Step 1: CIA assumption

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp C_i \mid X_i \qquad \text{Treatment assignment is as good as random within } X_i\text{-cells} \tag{70}$$

Step 2: Conditional mean comparison

$$\mathbb{E}[Y_i \mid X_i, C_i = 1] - \mathbb{E}[Y_i \mid X_i, C_i = 0] = \mathbb{E}[Y_{1i} \mid X_i, C_i = 1] - \mathbb{E}[Y_{0i} \mid X_i, C_i = 0] \tag{71}$$

$$= \mathbb{E}[Y_{1i} \mid X_i] - \mathbb{E}[Y_{0i} \mid X_i] \qquad \text{By CIA, potential outcom do not depend on } C_i \text{ give} \tag{72}$$

$$= \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i] \tag{73}$$

Step 3: Interpretation

$$\mathbb{E}[Y_{1i} - Y_{0i} \mid X_i, C_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i] \qquad \text{Effect for treated group equals effect for all with } X_i \tag{74}$$

Step 4: Marginalize over X_i

$$\mathbb{E}\Big[\, \mathbb{E}[Y_i \mid X_i, C_i = 1] - \mathbb{E}[Y_i \mid X_i, C_i = 0] \,\Big] = \mathbb{E}\Big[\mathbb{E}[Y_{1i} - Y_{0i} \mid X_i]\Big] \tag{75}$$

$$= \mathbb{E}[Y_{1i} - Y_{0i}] \qquad \text{Law of iterated expectations} \tag{76}$$

- Conditional-on-X_i, the comparison of treated vs untreated is unbiased.
- The CIA guarantees that within each X_i, treated and untreated differ only by treatment status.
- Averaging over X_i gives us the overall Average Treatment Effect (ATE).
- **Intuition:** The CIA tells us that people with the same X_i are comparable, regardless of whether they attend private college or not. This motivates methods like *matching*, where we compare treated and untreated individuals with identical or very similar characteristics.

Regression and the CIA

- The CIA allows regression to be interpreted causally.
- To focus on selection, assume a **constant treatment effect**.

Step 1: Define potential outcomes

$$Y_{0i} = \alpha + \eta_i \quad \text{Outcome if untreated with individual noise } \eta_i \tag{77}$$

$$Y_{1i} = Y_{0i} + \rho \quad \text{Outcome if treated = baseline + constant effect } \rho \tag{78}$$

Step 2: Observed outcome rule

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})C_i \quad \text{From potential outcomes see previous slides} \tag{79}$$

$$= Y_{0i} + \rho C_i \quad \text{Substitute } Y_{1i} - Y_{0i} = \rho \tag{80}$$

$$= (\alpha + \eta_i) + \rho C_i \quad \text{Substitute } Y_{0i} = \alpha + \eta_i \tag{81}$$

$$= \alpha + \rho C_i + \eta_i \quad \text{Collect terms: observed equation} \tag{82}$$

Step 3: Interpretation

- Equation above looks like a bivariate regression of Y_i on C_i:

$$Y_i = \alpha + \rho C_i + \eta_i$$

- But it is not an actual regression equation — here C_i may be correlated with the residual η_i (selection into treatment).
- The CIA assumption is what allows us to treat this as a causal regression: under CIA, $C_i \perp \eta_i$, so ρ is consistently estimated.
- **Intuition:** With constant effects, the observed outcome looks just like a regression of Y on C. The slope is the causal effect ρ. But unless assignment is "as good as random" (CIA), C may be correlated with unobserved factors in η_i, biasing the estimate. The CIA guarantees that the regression coefficient recovers the true causal effect.

Regression and the CIA II

Step 1: CIA in constant-effects setup

$$\mathbb{E}[\eta_i \mid C_i, X_i] = \mathbb{E}[\eta_i \mid X_i] \quad \text{CIA implies residual mean independent of treatment } C_i \tag{83}$$

Step 2: Assume linear form for conditional mean

$$\mathbb{E}[\eta_i \mid X_i] = X_i'\gamma \qquad \text{Model the expectation of } \eta_i \text{ as linear in covariates} \tag{84}$$

Step 3: Compute conditional expectation of Y_i

$$\mathbb{E}[Y_i \mid X_i, C_i] = \alpha + \rho C_i + \mathbb{E}[\eta_i \mid X_i] \qquad \text{From earlier: } Y_i = \alpha + \rho C_i + \eta_i \tag{85}$$

$$= \alpha + \rho C_i + X_i'\gamma \tag{86}$$

Step 4: Define regression error

$$Y_i = \alpha + \rho C_i + X_i'\gamma + \nu_i \tag{87}$$

$$\nu_i \equiv \eta_i - X_i'\gamma = \eta_i - \mathbb{E}[\eta_i \mid C_i, X_i] \tag{88}$$

Step 5: Orthogonality condition

$$\mathbb{E}[\nu_i \mid C_i, X_i] = 0 \qquad \text{Error term is uncorrelated with included regressors} \tag{89}$$

- Under CIA, the same ρ that appears in the potential outcomes setup also appears as the regression slope.
- The error ν_i is properly defined so it is orthogonal to both C_i and X_i.
- This establishes the equivalence between regression coefficients and causal effects under CIA.
- **Intuition:** Once we control for X_i, treatment C_i is "as good as random," so the regression picks up the true causal effect ρ. The regression error is just the part of η_i not explained by X_i, and it no longer contaminates the estimate.

The College Matching Matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

The college matching matrix shows applicant groups (A–D), their admission decisions across different types of private and public colleges, and their observed earnings in 1996. Note: Enrollment decisions are highlighted in gray.

- Groups A–D represent different applicant pools, with students admitted or rejected across college types (private vs public).
- Highlighted cells show the actual enrollment choices made by each student.
- Average earnings for private college enrollees exceed those for public enrollees by about $19,500.
- Within matched groups, differences vary: for Group A, private students earn $5,000 less, while for Group B they earn $30,000 more, averaging to about $12,500 overall.
- **Intuition:** Simple comparisons of private vs public students suggest a big payoff, but when we compare students within similar applicant groups (matching on admission profiles), the private advantage shrinks. This illustrates that much of the raw difference comes from selection rather than causal effects.

Private School Effects: Barron's Matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a private-school dummy and controls. Columns (1)–(3) exclude applicant selectivity controls; columns (4)–(6) include them. Standard errors in parentheses. Sample size: 5,583.

- Without selection controls, the private school effect appears positive and large (around 0,135 log points).
- Once controls are added (SAT scores, family background, demographics, high school rank, etc.), the private coefficient drops toward zero (around 0,007–0,013).
- Including applicant selectivity-group dummies (columns 4–6) confirms the result: little or no private school earnings advantage once comparable students are matched.
- The findings align with a **self-revelation model**: applicants have a good idea of their ability, and admissions outcomes largely reflect that, not private schooling per se.

- **Intuition:** At first glance, private colleges seem to pay off. But once we compare students of similar ability and background, the effect mostly vanishes. The apparent private premium was due to selection — stronger students go to private schools, not because private schools themselves cause higher earnings.

Private School Effects: Average SAT Controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a private-school dummy and controls. Columns (1)–(3) exclude selection controls, while columns (4)–(6) add them. Additional controls include average SAT score of schools applied to, number of applications, and background covariates. Sample size: 14,238. Standard errors in parentheses.

- Without selection controls (cols. 1–3), the private school effect appears positive (0.21–0.15 log points).

- With selection controls (cols. 4–6), the private school coefficient shrinks sharply (0.034–0.037, sometimes insignificant).
- The **average SAT score of schools applied to** is highly predictive of earnings (0.07–0.11), showing that application choices reveal ability and ambition.
- Other covariates (family income, gender, race, etc.) have smaller effects compared to these selection variables.
- **Intuition:** The raw private-school premium is mostly selection bias. Students who apply to more competitive (high-SAT) schools already have higher potential earnings. Once we control for this revealed-preference measure, the apparent causal effect of attending a private school essentially disappears.

School Selectivity Effects: Average SAT Controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School average SAT score ÷ 100	.109	.071	.076	−.021	−.031	.000
	(.026)	(.025)	(.016)	(.026)	(.026)	(.018)
Own SAT score ÷ 100		.049	.018		.037	.009
		(.007)	(.006)		(.006)	(.006)
Log parental income			.187			.161
			(.024)			(.025)
Female			−.403			−.396
			(.015)			(.014)
Black			−.023			−.034
			(.035)			(.035)
Hispanic			.015			.006
			(.052)			(.053)
Asian			.173			.155
			(.036)			(.037)
Other/missing race			−.188			−.193
			(.119)			(.116)
High school top 10%			.061			.063
			(.018)			(.019)
High school rank missing			.001			−.009
			(.024)			(.022)
Athlete			.102			.094
			(.025)			(.024)
Average SAT score of schools applied to ÷ 100				.138	.116	.089
				(.017)	(.015)	(.013)
Sent two applications				.082	.075	.063
				(.015)	(.014)	(.011)
Sent three applications				.107	.096	.074
				(.026)	(.024)	(.022)
Sent four or more applications				.153	.143	.106
				(.031)	(.030)	(.025)

This table reports estimates of the effect of attending a more selective institution (measured by the average SAT score of enrolled students) on earnings. Each column shows coefficients from a regression of log earnings on school average SAT and controls. Columns (1)–(3) exclude selection controls, while columns (4)–(6) include them. Sample size: 14,238. Standard errors in parentheses.

- Without selection controls (cols. 1–3), the coefficient on school average SAT is positive and significant (0,109–0,076), suggesting that attending a more selective school is associated with higher earnings.
- With selection controls (cols. 4–6), the coefficient collapses toward zero (−0,021 to −0,008), and is no longer statistically significant.

- This shows that the apparent payoff from school selectivity is explained by student characteristics and preferences (selection), not by causal benefits of attending a higher-SAT peer group.

- **Intuition:** At first glance, selective schools seem to "boost" earnings because they admit stronger students. But once we control for what types of students *apply to and attend* these institutions, the effect of school selectivity disappears. In other words, *it's not the classmates, it's the selection.*

Private School Effects: Omitted Variable Bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	
High school rank missing		.556			−.030	
		(.102)			(.023)	
Athlete		−.318			.037	
		(.147)			(.016)	
Average SAT score of schools applied to ÷ 100			.777			.063
			(.058)			(.014)
Sent two applications			.252			.020
			(.077)			(.010)
Sent three applications			.375			.042
			(.106)			(.013)
Sent four or more applications			.330			.079
			(.093)			(.014)

This table reports regressions of (1)–(3) student ability (own SAT score) and (4)–(6) family background (log parental income) on private school attendance and controls. Large coefficients on "Private school" indicate that private attendance is strongly correlated with student ability and family income, suggesting omitted variable bias in simple regressions of earnings on private schooling.

- Columns (1)–(3) use **Own SAT score** as the dependent variable. Private school attendance is associated with more than a full standard deviation higher SAT score ($\approx$ 1,1), even after adding controls.

- Columns (4)–(6) use **Log parental income** as the dependent variable. Private school attendance is also associated with much higher family income (0,128–0,138), though the effect shrinks with more controls (0,028 in col. 6).
- These results reveal that private schooling is not randomly assigned: it is strongly correlated with both ability and family background.
- **Intuition:** If we simply regress earnings on private school attendance, we will overstate the causal effect. Why? Because part of the "private school premium" comes from omitted variables — smarter students (higher SATs) and richer families are more likely to attend private schools.
- This is a textbook case of *omitted variable bias (OVB)*: the regression coefficient on private schooling loads onto ability and income differences that are not controlled for.

Conclusions

- **Regression** provides the best-in-class approximation to the *Conditional Expectation Function (CEF)*.
- **Regressions** are usually the *first step* in empirical analysis.
- **Regression** is our *primary tool* for confronting the identification problem.
- If the regression you have is not delivering the desired relationship, then the underlying model itself is *unsatisfactory*.
- In such cases: **Instrumental Variables (IV)** is the natural next step.
- **Intuition:** Regression is like the default starting point — it gives us the clearest linear view of how Y relates to X. If this simple picture is misleading or biased, we need stronger tools. That's where IV comes in: it helps us deal with hidden confounders and extract the true causal effect.