The Conditional Expectation Function

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's Reading

Required

- **Aronow & Miller**, §2.2.3–2.2.4: CEF, LIE, best predictor (pp. 72–88)
- **Blackwell**, Ch. 1: What is regression really doing?

This is the most important lecture of the probability unit.
Everything that follows—regression, OLS, causal inference—builds on the CEF.

The Practical Question

You're an analyst at a campaign. Your boss asks:

"Among voters with a college degree, what's the average level of support for our candidate?"

What your boss wants is: $\mathbb{E}[\text{Support} \mid \text{Education} = \text{College}]$

She doesn't want:

- The full distribution of support among college voters
- Just the overall average support
- A complicated model

She wants a single number that summarizes support, conditional on education.

The Conditional Expectation Function

Definition

The **Conditional Expectation Function** (CEF) is:

$$G_Y(x) = \mathbb{E}[Y|X = x]$$

What is this?

- For each value of x, compute the expected value of Y among units with $X = x$
- The result is a *function* of x
- It summarizes the conditional distribution with a single number

Other names: Conditional mean, regression function

Blackwell calls this "the thing regression is trying to estimate."

Computing the CEF

The formulas you need

For continuous Y:

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\, dy$$

For discrete Y:

$$\mathbb{E}[Y|X=x] = \sum_{y} y \cdot \Pr(Y=y \mid X=x)$$

Key point: The CEF is a *function of x*—plug in different values of x and you get different numbers. It's not a single number.

We already learned conditional distributions in 04a. The CEF just takes their expected value.

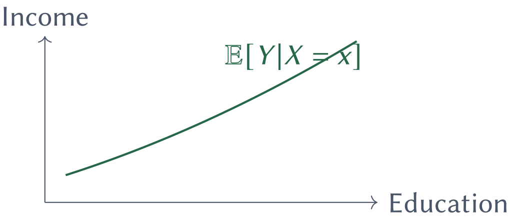
Example: Wages and Education

Setup: Y = annual income, X = years of education

The CEF $G_Y(x) = \mathbb{E}[\text{Income}|\text{Education} = x]$ answers:

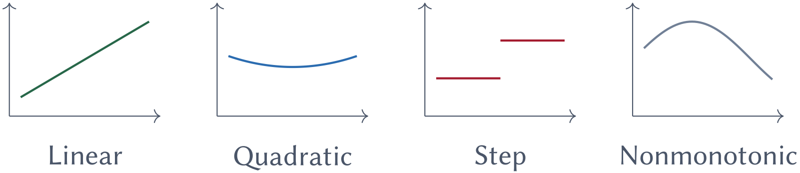
- What's the average income among people with 12 years of education?
- What's the average income among people with 16 years?
- What's the average income among people with 20 years?

The CEF traces out how average income changes with education.



The CEF Can Be Any Shape

Nothing requires the CEF to be linear.



Regression typically assumes linearity: $\mathbb{E}[Y|X = x] = \alpha + \beta x$
This is a *modeling assumption*, not a fact about the world.

When we get to OLS, we'll see it as approximating the true CEF with a line.

Why the CEF Matters: Best Prediction

Claim: The CEF is the *best predictor* of Y given X.

What do we mean by "best"?

Suppose you must predict Y using only X. You choose some function $g(X)$.

Define the **Mean Squared Error** of your prediction:

$$\mathsf{MSE}(g) = \mathbb{E}\left[(Y - g(X))^2\right]$$

Theorem: CEF is the MSE-Optimal Predictor

Among *all* functions $g(X)$, the CEF minimizes MSE:

$$\mathbb{E}[Y|X] = \arg\min_{g(X)} \mathbb{E}\left[(Y - g(X))^2\right]$$

Intuition: Why the CEF is Best

Think about what you're doing when you predict Y from X:

1. You observe $X = x$
2. You know the distribution of Y given $X = x$
3. You need to pick a single number as your guess

We already proved (Week 3): The best constant predictor of a random variable is its expected value.

Applying that here: Once we condition on $X = x$, the best prediction of Y is $\mathbb{E}[Y|X = x]$.

The CEF is just "pick the mean" applied separately for each $X = x$.

The CEF Residual

Define the CEF residual:

$$\varepsilon = Y - \mathbb{E}[Y|X]$$

This is what's "left over" after the CEF prediction.

Key Property of CEF Residuals

$$\mathbb{E}[\varepsilon|X] = 0$$

Why?

$$\begin{aligned}\mathbb{E}[\varepsilon|X] &= \mathbb{E}[Y - \mathbb{E}[Y|X] \mid X] \\ &= \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0\end{aligned}$$

The residual has mean zero *at every value of X*, not just overall.

Why This Matters

$\mathbb{E}[\varepsilon|X] = 0$ means the CEF "soaks up" all the predictable variation.

Implications:

- If $\mathbb{E}[\varepsilon|X] \neq 0$, we could improve our prediction
- The CEF captures everything X can tell us about Y
- What's left (ε) is genuinely unpredictable from X

This is why the CEF is the "best" predictor.

The Foundational Property

CEF Residual Orthogonality

$$\text{Cov}(\varepsilon, g(X)) = 0 \quad \text{for any function } g$$

In words: The CEF residual is uncorrelated with *any* function of X.

Why this matters:

- There is no remaining systematic relationship with X
- No transformation of X could improve the prediction
- This is the property regression tries to achieve

Regression residuals will satisfy a weaker version: $\text{Cov}(u, X) = 0$ (just linear).

The Law of Iterated Expectations (LIE)

Law of Iterated Expectations

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

In words: The overall mean of Y equals the average of the conditional means, weighted by the distribution of X.

Discrete case:

$$\mathbb{E}[Y] = \sum_x \mathbb{E}[Y|X = x] \cdot \Pr(X = x)$$

Also called the "law of total expectation" or "tower property."

LIE Example: Average Wages

Setup: Two groups—college grads and non-college.

Group	Share	Avg Wage
Non-College	0.60	$45,000
College	0.40	$75,000

What's the overall average wage?

Using LIE:

$$\begin{aligned}\mathbb{E}[\text{Wage}] &= \mathbb{E}[\text{Wage}|\text{No College}] \cdot \Pr(\text{No College}) \\ &\quad + \mathbb{E}[\text{Wage}|\text{College}] \cdot \Pr(\text{College}) \\ &= 45{,}000 \times 0.60 + 75{,}000 \times 0.40 \\ &= 27{,}000 + 30{,}000 = \$57{,}000\end{aligned}$$

LIE is Everywhere in Statistics

You'll use this constantly:

- Proving unbiasedness of estimators
- Deriving variance decompositions
- Understanding omitted variable bias
- Causal inference (potential outcomes, weighting)

Example preview (OVB derivation):
"What's the expected value of the short regression coefficient?"
"First condition on X, compute the expectation, then average over X."

Mastering LIE is essential for the rest of this course.

LIE with Extra Conditioning

The general version you'll need for proofs

Standard LIE (what we just saw):

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big]$$

With extra conditioning on Z:

$$\mathbb{E}[Y|Z] = \mathbb{E}\big[\mathbb{E}[Y|X,Z] \bigm| Z\big]$$

"Average first over X (holding Z fixed), then you have a function of Z only."

Conditioning on functions: If g is any function of X, then

$$\mathbb{E}[Y|g(X),X] = \mathbb{E}[Y|X]$$

Adding $g(X)$ provides no new information beyond X itself.

Example: If you know income (X), also knowing tax bracket ($g(X)$) doesn't help predict consumption.

The CEF Decomposition

We can always write:

$$Y = \mathbb{E}[Y|X] + \varepsilon$$

where $\mathbb{E}[\varepsilon|X] = 0$.

This is a **decomposition** of Y into:

- **Systematic part**: $\mathbb{E}[Y|X]$ — what X predicts
- **Idiosyncratic part**: ε — unpredictable from X

Regression does the same thing, but with a linear approximation:

$$Y = \alpha + \beta X + u$$

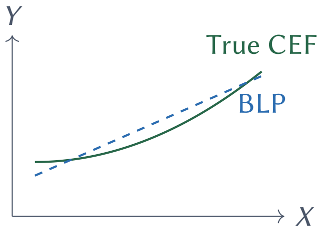
We'll make this connection precise in Week 8.

Blackwell's Take (Chapter 1)

From Blackwell:

> *"Linear regression is a method for finding the best linear approximation to the conditional expectation function."*

Key insight: Regression doesn't assume the CEF is linear.
It finds the *line* that gets closest to the true CEF, whatever shape it is.



The Variance Decomposition

Another use of the CEF: Decomposing variance.

Law of Total Variance

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$$

In words:

- Total variance = Within-group variance + Between-group variance
- $\mathbb{E}[\text{Var}(Y|X)]$ = Average variance of Y within each X group
- $\text{Var}(\mathbb{E}[Y|X])$ = Variance of the group means

This is the foundation of R-squared in regression.

Example: Wage Variance

Setup: Same as before, but now with within-group variance.

Group	Share	Mean Wage	SD of Wage
Non-College	0.60	\$45,000	\$15,000
College	0.40	\$75,000	\$25,000

Within-group variance: $\mathbb{E}[\text{Var}(Y|X)]$

$$= 0.60 \times (15{,}000)^2 + 0.40 \times (25{,}000)^2 = 385{,}000{,}000$$

Between-group variance: $\text{Var}(\mathbb{E}[Y|X])$ — variance of (45K, 75K) with weights (0.6, 0.4)

$$= 0.60 \times (45{,}000 - 57{,}000)^2 + 0.40 \times (75{,}000 - 57{,}000)^2 = 216{,}000{,}000$$

Total variance: $385M + 216M = 601{,}000{,}000$

Applications in Political Science

The CEF is everywhere in our research:

- $\mathbb{E}[\text{Vote Share}|\text{Incumbent}]$: Average vote share for incumbents vs. challengers
- $\mathbb{E}[\text{Turnout}|\text{Age}]$: How turnout varies with age
- $\mathbb{E}[\text{Approval}|\text{Economy}]$: Presidential approval as a function of economic conditions
- $\mathbb{E}[\text{Policy Position}|\text{Party}]$: Average policy positions by party

Regression estimates these relationships from data.

How Would You Estimate the CEF?

In practice, you have data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Simple approach (if X is discrete):

- For each value x, compute the sample mean of Y among observations with $X_i = x$
- This is the **sample analog** of the CEF

If X is continuous:

- Bin X and compute means within bins
- Or: fit a regression line (linear approximation to CEF)
- Or: use nonparametric methods (kernel regression, loess)

Regression = Linear approximation + estimation from sample data

Key Takeaways

1. **The CEF** $\mathbb{E}[Y|X = x]$ is the best predictor of Y given X
2. **CEF residuals** satisfy $\mathbb{E}[\varepsilon|X] = 0$ — no predictable part left
3. **Regression** approximates the CEF with a linear function

The big idea: The CEF is what regression is trying to estimate.

Next week: How do we learn about populations from samples?

For Monday

Topic: From Population to Sample

We've defined population quantities: $\mathbb{E}[Y]$, $\text{Var}(Y)$, $\mathbb{E}[Y|X]$.

But we only have sample data. How do we learn about populations from samples?

Reading: A&M §3.1–3.2, Blackwell Ch. 3