Joint Distributions, Conditioning, and the CEF

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today’s Reading

Required

- **Aronow & Miller**, §1.3: Joint, marginal, conditional distributions (pp. 31–44)
- **Aronow & Miller**, §2.2.1–2.2.4: Covariance, correlation, CEF, LIE (pp. 59–88)
- **Blackwell**, Ch. 1: What is regression really doing?

Today we build the full toolkit: joint distributions, conditional distributions, and the CEF—the function that regression is trying to estimate.

Galton’s Data: Height and Heredity

What tall parents and short parents tell us

Last week we met **Francis Galton** (1822–1911), the Victorian polymath who collected height data on hundreds of families in 1880s England to study heredity.

The obvious finding: Tall parents tend to have tall children. Short parents tend to have short children.

The surprising finding: Tall parents have tall children, but *not as tall as them.* Short parents have short children, but *not as short as them.*

The distribution of heights wasn’t spreading out over generations. It was **stable**.

Why? This puzzled Galton.

Galton's Insight: Regression to the Mean

The normal distribution explains the pattern

Recall the normal distribution: most probability mass is packed near the center.

- ∼68% of probability within 1 SD of the mean
- ∼95% within 2 SDs
- Only ∼2.5% in each tail beyond ±2 SDs

A father at the 99th percentile most likely has a son *closer to the middle*—not because a "force" pulls him back, but because there is so much more probability mass near the center.

Galton called this **"regression to the mean."**

The word "regression" predates OLS—Galton literally meant "regressing back toward" the average. It's a statistical phenomenon, not a causal one.

Galton’s Problem: Two Variables at Once

From “regression to the mean” to joint distributions

Galton realized he needed new mathematical machinery.

Knowing the *marginal* distribution of fathers’ heights and the *marginal* distribution of sons’ heights wasn’t enough. He needed to describe **how they move together**.

This required:

- A way to describe **two variables simultaneously** → *joint distributions*
- A way to ask **“given father’s height, what’s the son’s height?”** → *conditional distributions*
- A single number summarizing **how strongly they’re related** → *covariance and correlation*

Galton saw the need; Karl Pearson later formalized the math.

Today we build all of this machinery.

Why Do Political Scientists Care?

We always observe multiple variables together

Every interesting question involves relationships:

- Education and party identification
- Income and voter turnout
- War duration and casualties
- Campaign spending and vote share

Knowing each variable's distribution separately—even completely—misses the *relationship*.

We need a way to describe how two variables behave **together**.

A Concrete Question

Does education predict party identification?

Here's a question a political scientist might ask:

Are college-educated Americans more likely to identify as Republican than Americans without a college degree?

To answer this, we need to know how **education** and **party ID** are distributed *together*—not just separately.

Knowing that 50% of people are college-educated and 35% are Republican doesn't answer the question. We need to know: **of the college-educated, what fraction are Republican?**

Let's look at some data.

Example: Education and Party ID

Suppose we survey the population on education level and party identification.

	Party			
Education	Dem	Ind	Rep	Row Total
No College	0.20	0.15	0.15	0.50
College	0.18	0.12	0.20	0.50
Col Total	0.38	0.27	0.35	1.00

Each cell is the fraction of the population with that combination.

Example: 20% of people have no college degree and identify as Democrat.

All the cells sum to 1. This table captures everything about how these two variables relate.

What Just Happened?

A joint distribution is just a PMF on pairs

Education has **2 outcomes**. Party has **3 outcomes**.

But the pair (Education, Party) is itself a random variable—with $2 \times 3 =$ **6 outcomes** in its support:

$$\begin{gathered}(\text{No College, Dem}), \quad (\text{No College, Ind}), \quad (\text{No College, Rep}),\\ (\text{College, Dem}), \quad (\text{College, Ind}), \quad (\text{College, Rep})\end{gathered}$$

Each outcome gets a probability. They sum to 1. **That's just a PMF.**

The marginal distributions—Education alone, Party alone—are hiding inside this table. You get them by summing out the variable you don't care about.

Key point: You can recover the marginals from the joint. But you can't go the other way.

The Big Picture

That table has a name: it's a **joint distribution**.

What stays the same: Everything we've been doing still works.

- PMF? We have one—it's defined over pairs now.
- Expected values? Same idea—we'll use LOTUS on (x, y) pairs later today.
- Variance? Still applies.

What's new: Pairs let us ask questions single variables can't.

- **Marginals**: What does each variable look like on its own?
- **Conditionals**: If I know X, what does that tell me about Y?
- **Covariance and correlation**: How strongly are they related?

We'll build up marginals and conditionals first, then circle back to expected values and LOTUS.

Joint Distribution: Discrete Case

The education/party table is an example of a **joint PMF**:

Joint Probability Mass Function

For discrete random variables X and Y, the **joint PMF** is:

$$f(x, y) = \Pr(X = x \text{ and } Y = y)$$

Properties:

- $f(x, y) \geq 0$ for all x, y
- $\sum_x \sum_y f(x, y) = 1$

Every cell is a probability; they all sum to 1—just like our table.

Marginal Distributions

Question: What if we only care about X (ignoring Y)?

Marginal PMF

The **marginal distribution** of X is obtained by summing over Y:

$$f_X(x) = \sum_y f(x, y) = \Pr(X = x)$$

From our example:

- $f_X(\text{No College}) = 0.20 + 0.15 + 0.15 = 0.50$
- $f_X(\text{College}) = 0.18 + 0.12 + 0.20 = 0.50$

“Marginal” because these appear in the margins of the table.

The Marginal Is a Full Distribution

Not a single number—a complete PMF

We just computed:

- $f_X(\text{No College}) = 0.50$
- $f_X(\text{College}) = 0.50$

Those two numbers **together** are the marginal distribution of Education.

It's a complete PMF—with its own support {No College, College}, probabilities summing to 1, and you could compute $\mathbb{E}[X]$, $\text{Var}(X)$ from it. Everything we've been doing all semester.

Where Did the Marginal Come From?

Extracting a distribution from the joint table

What we did: For each value of X, we summed over all values of Y.

That operation *extracted* a full distribution from the joint table—one variable at a time.

The marginal is the distribution you'd see if you could only observe one variable.

You can always recover the marginals from the joint. But you can't recover the joint from the marginals—the relationship gets lost.

Visualizing Marginalization

	Dem	Ind	Rep	→ f_X
No Col	0.20	0.15	0.15	**0.50**
College	0.18	0.12	0.20	**0.50**

Sum across rows → marginal distribution of X

Conditional Distribution

Key question: Given that we *know* $X = x$, what's the distribution of Y?

Conditional PMF

The **conditional distribution** of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$$

Intuition:

- Zoom in on the row where $X = x$
- Renormalize so probabilities sum to 1

Same structure as conditional probability for events: $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$—now extended to random variables.

Example: Party ID Given Education

What's the distribution of Party among college graduates?

We need $f_{Y|X}(y|\text{College})$ for each party:

$$f_{Y|X}(\text{Dem}|\text{College}) = \frac{f(\text{College}, \text{Dem})}{f_X(\text{College})} = \frac{0.18}{0.50} = 0.36$$

$$f_{Y|X}(\text{Ind}|\text{College}) = \frac{0.12}{0.50} = 0.24$$

$$f_{Y|X}(\text{Rep}|\text{College}) = \frac{0.20}{0.50} = 0.40$$

Check: $0.36 + 0.24 + 0.40 = 1.00$ ✓

Among college grads: 36% Dem, 24% Ind, 40% Rep

Comparing Conditional Distributions

	Dem	Ind	Rep
$f_{Y\|X}(y\|\text{No College})$	0.40	0.30	0.30
$f_{Y\|X}(y\|\text{College})$	0.36	0.24	0.40

What do we learn?

- The two conditional distributions are *different*
- Knowing education level changes our beliefs about party ID
- The conditional distribution of *Y depends on X*

This dependence is what regression studies.

(These are stylized numbers for illustration, not real survey data.)

Conditional PDF: Interpretation

Two ways to think about joint distributions

Probability interpretation:

$$\Pr(a < Y < b \mid X = x) = \int_a^b f_{Y|X}(y|x)\, dy$$

Factorization of the joint PDF:

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) \cdot f_X(x)$$

Joint = Conditional $\times$ Marginal

Symmetrically: $f_{X,Y}(x, y) = f_{X|Y}(x|y) \cdot f_Y(y)$

Sampling from a Joint Distribution

The factorization tells you how to generate data

The factorization $f_{X,Y}(x, y) = f_{Y|X}(y|x) \cdot f_X(x)$ gives a recipe:

1. **Draw** $X \sim f_X$ (from the marginal)
2. **Draw** $Y \sim f_{Y|X}(\cdot \mid X)$ (from the conditional, given the X you drew)

Example: To simulate (Education, Party ID):

1. Draw Education: 50% chance College, 50% No College
2. If College $\longrightarrow$ draw Party from (0.36, 0.24, 0.40)
3. If No College $\longrightarrow$ draw Party from (0.40, 0.30, 0.30)

Every joint distribution implies a data-generating process.

Independence of Random Variables

When does knowing X tell us nothing about Y?

Definition: Independence

X and Y are **independent**, written $X \perp\!\!\!\perp Y$, if:

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \quad \text{for all } x, y$$

The joint is just the product of the marginals—no interaction, no dependence.

In our education/party example, X and Y are NOT independent—the conditional distributions differed.

But what does this definition actually *imply*?

What Independence Really Means

A short proof that connects everything we've built

We proved that joint distributions factorize:

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) \cdot f_X(x) \quad \text{(factorization)}$$

Independence says the joint is the product of marginals:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) \quad \text{(independence)}$$

Set the right-hand sides equal:

$$f_{Y|X}(y|x) \cdot f_X(x) = f_X(x) \cdot f_Y(y)$$

The $f_X(x)$ cancels:

$$\boxed{f_{Y|X}(y|x) = f_Y(y)}$$

The conditional distribution equals the marginal—conditioning on X changes nothing about Y.

In causal inference, independence assumptions like this are everywhere.

From Tables to Surfaces

Everything carries over from discrete to continuous

So far we've worked with **discrete** variables—tables with finitely many cells.

In practice, most variables we care about are **continuous**. The good news: every concept we just built carries over.

Concept	**Discrete**	**Continuous**
Joint distribution	PMF (table)	PDF (surface)
Marginal	Sum across row/column	Integrate out a variable
Conditional	Divide by row/column sum	Divide by marginal PDF
Independence	Product of PMFs	Product of PDFs
Joint CDF	$F(x, y) = \Pr(X \leq x, Y \leq y)$	Same formula

Same ideas, new notation. The CDF doesn't even change form. Let's see what this looks like.

Joint Distribution: Continuous Case

Joint Probability Density Function

For continuous X and Y, the **joint PDF** $f(x, y)$ satisfies:

$$\Pr(X \in A, Y \in B) = \iint_{A \times B} f(x, y)\, dx\, dy$$

Properties:

- $f(x, y) \geq 0$ (density is never negative)
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$ (integrate over everything = all probability = 1, just like summing all 6 cells)

Marginal: $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$

Conditional: $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$

What Is a “Surface”?

From curves to landscapes

A univariate PDF is a **curve**: for each value of x, the height $f(x)$ tells you the density.

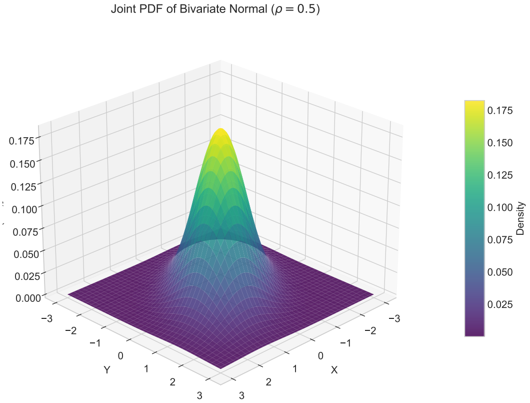
A joint PDF is a **surface**: for each pair (x, y), the height $f(x, y)$ tells you the density at that point. Think of it as a landscape—hills where outcomes are likely, valleys where they’re rare.

- In the univariate case, probability = **area** under the curve
- In the bivariate case, probability = **volume** under the surface

Let’s see what these surfaces actually look like.

Visualizing the Joint PDF: A 3D Surface

Height = density at each (x, y) point



Key insight: The joint PDF is a *surface* over the (x, y) plane. Higher = more likely.

Density Is Not Probability

The key difference between discrete and continuous

In the **discrete** case, the PMF gives you probabilities directly:

$$f(\text{College}, \text{Dem}) = 0.18 = \Pr(X = \text{College},\ \ Y = \text{Dem})$$

In the **continuous** case, the PDF gives you *density*, not probability. To get a probability, you have to **integrate**—just like the univariate case.

With one variable, probability = area under the curve.
With two variables, probability = **volume** under the surface.

Example: From Density to Probability

Setting up the double integral

Suppose X = voter ideology and Y = campaign spending, both continuous, with some joint PDF $f(x,y)$.

Question: What is the probability that a voter is moderate ($X \in [0.4, 0.6]$) *and* spending is moderate ($Y \in [1000, 5000]$)?

We integrate the density over that region:

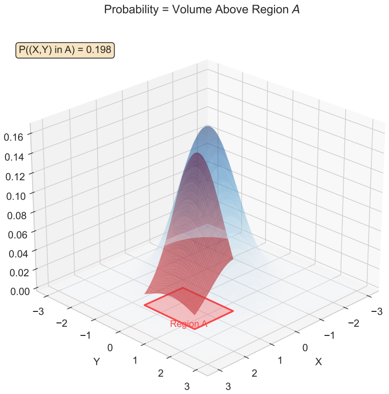
$$\Pr(X \in [0.4, 0.6],\ \ Y \in [1000, 5000]) = \int_{0.4}^{0.6} \int_{1000}^{5000} f(x,y)\, dy\, dx$$

Two variables, two integrals. The inner integral handles y; the outer handles x.

This is the volume under the surface over a rectangular region.

Probability = Volume Under the Surface

Connecting geometry to integrals



$$\Pr((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy = \textbf{volume above region } A$$

Example: Computing a Joint Probability

Setting up the problem

Let X = ideology (scaled 0 to 1) and Y = turnout propensity (0 to 1), with joint PDF:

$$f(x, y) = 2x \quad \text{for } 0 \leq x \leq 1,\ 0 \leq y \leq 1$$

Notice: density depends on x but not y—more ideologically extreme people have higher density regardless of their turnout propensity.

Question: What is $\Pr(0.4 \leq X \leq 0.6,\ 0.4 \leq Y \leq 0.6)$?

We need to integrate the density over this region—a double integral:

$$\Pr = \int_{0.4}^{0.6} \int_{0.4}^{0.6} 2x \, dy \, dx$$

Example: Solving the Double Integral

Inner integral first, then outer

$$\int_{0.4}^{0.6}\int_{0.4}^{0.6} 2x\,dy\,dx$$

Step 1: Inner integral—integrate over y, treating x as a constant:

$$\int_{0.4}^{0.6} 2x\,dy = 2x \cdot (0.6 - 0.4) = 0.4x$$

Step 2: Outer integral—now integrate over x:

$$\int_{0.4}^{0.6} 0.4x\,dx = 0.4 \cdot \left.\frac{x^2}{2}\right|_{0.4}^{0.6} = 0.4 \cdot (0.18 - 0.08) = \boxed{0.04}$$

A 4% chance of being moderate *and* moderate-turnout. Work inside out.

What If You Only Care About One Variable?

From joint probabilities to marginal distributions

We just computed $\Pr(X \in [0.4, 0.6],\ Y \in [0.4, 0.6])$—a probability about *both* variables over specific ranges.

But sometimes you only care about one. What's the distribution of ideology *regardless* of turnout?

In the **discrete** case, we did this: to find Pr(College), we summed across the entire College row.

In the **continuous** case, same idea—integrate out the variable you don't care about over its *entire* range:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

For joint probability, we integrated both variables over *specific* ranges.
For marginals, we integrate one variable over *everything*—it disappears.

Example: Finding the Marginal Distributions

Using our same joint PDF $f(x, y) = 2x$

Marginal of X (ideology): integrate out Y

$$f_X(x) = \int_0^1 2x\, dy = 2x \cdot (1 - 0) = 2x$$

More density at higher x—ideology is not uniformly distributed.

Marginal of Y (turnout): integrate out X

$$f_Y(y) = \int_0^1 2x\, dx = x^2\Big|_0^1 = 1$$

A uniform distribution! Turnout propensity is equally likely everywhere.

This makes sense: $f(x, y) = 2x$ never depended on y, so the marginal of Y is flat.

Expected Values from the Marginals

Now that we have f_X and f_Y, compute their means

Expected ideology:

$$\mathbb{E}[X] = \int_0^1 x \cdot f_X(x)\, dx = \int_0^1 x \cdot 2x\, dx = \int_0^1 2x^2\, dx = \frac{2x^3}{3}\bigg|_0^1 = \frac{2}{3}$$

The average ideology is 2/3, skewed right—consistent with more density at higher x.

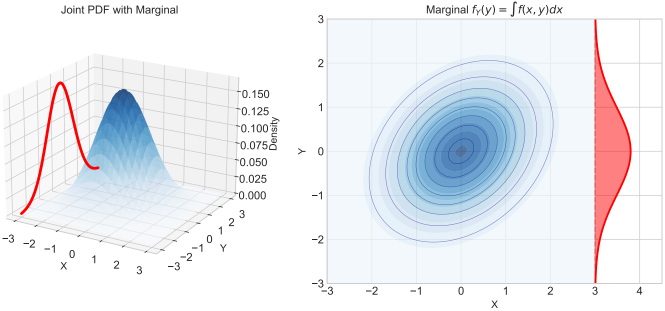
Expected turnout:

$$\mathbb{E}[Y] = \int_0^1 y \cdot f_Y(y)\, dy = \int_0^1 y \cdot 1\, dy = \frac{y^2}{2}\bigg|_0^1 = \frac{1}{2}$$

The average turnout is 1/2—exactly what you'd expect from a uniform distribution.

Marginal Distribution = “Projection”

Collapse the surface onto one axis



$f_Y(y) = \int f(x, y)\, dx$ — integrate out X, “project” onto the Y-axis.

Same idea as summing across rows in the discrete case.

What If You Fix One Variable?

From marginals to conditionals

With marginals, we *eliminated* a variable—integrated it out entirely.

Now a different question: what does Y look like *if we know* $X = x$?

In the **discrete** case, we did this: to find the distribution of Party among college graduates, we divided joint probabilities by the marginal $\Pr(\text{College})$.

In the **continuous** case, same logic:

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Marginal = eliminate a variable (integrate it out).
Conditional = fix a variable (divide by its marginal).

Example: Finding the Conditional Distribution

Using our same joint PDF $f(x, y) = 2x$

We found $f_X(x) = 2x$ and $f_Y(y) = 1$. Now compute the conditionals:

Conditional of Y given X:

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{2x}{2x} = 1$$

Knowing ideology tells you nothing about turnout—it's still uniform.

Conditional of X given Y:

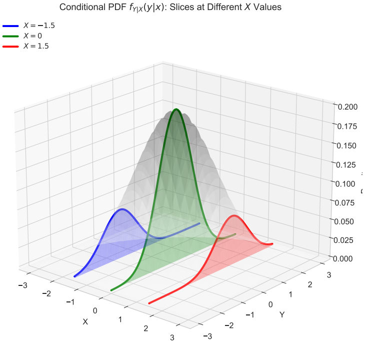
$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{2x}{1} = 2x$$

Knowing turnout tells you nothing about ideology—it's still $2x$.

Both conditionals equal the marginals. We proved earlier that this means $X \perp\!\!\!\perp Y$—and indeed $f(x, y) = 2x \cdot 1 = f_X(x) \cdot f_Y(y)$.

Conditional Distribution = “Slice”

Fix $X = x$ and look at the cross-section



Conditional PDF $f_{Y|X}(y|x)$: Slices at Different X Values

$f_{Y|X}(y|x)$ = the slice at $X = x$, renormalized to integrate to 1.

If the slice changes shape as x changes, X and Y are dependent. This is what regression studies!

From Joint CDF to Joint PDF

The CDF–PDF relationship generalizes too

In the univariate case, we went from CDF to PDF by differentiating:

$$f(x) = \frac{d}{dx} F(x)$$

With two variables, we take **partial derivatives** in both:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \, \partial y}$$

The CDF accumulates probability in a region; the PDF is the rate of change in both directions.

You'll use this on problem sets—given a joint CDF, differentiate to get the PDF.

Multivariate Expectation and 2D LOTUS

Computing expectations of functions of two random variables

Expected value of a function of (X, Y):

$$\mathbb{E}[g(X, Y)] = \iint g(x, y) f(x, y) \, dx \, dy$$

2D LOTUS (Law of the Unconscious Statistician):

- No need to find the distribution of $g(X, Y)$
- Integrate $g(x, y)$ directly against the joint PDF

Key applications:

- $\mathbb{E}[XY]$ — needed for covariance (coming next!)
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ — linearity still holds

Example: Computing $\mathbb{E}[XY]$

Using 2D LOTUS with $f(x, y) = 2x$

Set $g(x, y) = xy$ and apply 2D LOTUS:

$$\mathbb{E}[XY] = \int_0^1 \int_0^1 xy \cdot 2x \, dy \, dx = \int_0^1 \int_0^1 2x^2 y \, dy \, dx$$

Inner integral (over y):

$$\int_0^1 2x^2 y \, dy = 2x^2 \cdot \frac{y^2}{2} \Big|_0^1 = x^2$$

Outer integral (over x):

$$\int_0^1 x^2 \, dx = \frac{x^3}{3} \Big|_0^1 = \boxed{\frac{1}{3}}$$

We already found $\mathbb{E}[X] = \frac{2}{3}$ and $\mathbb{E}[Y] = \frac{1}{2}$, so $\mathbb{E}[X]\,\mathbb{E}[Y] = \frac{1}{3} = \mathbb{E}[XY]$.
Independence again: $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. This quantity is the key ingredient for covariance.

Why Does Dependence Matter?

Independence assumptions are everywhere in statistics

We constantly assume independence:

- Poll responses assumed independent
- RCT: treatment assignment $\perp\!\!\!\perp$ background characteristics
- Regression errors assumed independent across observations

Lack of independence is a blessing or a curse:

- **Blessing**: Two variables not independent $\Rightarrow$ potentially interesting relationship
- **Curse**: In observational studies, treatment is usually not independent of background $\Rightarrow$ confounding

Question: How do we *measure* dependence?

Covariance: Measuring Dependence

How often do high values of X occur with high values of Y?

Definition: Covariance

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Equivalent formula (often easier to compute):

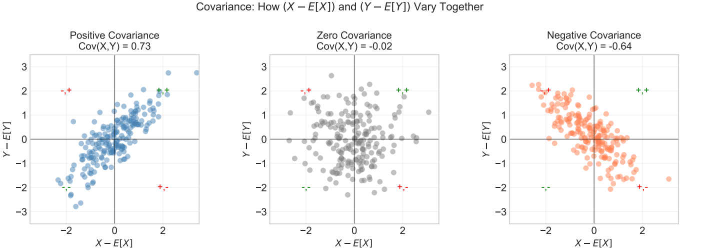
$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

Key property: If $X \perp\!\!\!\perp Y$, then $\text{Cov}(X, Y) = 0$

But the converse is NOT true! Zero covariance does not imply independence.

Covariance: Quadrant Intuition

The sign depends on where points cluster



- **Positive Cov**: Points cluster in $(+,+)$ and $(-,-)$ quadrants
- **Negative Cov**: Points cluster in $(+,-)$ and $(-,+)$ quadrants
- **Zero Cov**: Balanced across all quadrants

Properties of Covariance

1. $\text{Cov}(X, X) = \text{Var}(X)$ (variance is covariance with itself!)
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ (symmetric)
3. $\text{Cov}(X, c) = 0$ for any constant c
4. $\text{Cov}(aX, Y) = a \cdot \text{Cov}(X, Y)$
5. $\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$

Important Consequence

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$$

Unlike expected values, variances are NOT linear—covariance is the "correction term."

Correlation: Scale-Free Dependence

Covariance depends on units; correlation doesn't

Problem: Covariance depends on the scale of X and Y.

Definition: Correlation

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)}$$

Equivalent form:

$$\rho(X, Y) = \text{Cov}\left(\frac{X - \mathbb{E}[X]}{\text{SD}(X)}, \frac{Y - \mathbb{E}[Y]}{\text{SD}(Y)}\right)$$

Correlation = covariance of standardized variables

Correlation: Properties

Key properties:

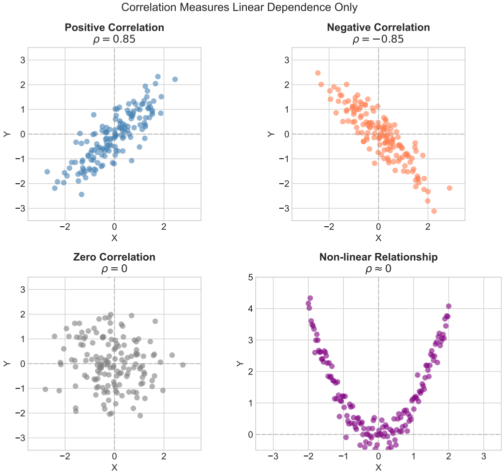
- $-1 \leq \rho(X, Y) \leq 1$
- $|\rho(X, Y)| = 1$ if and only if $Y = a + bX$ for some constants
  - $\rho = 1$: perfect positive linear relationship
  - $\rho = -1$: perfect negative linear relationship
- $\rho = 0$: no *linear* relationship

Critical caveat: Correlation measures **linear** dependence only.

Two variables can have $\rho = 0$ but still be strongly dependent! (e.g., $Y = X^2$ where X is symmetric around 0)

Correlation: Examples

What different ρ values look like



Putting It All Together

One distribution that uses everything we've built

We now have a full toolkit: joint distributions, marginals, conditionals, independence, covariance, and correlation.

We've been waiting to introduce this distribution because it *requires* correlation as a parameter. Now that we have ρ, we're ready.

The bivariate normal.

- Its joint distribution is a bell-shaped **surface** (the 3D hills we saw earlier)
- Its marginals are normal
- Its conditionals are normal
- Its dependence is fully captured by one number: ρ

This is where joint distributions meet regression.

Simulating the Bivariate Normal in R

What does the joint distribution look like as data?

On the left: X and Y are independent normals ($\rho = 0$).
On the right: X and Y are correlated ($\rho = 0.7$).

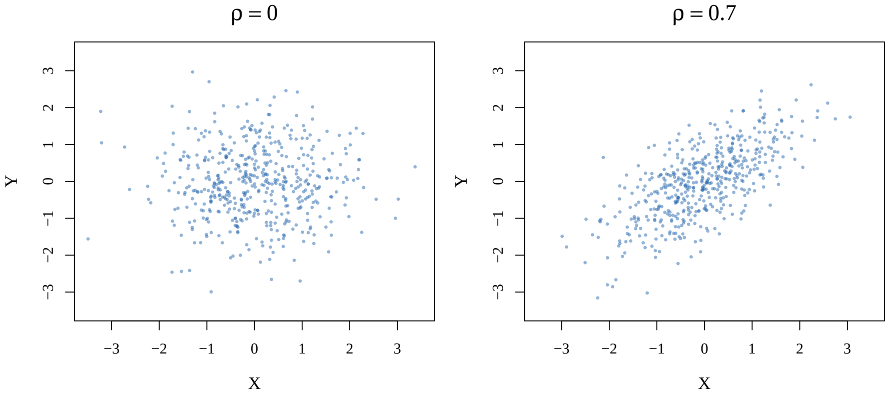
```
library(MASS)
# Independent (rho = 0)
Sigma0 <- matrix(c(1, 0, 0, 1), 2, 2)
xy0 <- mvrnorm(500, mu = c(0,0), Sigma = Sigma0)

# Correlated (rho = 0.7)
Sigma7 <- matrix(c(1, 0.7, 0.7, 1), 2, 2)
xy7 <- mvrnorm(500, mu = c(0,0), Sigma = Sigma7)
```

The joint distribution is that cloud of dots—where they cluster is where density is high.

The Bivariate Normal: Independence vs. Correlation

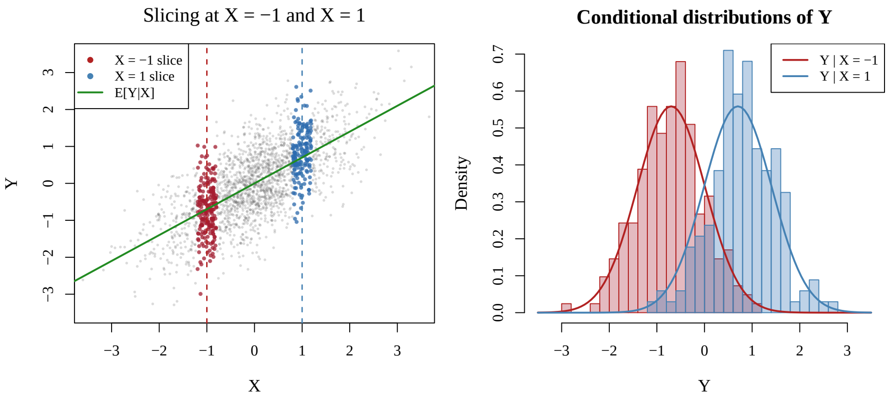
500 draws from each distribution



$\rho = 0$: circular cloud (no pattern). $\rho = 0.7$: tilted ellipse (knowing X tells you about Y).

Seeing the Conditional Distribution

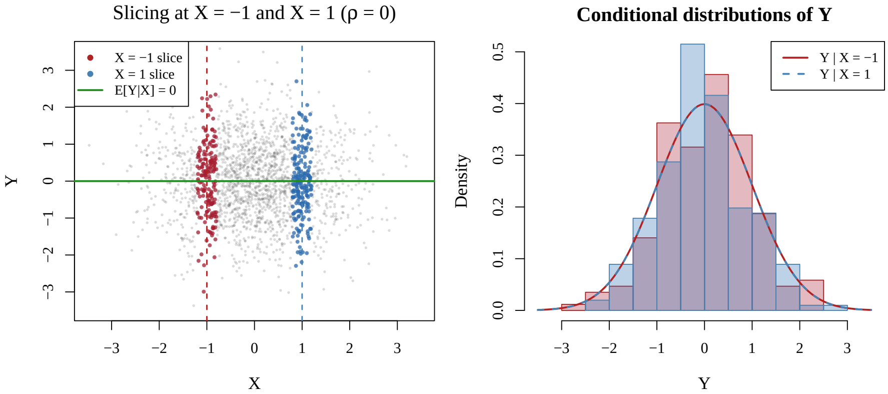
Slice the $\rho = 0.7$ cloud at $X = -1$ and $X = 1$



Each slice is normal, but the **center shifts**: $\mathbb{E}[Y|X = x] = \rho \cdot x$ (the green line). The conditional variance $\sigma_Y^2(1 - \rho^2)$ is the same at every slice.

Now Compare: Independence ($\rho = 0$)

Same slices, but now the centers don't move



Same center, same spread. $f_{Y|X}(y|x) = f_Y(y)$—exactly what we proved independence means.

Formalizing What We Just Saw

The conditional distribution of the bivariate normal

Those histograms were normal at every slice. Here is the formula:

Conditional Distribution

$$Y|X = x \sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X),\ \sigma_Y^2(1 - \rho^2)\right)$$

You saw each of these in the simulation:

- **Mean is linear in** x: $\mathbb{E}[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ (the green line)
- **Variance is constant**: $\sigma_Y^2(1 - \rho^2)$ (both histograms had the same spread)
- If $\rho = 0$, conditional mean $= \mu_Y$ (knowing X tells us nothing)

This is regression. The conditional mean traces a line through (x, y) space.

But Don't Forget

A common logical error

We just showed:

$X \perp\!\!\!\perp Y \implies \rho = 0$ ✓ True

Many people think the reverse also holds:

$\rho = 0 \implies X \perp\!\!\!\perp Y$ **FALSE**

This is the logical fallacy of *affirming the consequent*: "A implies B" does **not** mean "B implies A."

Why? Correlation only measures *linear* dependence. Two variables can have $\rho = 0$ but be strongly dependent through a nonlinear relationship (e.g., $Y = X^2$ where X is symmetric around 0).

Independence rules out *all* relationships—linear and nonlinear. Zero correlation only rules out linear ones.

Where We Stand

We've built the full toolkit for two variables:

1. **Joint → Marginal**: Sum/integrate out what you don't care about
2. **Joint → Conditional**: Divide by marginal to "zoom in" on a slice
3. **Covariance**: Measures how X and Y move together
4. **Correlation**: Scale-free measure of *linear* dependence
5. **Bivariate normal**: Conditional distribution is normal, mean is linear in X

Now the question: How do we *summarize* a conditional distribution?

From Conditional Distributions to Conditional Means

A short history of the most important function in statistics

We just learned: the conditional distribution $f_{Y|X}(y|x)$ captures everything about the relationship between X and Y.

But a full distribution is a lot of information.

In 1805, **Adrien-Marie Legendre** published the method of least squares. His question was practical: given noisy astronomical observations, how do you find the best-fitting curve?

His answer—minimize squared prediction errors—implicitly targets the **conditional mean**. The CEF was hiding inside regression for 200 years before anyone named it.

Legendre didn't think in terms of conditional distributions. The formal connection came much later, through the work of Kolmogorov (1933) and the modern probability framework.

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 earlier today. The CEF just takes their expected value.

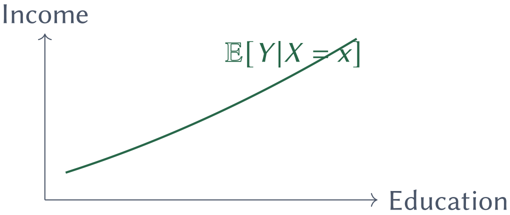
Example: Wages and Education

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

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

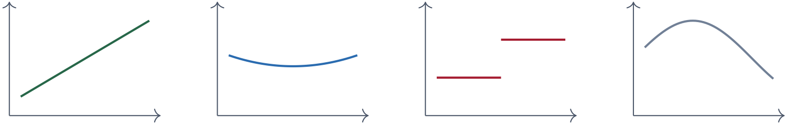
- What's the average income among people with 12 years of education?
- What's the average income among people with 16 years?
- What's the average income among people with 20 years?

The CEF traces out how average income changes with education.



The CEF Can Be Any Shape

Nothing requires the CEF to be linear.



Linear Quadratic Step Nonmonotonic

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.

The Foundational Property: Orthogonality

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 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 when we cover OLS.

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

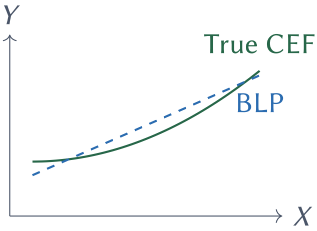
$$= 0.60 \times (45{,}000 - 57{,}000)^2 + 0.40 \times (75{,}000 - 57{,}000)^2 = 216{,}000{,}000$$

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

Blackwell’s Take: CEF vs. Linear Regression

From Blackwell (Ch. 1):

Linear regression finds the best *linear* approximation to the CEF, whatever shape the CEF has.



Key insight: Regression doesn’t *assume* the CEF is linear. It finds the line that gets closest to the true CEF.

When the CEF *is* linear (as in the bivariate normal), the BLP and the CEF coincide.

Applications in Political Science

The CEF is everywhere in political science 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. **Joint → Marginal → Conditional**: The hierarchy of distributions
2. **The CEF** $\mathbb{E}[Y|X = x]$ is the best predictor of Y given X
3. **CEF residuals** satisfy $\mathbb{E}[\varepsilon|X] = 0$—no predictable part left
4. **LIE**: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$—average the conditional averages
5. **Regression** approximates the CEF with a linear function

The big idea: Everything in regression traces back to the CEF.

Next Time: 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