

# Why Regression?

EC 425/525, Set 3

Edward Rubin

16 April 2019

# Prologue

# Schedule

## Last time

- The Experimental Ideal
- Fundamentals of R (wrap up Lab 1).

## Today

What's so great about linear regression and OLS?

**Read** MHE 3.1

## Upcoming

**Assignment** First step of project proposal due April 15<sup>th</sup>.

# Follow up

`return()`

1. `function()` automatically returns the last evaluated value—regardless of `return()`.
2. Hadley Wickham<sup>†</sup> suggests reserving `return` for "early" returns.

<sup>†</sup> An R god

# Regression

# Regression

## Why?

In our previous discussion, we began moving from simple differences to a regression framework.

**Q** Why do we<sup>†</sup> care so much about linear regression and OLS?

**A** As we discussed, regression allows us to control for covariates that can assist with (1) causal identification and (2) inference.

There's a deeper reason that we care about *linear* regression and ordinary least squares (OLS): ***the conditional expectation function (CEF)***.

<sup>†</sup> we = empirically inclined applied economists

# Regression

## Why?

Even ignoring causality, we can show important relationships between

1. **the CEF** (the conditional expectation function),
2. the **population regression function**,
3. and the **sampling distribution of regression estimates**.

# Regression

## The CEF

**Definition** The **conditional expectation function** for a dependent variable  $\mathbf{Y}_i$ , given a  $K \times 1$  vector of covariates  $\mathbf{X}_i$ , tells us **the expected value** (population average) of  $\mathbf{Y}_i$  with  $\mathbf{X}_i$  held constant.

Written as  $E[\mathbf{Y}_i | \mathbf{X}_i]$ , the CEF is a function of  $\mathbf{X}_i$ .<sup>†</sup>

## Examples

- $E[\text{Income}_i | \text{Education}_i]$
- $E[\text{Wage}_i | \text{Gender}_i]$
- $E[\text{Birth weight}_i | \text{Air quality}_i]$

<sup>†</sup> We'll generally assume  $\mathbf{X}_i$  is a random variable, which implies that  $E[\mathbf{Y}_i | \mathbf{X}_i]$  is also a random variable.

# Regression

## The CEF

Formally, for continuous  $\mathbf{Y}_i$  with conditional density  $f_y(t|\mathbf{X}_i = x)$ ,

$$E[\mathbf{Y}_i \mid \mathbf{X}_i = x] = \int t f_y(t|\mathbf{X}_i = x) dt$$

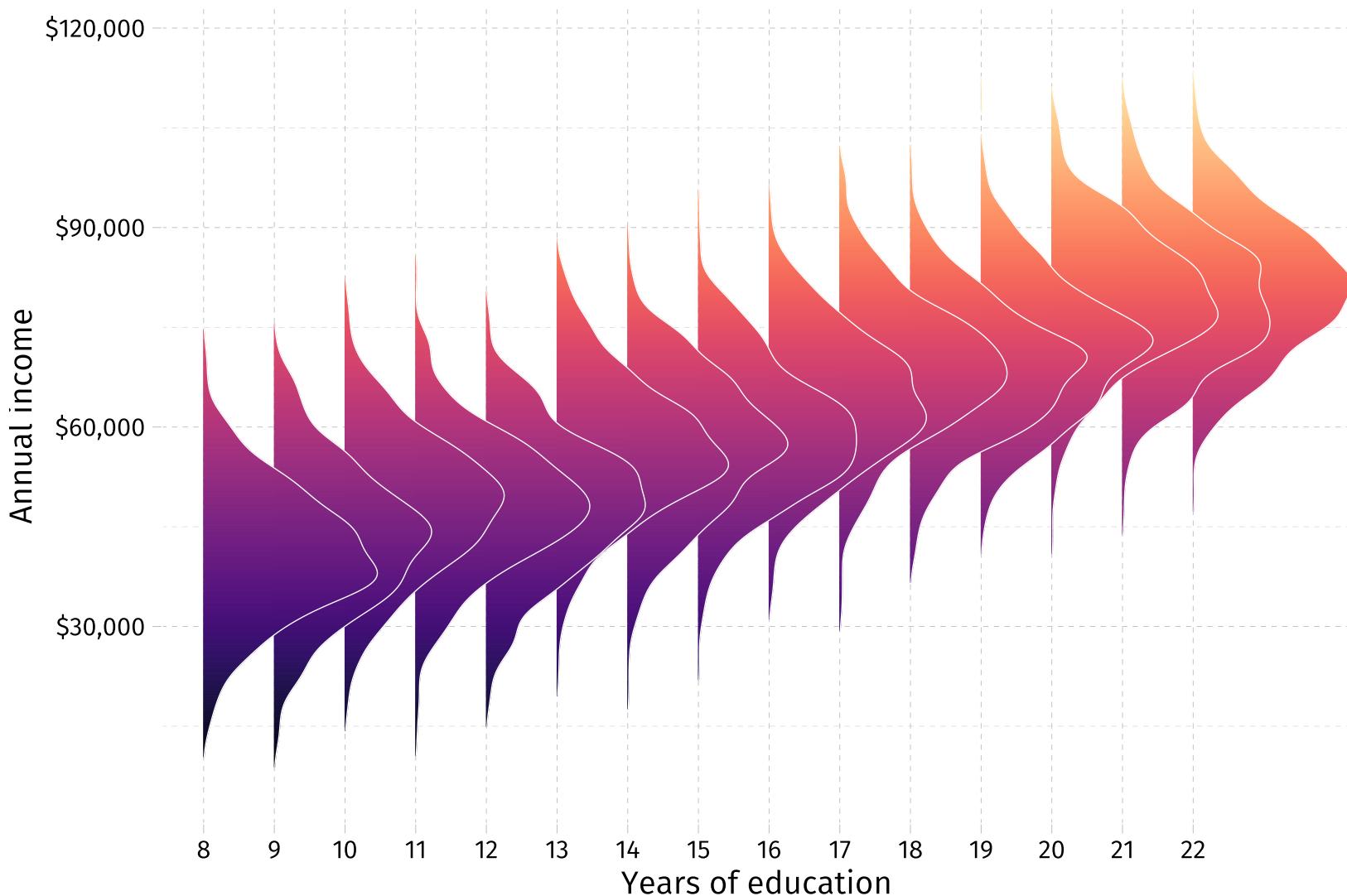
and for discrete  $\mathbf{Y}_i$  with conditional p.m.f.  $\Pr(\mathbf{Y}_i = t|\mathbf{X}_i = x)$ ,

$$E[\mathbf{Y}_i \mid \mathbf{X}_i = x] = \sum_t t \Pr(\mathbf{Y}_i = t|\mathbf{X}_i = x)$$

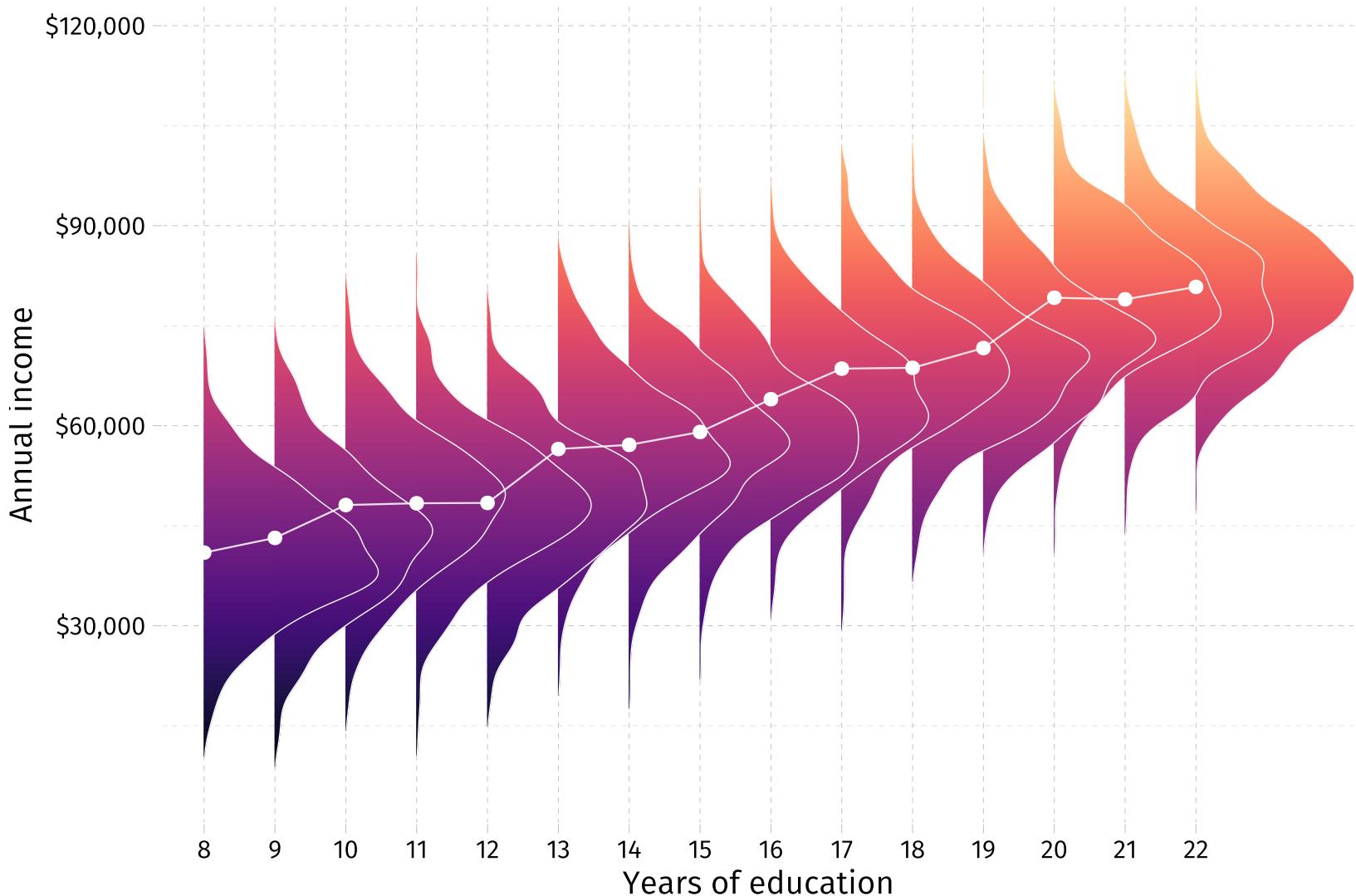
**Notice** We are focusing on the **population**. We want to build our intuition about the parameters that we will eventually estimate.

Graphically...

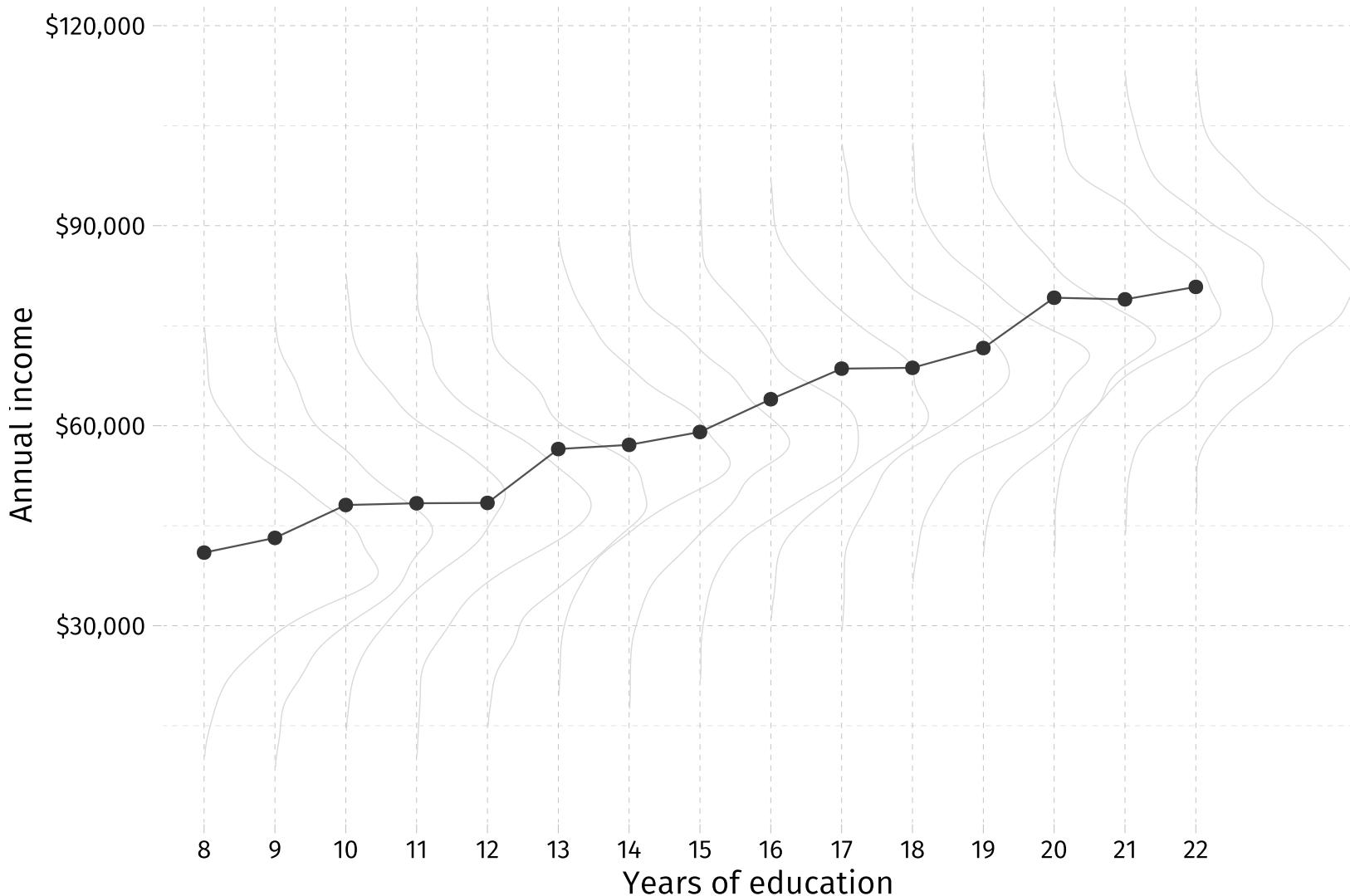
The conditional distributions of  $Y_i$  for  $X_i = x$  in 8, ..., 22.



The CEF,  $E[Y_i | X_i]$ , connects these conditional distributions' means.



Focusing in on the CEF,  $E[Y_i | X_i]$ ...



**Q** How does the CEF relate to/inform regression?

# Regression

## The CEF

As we derive the properties and relationships associated with the CEF, regression, and a host of other estimators, we will frequently rely upon ***the Law of Iterated Expectations*** (LIE).

$$E[Y_i] = E\left(E[Y_i | X_i]\right)$$

which says that the **unconditional expectation** is equal to the **unconditional average** of the **conditional expectation function**.

# Regression

## A proof of the LIE

First, we need notation...

Let  $f_{x,y}(u, t)$  denote the joint density for continuous RVs  $(\mathbf{X}_i, \mathbf{Y}_i)$ .

Let  $f_{y|x}(t | \mathbf{X}_i = u)$  denote the conditional distribution of  $\mathbf{Y}_i$  given  $\mathbf{X}_i = u$ .

And let  $g_y(t)$  and  $g_x(u)$  denote the marginal densities of  $\mathbf{Y}_i$  and  $\mathbf{X}_i$ .

# Regression

## A proof of the LIE

$$\begin{aligned} E\left(E[Y_i \mid X_i]\right) &= \int E[Y_i \mid X_i = u] g_x(u) du \\ &= \int \left[ \int t f_{y|x}(t \mid X_i = u) dt \right] g_x(u) du \\ &= \int \int t f_{y|x}(t \mid X_i = u) g_x(u) du dt \\ &= \int t \left[ \int f_{y|x}(t \mid X_i = u) g_x(u) du \right] dt \\ &= \int t \left[ \int f_{x,y}(u, t) du \right] dt \\ &= \int t g_y(t) dt \\ &= E[Y_i] \quad \text{🥳} \end{aligned}$$

Great. What's the point?

# Regression

## The *LIE* and the *CEF*

**Theorem** The CEF decomposition property (3.1.1)

The LIE allows us to **decompose random variables** into two pieces

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

1. **the conditional expectation function**
2. **a residual** with special powers<sup>†</sup>
  - i.  $\varepsilon_i$  is mean independent of  $X_i$ , i.e.,  $E[\varepsilon_i | X_i] = 0$ .
  - ii.  $\varepsilon_i$  is uncorrelated with any function of  $X_i$ .

**Important** It might not seem like much, but these results are **huge** for building intuition, theory, and application. Put a  here!

<sup>†</sup> Angrist and Pischke go with *special properties*.

# Regression

## The *LIE* and the *CEF*

**Proof** The CEF decomposition property (properties i. and ii. of  $\varepsilon_i$ )

**Mean independence**,  $E[\varepsilon_i | \mathbf{X}_i] = 0$

**Zero correlation** btn.  $\varepsilon_i$  and  $h(\mathbf{X}_i)$

$$E[\varepsilon_i | \mathbf{X}_i]$$

$$E[h(\mathbf{X}_i)\varepsilon_i]$$

$$= E\left(Y_i - E[Y_i | \mathbf{X}_i] \middle| \mathbf{X}_i\right)$$

$$= E\left(E[h(\mathbf{X}_i)\varepsilon_i | \mathbf{X}_i]\right)$$

$$= E[Y_i | \mathbf{X}_i] - E\left(E[Y_i | \mathbf{X}_i] \middle| \mathbf{X}_i\right)$$

$$= E\left(h(\mathbf{X}_i) E[\varepsilon_i | \mathbf{X}_i]\right)$$

$$= E[Y_i | \mathbf{X}_i] - E[Y_i | \mathbf{X}_i]$$

$$= E[h(\mathbf{X}_i) \times 0]$$

$$= 0$$

$$= 0$$

# Regression

## The *LIE* and the *CEF*

### The **CEF decomposition property**

says that we can decompose any random variable (e.g.,  $\mathbf{Y}_i$ ) into

1. a part that is *explained by  $\mathbf{X}_i$*  (i.e., the CEF  $E[\mathbf{Y}_i | \mathbf{X}_i]$ ),
2. a part that is *orthogonal to<sup>†</sup>* any function of  $\mathbf{X}_i$  (i.e.,  $\varepsilon_i$ ).

### Why the **CEF**?

The **CEF** also presents an intuitive summary of the relationship between  $\mathbf{Y}_i$  and  $\mathbf{X}_i$ , since we often use means to characterize random variables.

But (of course) there are more reasons to use the CEF...

<sup>†</sup> "orthogonal to" = "uncorrelated with"

# Regression

## The *LIE* and the *CEF*

**Theorem** The CEF prediction property (3.1.2)

Let  $m(\mathbf{X}_i)$  be *any* function of  $\mathbf{X}_i$ . The CEF solves

$$E[\mathbf{Y}_i \mid \mathbf{X}_i] = \arg \min_{m(\mathbf{X}_i)} E[(\mathbf{Y}_i - m(\mathbf{X}_i))^2]$$

In other words, the **CEF** is the minimum mean-squared error (MMSE) predictor of  $\mathbf{Y}_i$  given  $\mathbf{X}_i$ .

### **Notice**

1. We haven't restricted  $m$  to any class of functions—it can be nonlinear.
2. We're talking about *prediction* (specifically predicting  $\mathbf{Y}_i$ ).

**Proof** The CEF prediction property

$$\left( Y_i - m(X_i) \right)^2 \tag{1}$$

$$= \left( \{ Y_i - E[Y_i | X_i] \} + \{ E[Y_i | X_i] - m(X_i) \} \right)^2$$

$$= \left( Y_i - E[Y_i | X_i] \right)^2 \tag{a}$$

$$+ 2 \left( E[Y_i | X_i] - m(X_i) \right) \times \left( Y_i - E[Y_i | X_i] \right) \tag{b}$$

$$+ \left( E[Y_i | X_i] - m(X_i) \right)^2 \tag{c}$$

**Recall:** We want to choose the  $m(X_i)$  that minimizes (1) in expectation.

(a) is irrelevant, i.e., it does not depend upon  $m(X_i)$ .

(b) equals zero in expectation:  $E[h(X_i) \times \varepsilon_i] = 0$ .

(c) is minimized by  $m(X_i) = E[Y_i | X_i]$ , i.e., when  $m(X_i)$  is the CEF.

# Regression

## The *LIE* and the *CEF*

.  
∴ the **CEF** is the function that minimizes the mean-squared error (MSE)

$$E[Y_i | X_i] = \arg \min_{m(X_i)} E\left[(Y_i - m(X_i))^2\right]$$

# Regression

## The *LIE* and the *CEF*

One final property of the *CEF* (very similar to the decomposition property)

**Theorem** The ANOVA theorem (3.1.3)

$$\text{Var}(\mathbf{Y}_i) = \text{Var}(\mathbf{E}[\mathbf{Y}_i \mid \mathbf{X}_i]) + \mathbf{E}[\text{Var}(\mathbf{Y}_i \mid \mathbf{X}_i)]$$

which says that we can decompose the variance in  $\mathbf{Y}_i$  into

1. the variance in the *CEF*
2. the variance of the residual

**Example** Decomposing wage variation into (1) variation explained by workers' characteristics and (2) unexplained (residual) variation

The proof centers on the fact that the decomposition property of the CEF.

We now understand the CEF a bit better.

But how does the CEF actually relate to regression?

# Regression

## The *CEF* and regression

We've discussed how the *CEF* summarizes empirical relationships.

Previously we discussed how regression provides simple empirical insights.

Let's link these two concepts.

# Regression

## The CEF and regression

### Population least-squares regression

We will focus on  $\beta$ , the vector (a  $K \times 1$  matrix) of population, least-squares regression coefficients, *i.e.*,

$$\beta = \arg \min_b E \left[ (\mathbf{Y}_i - \mathbf{X}'_i b)^2 \right]$$

where  $b$  and  $\mathbf{X}_i$  are also  $K \times 1$ , and  $\mathbf{Y}_i$  is a scalar.

Taking the first-order condition gives

$$E \left[ \mathbf{X}_i (\mathbf{Y}_i - \mathbf{X}'_i b) \right] = 0$$

# Regression

## The CEF and regression

From the first-order condition

$$E[\mathbf{X}_i (\mathbf{Y}_i - \mathbf{X}'_i b)] = 0$$

we can solve for  $b$ . We've defined the optimum as  $\beta$ . Thus,

$$\beta = E[\mathbf{X}_i \mathbf{X}'_i]^{-1} E[\mathbf{X}_i \mathbf{Y}_i]$$

**Note** The first-order conditions tell us that our least-squares population regression residuals ( $e_i = \mathbf{Y}_i - \mathbf{X}'_i \beta$ ) are uncorrelated with  $\mathbf{X}_i$ .

# Regression

## Anatomy

Our "new" result:  $\beta = E[\mathbf{X}_i \mathbf{X}'_i]^{-1} E[\mathbf{X}_i \mathbf{Y}_i]$

In **simple linear regression** (an intercept and one regressor  $x_i$ ),

$$\beta_1 = \frac{\text{Cov}(\mathbf{Y}_i, x_i)}{\text{Var}(x_i)} \quad \beta_0 = E[\mathbf{Y}_i] - \beta_1 E[x_i]$$

For **multivariate regression**, the coefficient on the  $k^{\text{th}}$  regressor  $x_{ki}$  is

$$\beta_k = \frac{\text{Cov}(\mathbf{Y}_i, \tilde{x}_{ki})}{\text{Var}(\tilde{x}_{ki})}$$

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

# Regression

## Anatomy

This alternative formulation of least-squares coefficients is quite powerful.

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

**Why?** This expression illustrates how each coefficient in a least-squares regression represents the bivariate slope coefficient **after controlling for the other covariates**.

# Regression

## Anatomy

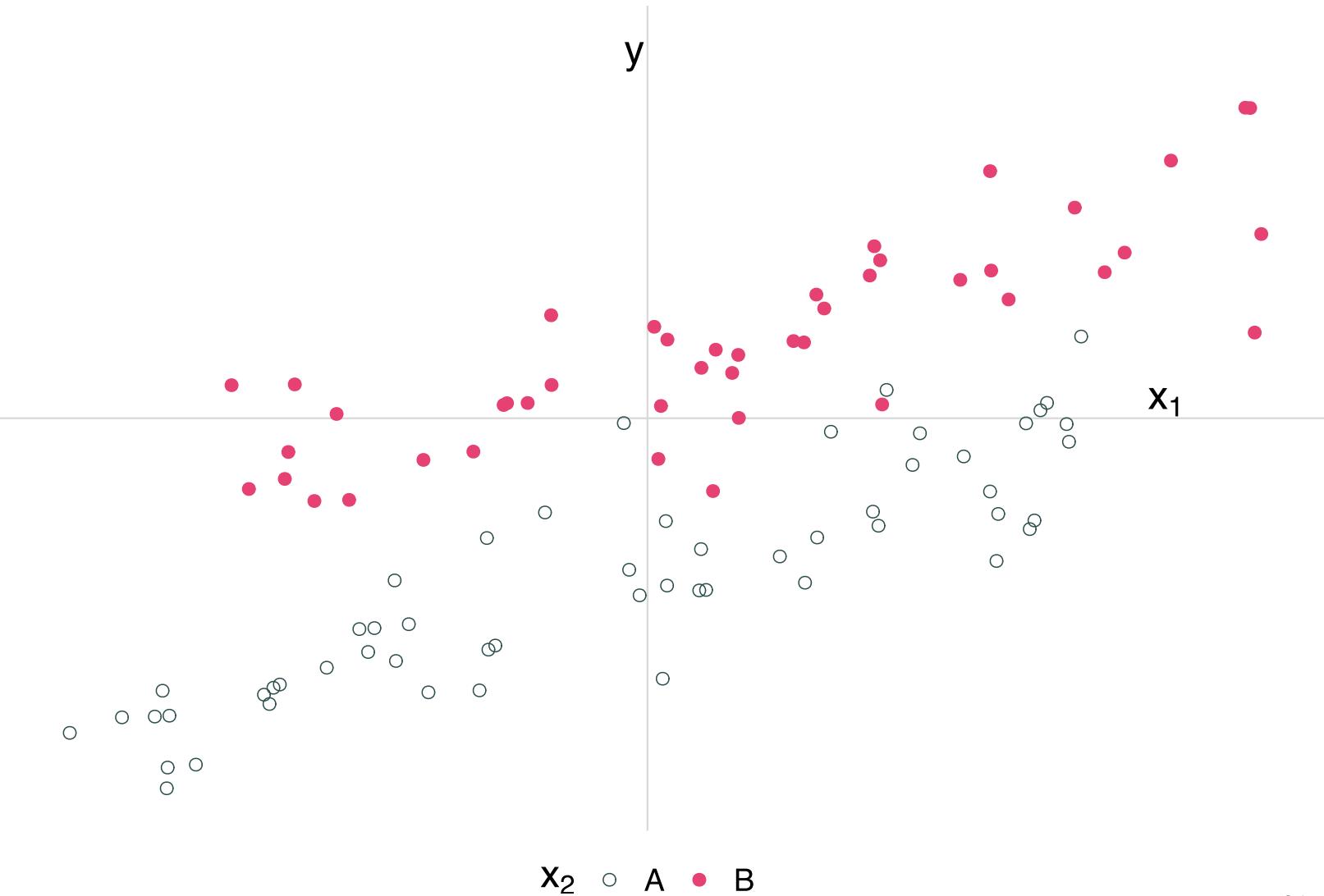
In fact, we can re-write our coefficients to further emphasize this point

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

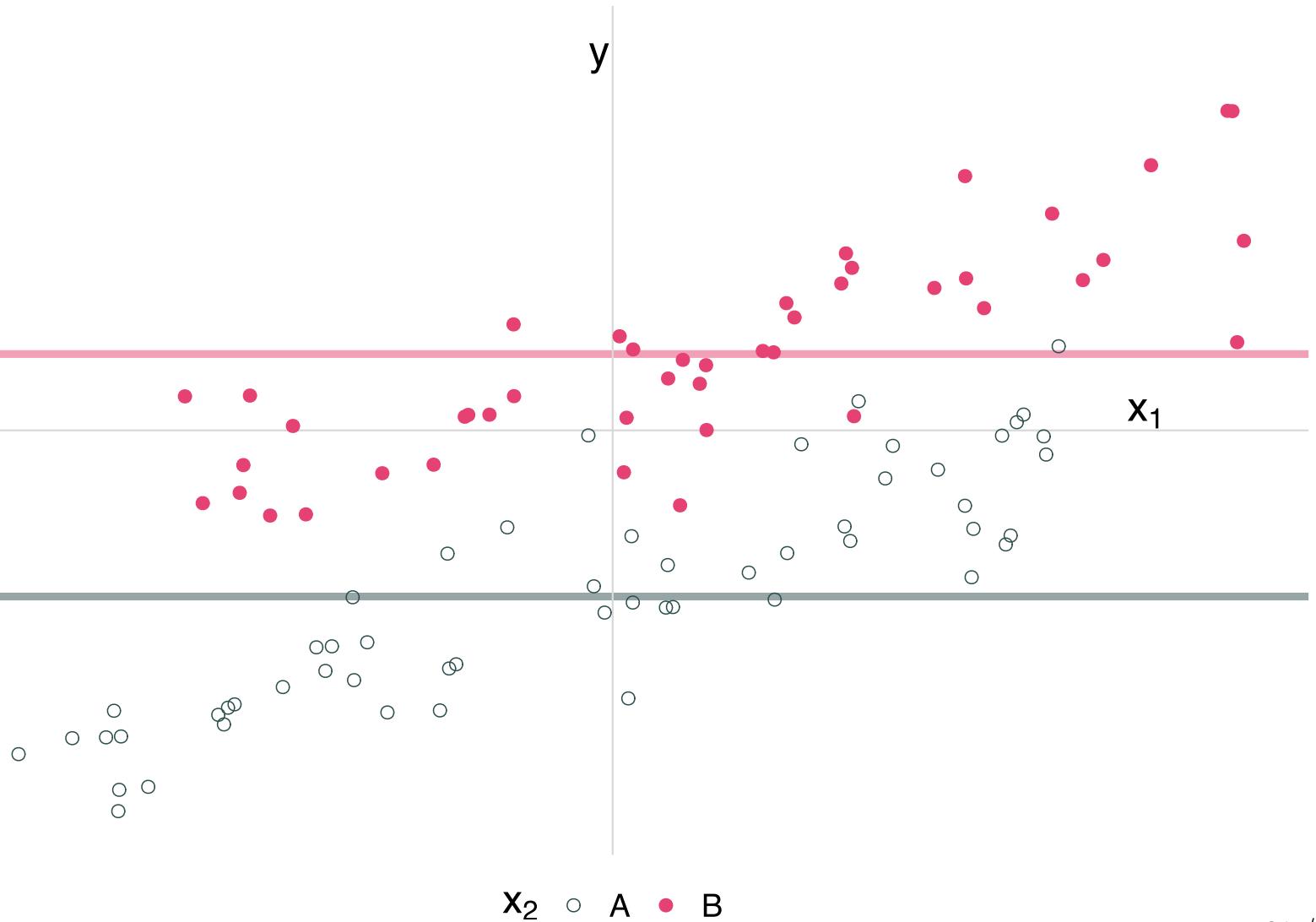
$\tilde{Y}_i$  denotes the residual from regressing  $Y_i$  on all regressors except  $x_{ki}$ .

Graphical example

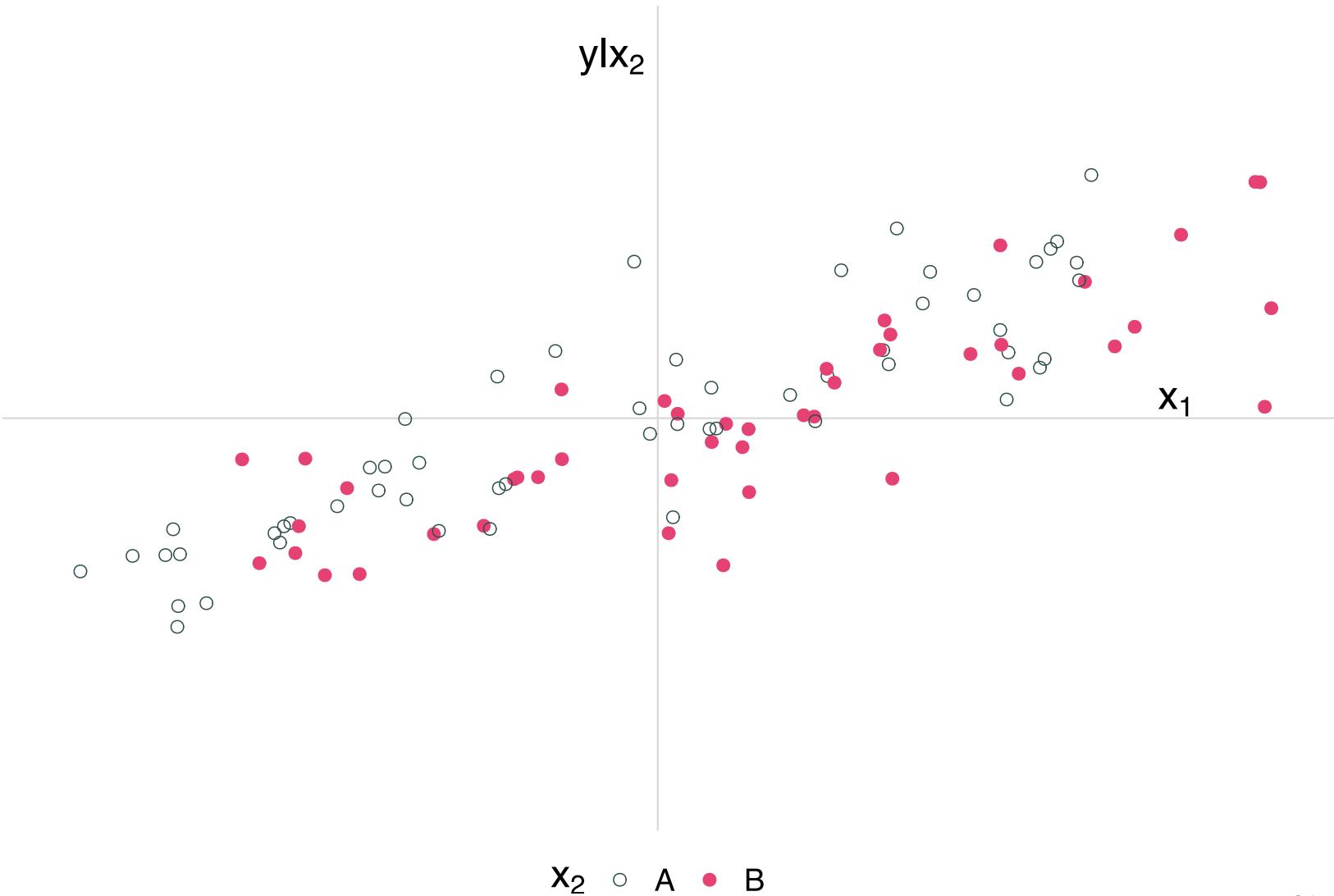
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$



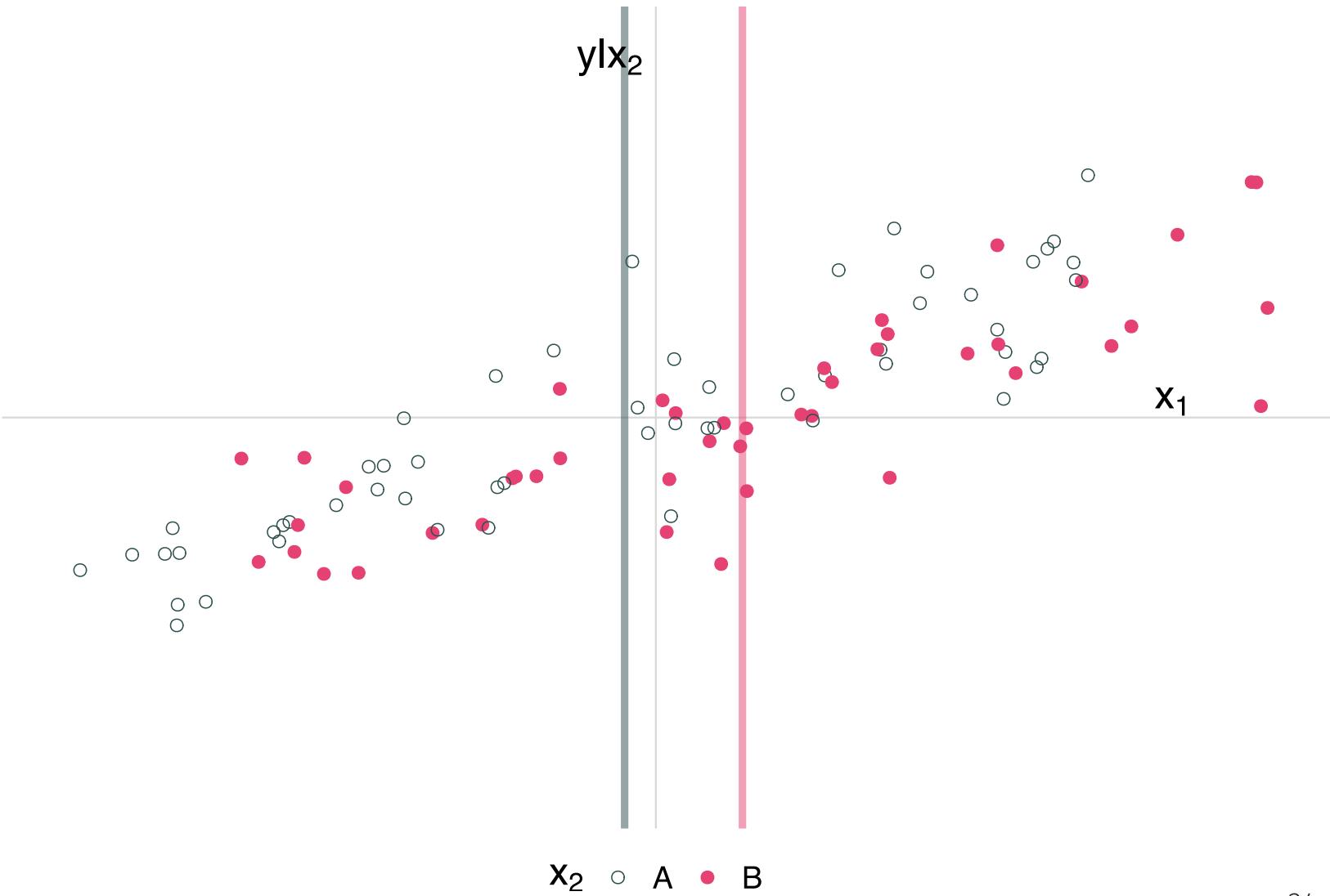
$\beta_1$  gives the relationship between  $y$  and  $x_1$  after controlling for  $x_2$



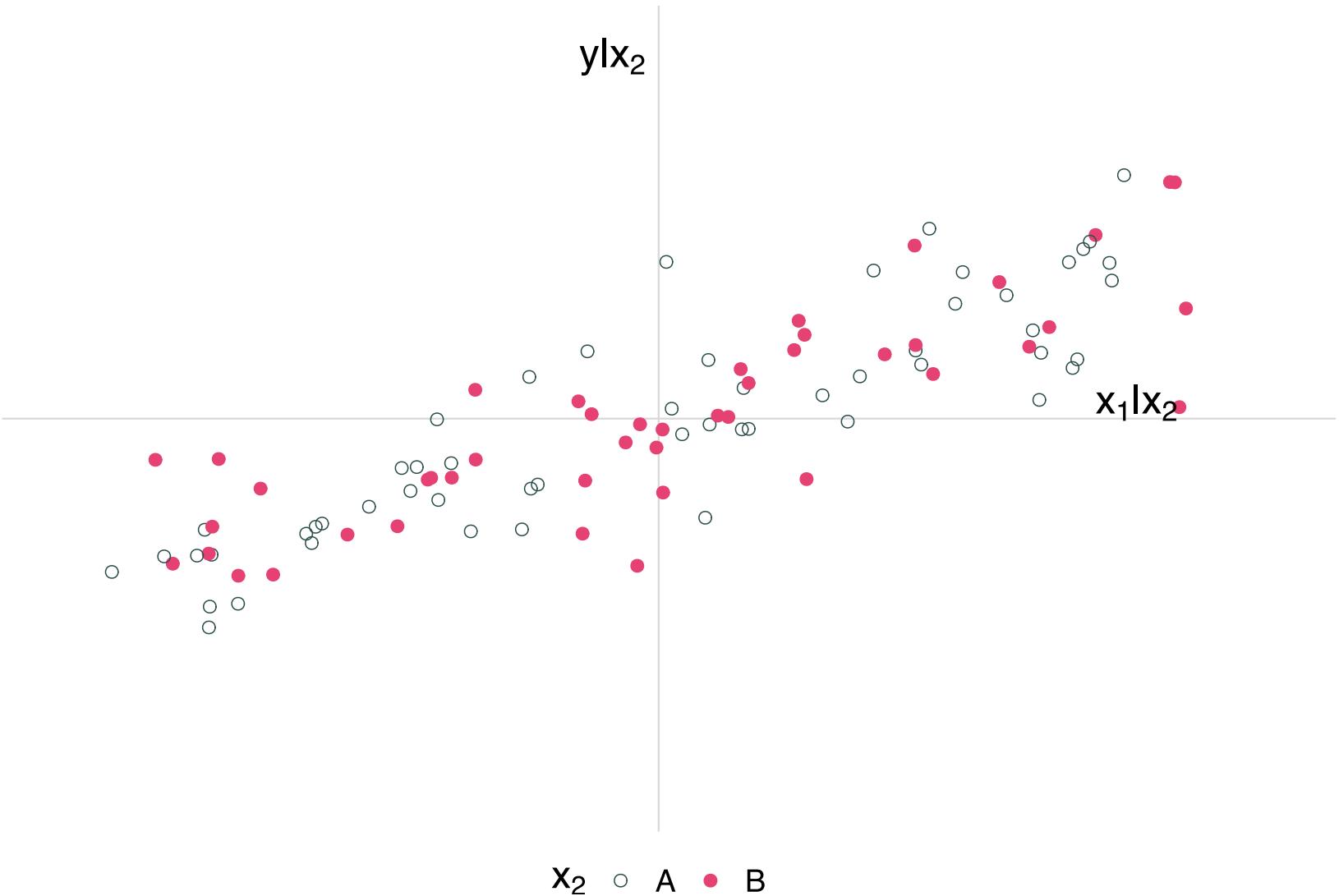
$\beta_1$  gives the relationship between  $y$  and  $x_1$  after controlling for  $x_2$



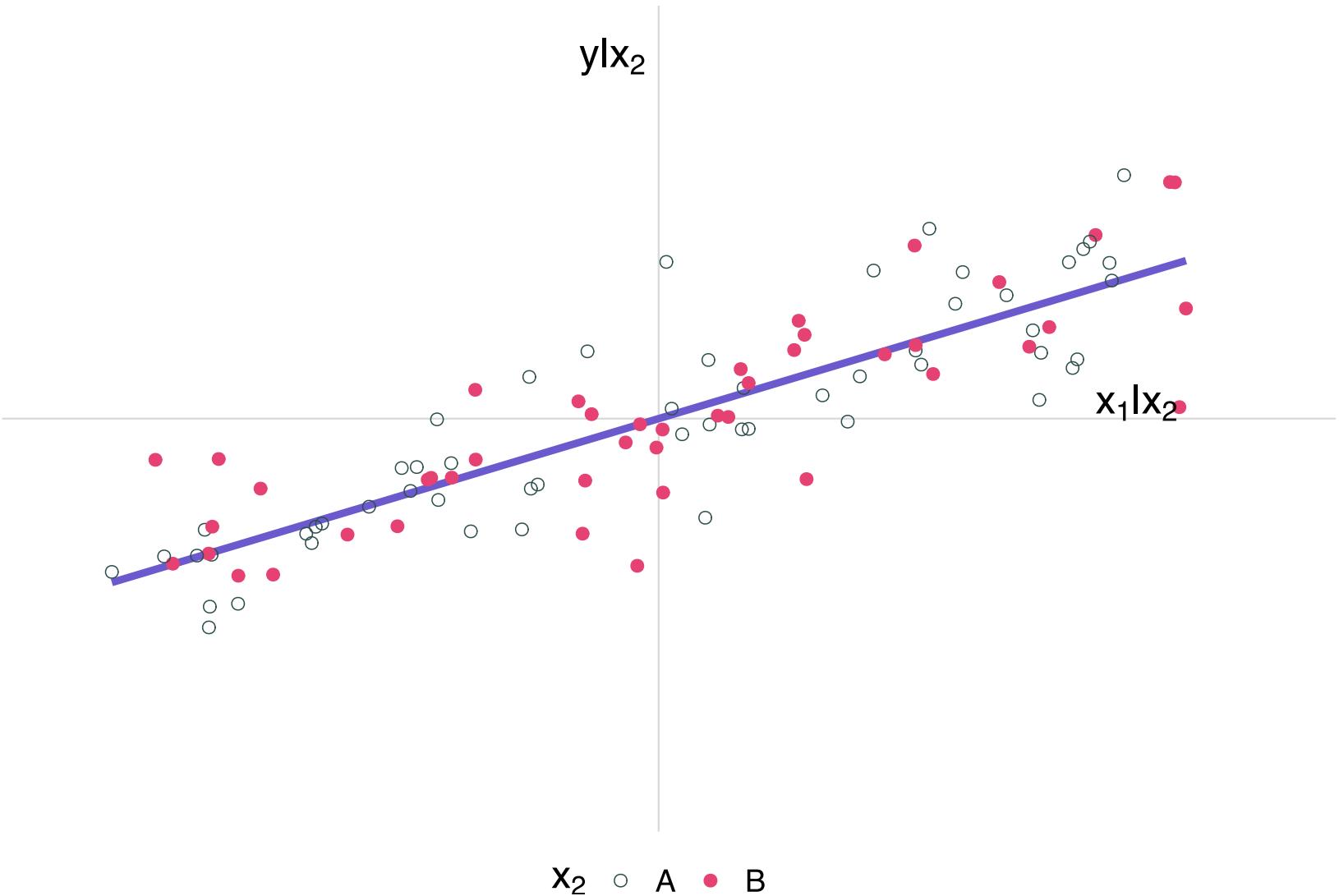
$\beta_1$  gives the relationship between  $y$  and  $x_1$  after controlling for  $x_2$



$\beta_1$  gives the relationship between  $y$  and  $x_1$  after controlling for  $x_2$



$\beta_1$  gives the relationship between  $y$  and  $x_1$  after controlling for  $x_2$



Now that we've refreshed/deepened our regression knowledge, let's connect regression and the CEF.

# Regression

## Regression and the CEF

Angrist and Pischke make the case that

... you should be interested in regression parameters if you are interested in the CEF. (*MHE*, p.36)

**Q** What is the reasoning/connection?

**A** We'll cover three reasons.

1. If the CEF is linear, then the population regression line is the CEF.
2. The function  $\mathbf{X}'_i \beta$  is the min. MSE linear predictor of  $\mathbf{Y}_i$  given  $\mathbf{X}_i$ .
3. The function  $\mathbf{X}'_i \beta$  gives the min. MSE linear approximation to the CEF.

# Regression

## Regression and the CEF

**Theorem** The linear CEF theorem (3.1.4)

If the CEF is linear, then the population regression is the CEF.

**Proof** Let the CEF equal some linear function, i.e.,  $E[\mathbf{Y}_i \mid \mathbf{X}_i] = \mathbf{X}'_i \beta^*$ .

From the CEF decomposition property, we know  $E[\mathbf{X}_i \varepsilon_i] = 0$ .

$$\implies E[\mathbf{X}_i (\mathbf{Y}_i - E[\mathbf{Y}_i \mid \mathbf{X}_i])] = 0$$

$$\implies E[\mathbf{X}_i (\mathbf{Y}_i - \mathbf{X}'_i \beta^*)] = 0$$

$$\implies E[\mathbf{X}_i \mathbf{Y}_i] - E[\mathbf{X}_i \mathbf{X}'_i \beta^*] = 0$$

$$\implies \beta^* = E[\mathbf{X}_i \mathbf{X}'_i]^{-1} E[\mathbf{X}_i \mathbf{Y}_i] = \beta, \text{ our population regression coefficients.}$$

# Regression

## Regression and the CEF

**Theorem** The linear CEF theorem (3.1.4)

If the CEF is linear, then the population regression is the CEF.

Linearity can be a strong assumption. When might we expect linearity?

1. Situations in which  $(Y_i, X_i)$  follows a multivariate normal distribution.

**Concern** Might be limited—especially when  $Y_i$  or  $X_i$  are not continuous.

1. Saturated regression models

**Example** A model with two binary indicators and their interaction.

# Regression

## Regression and the CEF

**Theorem** The best linear predictor theorem (3.1.5)

The function  $\mathbf{X}'_i \beta$  the best linear predictor of  $\mathbf{Y}_i$  given  $\mathbf{X}_i$  (minimizes MSE).

**Proof** We defined  $\beta$  as the vector that minimizes MSE, *i.e.*,

$$\beta = \arg \min_b E[(\mathbf{Y}_i - \mathbf{X}'_i b)^2]$$

so  $\mathbf{X}'_i \beta$  is literally defined as the minimum MSE linear predictor of  $\mathbf{Y}_i$ .

- The population-regression function ( $\mathbf{X}'_i \beta$ ) is the best (min. MSE) *linear* predictor of  $\mathbf{Y}_i$  given  $\mathbf{X}_i$ .
- The CEF ( $E[\mathbf{Y}_i | \mathbf{X}_i]$ ) is the best predictor (min. MSE) of  $\mathbf{Y}_i$  given  $\mathbf{X}_i$  across *all classes* of functions.

# Regression

## Regression and the CEF

**Q** If  $\mathbf{X}'_i \beta$  is **the best linear predictor** of  $\mathbf{Y}_i$  given  $\mathbf{X}_i$ , then why is there so much interest machine learning for prediction (opposed to regression)?

**A** A few reasons

1. Relax *linearity*
2. Model selection
  - choosing  $\mathbf{X}_i$  is not always obvious
  - overfitting is bad
3. It's fancy/shiny and new
4. Some ML methods boil down to regression
5. Others?

**Counter Q** Why are we (still) using regression?

# Regression

## Regression and the CEF

**Theorem** The regression CEF theorem (3.1.6)

The population regression function  $\mathbf{X}'_i \beta$  provides the minimum MSE linear approximation to the CEF  $E[\mathbf{Y}_i | \mathbf{X}_i]$ , i.e.,

$$\beta = \arg \min_b E \left\{ \left( E[\mathbf{Y}_i | \mathbf{X}_i] - \mathbf{X}'_i b \right)^2 \right\}$$

**Put simply** Regression gives us the *best* linear approximation to the CEF.

**Proof** First, recall that, in expectation,  $\beta$  is the  $b$  that minimizes  $(Y_i - X'_i b)^2$

$$(Y_i - X'_i b)^2 \tag{1}$$

$$= \left( \{Y_i - E[Y_i | X_i]\} + \{E[Y_i | X_i] - X'_i b\} \right)^2$$

$$= \left( Y_i - E[Y_i | X_i] \right)^2 \tag{a}$$

$$+ \left( E[Y_i | X_i] - X'_i b \right)^2 \tag{b}$$

$$+ 2 \left( Y_i - E[Y_i | X_i] \right) \left( E[Y_i | X_i] - X'_i b \right) \tag{c}$$

We want to minimize (b), and we know  $\beta$  minimizes (1).

(a) is irrelevant, i.e., it does not depend upon  $b$ .

(c) can be written as  $2\varepsilon_i h(X_i)$ , which equals zero in expectation.

.: (In expectation) If  $b = \beta$  minimizes (1), then  $b = \beta$  minimizes (b).

# Regression

## Regression and the CEF

Let's review our new(-ish) regression results

1. When the CEF is linear, the regression function *is* the CEF.  
**Too small** Very specific circumstances—or big assumptions.
2. Regression gives us the best *linear* predictor of  $\mathbf{Y}_i$  (given  $\mathbf{X}_i$ )  
**Off point** We're often interested in  $\beta$ —not  $\hat{\mathbf{Y}}_i$ .
3. Regression provides the best *linear* approximation of the CEF.  
**Just right?** (Depends on your goals)

# Regression

## Regression and the CEF

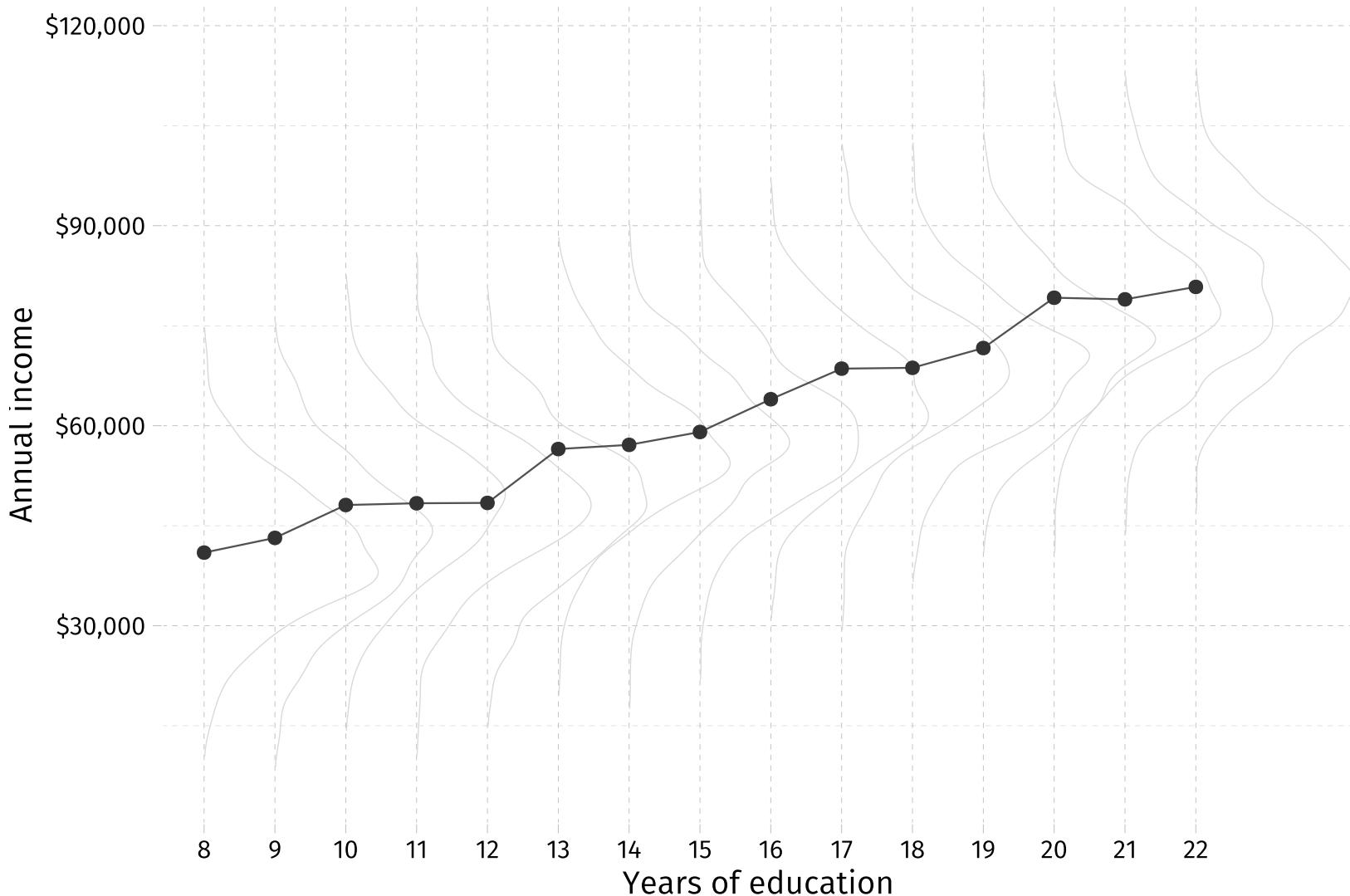
Motivation (3) tends to be the most compelling.

Even when the CEF is not linear, regression recovers the best linear approximation to the CEF.

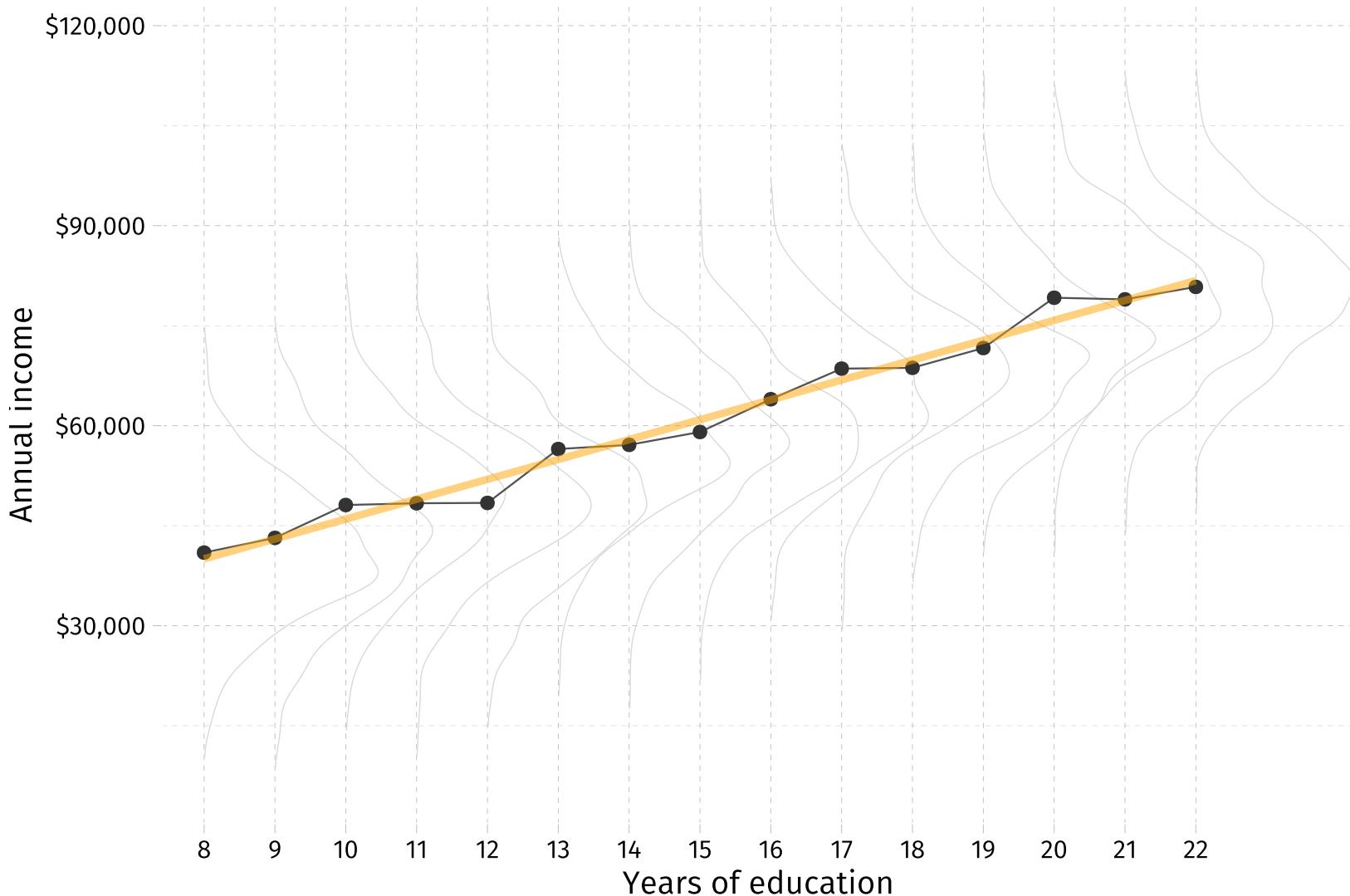
The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships without necessarily trying to pin them down exactly. (MHE, p.39, emphasis added)

Let's dig into this linear-approximate to the CEF a little more...

## Returning to our **CEF**



## Adding the population regression function



# Regression

## Regression and the CEF

As the previous figure suggests, one way to think about least-squares regression is **estimating a weighted regression on the CEF** rather than the individual observations.

TLDR Use  $E[\mathbf{Y}_i \mid \mathbf{X}_i]$  as the outcome, rather than  $\mathbf{Y}_i$ , and properly weight.

Suppose  $\mathbf{X}_i$  is discrete with pmf  $g_x(u)$

$$E\left[\left(E[\mathbf{Y}_i \mid \mathbf{X}_i] - \mathbf{X}'_i b\right)^2\right] = \sum_u \left(E[\mathbf{Y}_i \mid \mathbf{X}_i = u] - u'b\right)^2 g_x(u)$$

i.e.,  $\beta$  can be expressed as weighted-least squares regression of  $E[\mathbf{Y}_i \mid \mathbf{X}_i = u]$  on  $u$  (the values of  $\mathbf{X}_i$ ) weighted by  $g_x(u)$ .

# Regression

## Regression and the *CEF*

We can also use LIE here

$$\begin{aligned}\beta &= E \left[ X_i X'_i \right]^{-1} E[X_i Y_i] \\ &= E \left[ X_i X'_i \right]^{-1} E[X_i E(Y_i | X_i)]\end{aligned}$$

**Pro** Useful for aggregated data when microdata are sensitive/big.

**Con** You **will not** get the same standard errors.

# Table of contents

## Admin

1. Schedule
2. `return`

## Regression

1. Why?
2. The CEF
  - Definition
  - Graphically
  - Law of iterated expectations
  - Decomposition
  - Prediction
3. Population least squares
4. Anatomy
5. Regression-CEF theorem
6. WLS