Unit 1. Introduction

Conditional expectations and linear projections

Henry Redondo

Dpto. de Fundamentos del Análisis Económico, Universidad de Alicante

Econometrics II, 2026/27

Motivation

Does one more year of education cause higher wages?

- Wages and education are clearly correlated.
- But more educated workers also differ in ability, family background, region, …
- Before we can talk about estimators, we need precise language for what we want to estimate: the conditional expectation $E(y \mid \mathbf{x})$ and the linear projection $L(y \mid 1, \mathbf{x})$.

Outline

1. Introduction

2. Conditional Expectation

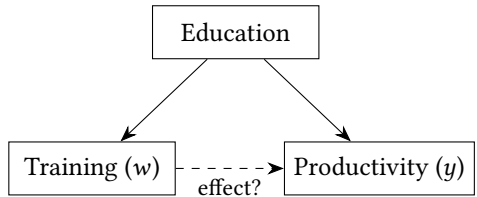
3. Linear Projections

Causal relationships

- A central goal of empirical economics: does a change in w cause a change in y?
  - Does an additional year of education increase wages?
  - Does a training programme increase workers' productivity?
- **Ceteris paribus**: holding all other relevant factors fixed. This is what a causal statement requires; a correlation between two variables is, in general, not enough.
- We focus on the average response. A ceteris paribus analysis requires $E(y \mid w, \mathbf{c})$, where $\mathbf{c}$ is the vector of **controls** held fixed while we vary w.

Why we need controls

- We control for **c** because w is correlated with other factors that also influence y.
- Training example: education affects both
  - the probability of joining the programme, and
  - productivity itself.
- If we ignore education, part of its effect is wrongly attributed to training.



What can go wrong

1. **Choice of controls.** Which is the appropriate set of controls $\mathbf{c}$?
2. **Unobservables.** Once $\mathbf{c}$ is chosen, some of its elements may be unobservable (e.g. ability).
3. **Measurement error.** Some variables may be measured with error (e.g. self-reported hours or income).
4. **Simultaneity.** We may only observe equilibrium values of w and y when both are determined jointly (e.g. prices and quantities).

Much of Econometrics II is about tools to deal with these problems.

Data structures

- **Time series**: one unit, many periods (Spain's quarterly unemployment rate).
- **Cross section**: many units, one period (workers surveyed in a single quarter).
- **Panel data**: many units, many periods (workers' employment histories over years).

In most of the course we rely on **random sampling**:

- there is a population model of interest that has been specified;
- an independent and identically distributed (iid) sample can be drawn from that population.

Random sampling is often reasonable for cross-section data.

Exercise. Why is random sampling usually not a good description of time-series data?

The conditional expectation

Many applied studies estimate or test hypotheses about the expectation of y conditional on $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$.

- y: explained, dependent or response variable.
- $\mathbf{x}$: explanatory variables, independent variables, regressors, covariates or controls.

$E(y \mid \mathbf{x})$ is the **mean of the conditional distribution** of y given $\mathbf{x}$: integrate with the conditional pdf if y is continuous, sum with the conditional pmf if it is discrete.

In a sample: group people with the same $\mathbf{x}$ (e.g. years of education) and average y within each group.

The conditional expectation as a random variable

- The conditional expectation
$$\mu(\mathbf{x}) = E(y \mid \mathbf{x})$$
is a function of $\mathbf{x}$; since $\mathbf{x}$ is a random vector, $\mu(\mathbf{x})$ is itself a random variable.
- Often we assume $\mu(\mathbf{x})$ depends on a finite number of unknown parameters: a **parametric model**. The goal is to estimate those parameters.

Three models used throughout the unit:

$$E(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$
$$E(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \tag{2}$$
$$E(y \mid x_1, x_2) = \exp\left(\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2\right) \tag{3}$$

Partial effects

The **partial effect** of a continuous x_j on $E(y \mid \mathbf{x})$ (loosely, "on y") is

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_j}.$$

For a discrete x_j, compare $E(y \mid \mathbf{x})$ at different values of x_j. If x_j is binary:

$$E(y \mid x_1, .., x_{j-1}, x_j = 1, x_{j+1}, .., x_k) - E(y \mid x_1, .., x_{j-1}, x_j = 0, x_{j+1}, .., x_k)$$

In both cases $x_1, .., x_{j-1}, x_{j+1}, .., x_k$ are held fixed: this is ceteris paribus, formally.

Partial effects in models (1)–(3)

Assume x_1 is continuous. The partial effect of x_1 on $E(y \mid \mathbf{x})$ is:

	Model	$\dfrac{\partial E(y \mid \mathbf{x})}{\partial x_1}$
(1)	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	β_1
(2)	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$	$\beta_1 + 2\beta_3 x_1 + \beta_5 x_2$
(3)	$\exp(\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2)$	$\exp(\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2)\,\dfrac{\beta_1}{x_1}$

Outside model (1) the partial effect is not a single number: we must say at which $\mathbf{x}$ we evaluate it (e.g. at the means).

Elasticities

The **partial elasticity** of $E(y \mid \mathbf{x})$ with respect to a continuous x_j, holding the other regressors fixed, is

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_j} \frac{x_j}{E(y \mid \mathbf{x})}.$$

If $E(y \mid \mathbf{x}) > 0$ and $x_j > 0$:

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_j} \frac{x_j}{E(y \mid \mathbf{x})} = \frac{\partial \log E(y \mid \mathbf{x})}{\partial \log x_j}$$

It is the approximate percentage change in $E(y \mid \mathbf{x})$ when x_j increases by 1%, holding $x_1, .., x_{j-1}, x_{j+1}, .., x_k$ fixed.

Elasticities in models (1) and (3)

Model (1). $E(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

$$\frac{\partial E(y \mid \mathbf{x})}{\partial x_1} \frac{x_1}{E(y \mid \mathbf{x})} = \frac{\beta_1 x_1}{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

It depends on where we evaluate it.

Model (3).
$E(y \mid x_1, x_2) = \exp(\beta_0 + \beta_1 \log(x_1) + \beta_2 x_2)$

$$\frac{\partial \log E(y \mid \mathbf{x})}{\partial \log x_1} = \beta_1$$

A constant elasticity: we read it directly off the coefficient.

Linearity

Property 1. The conditional expectation is a linear operator:

$$E\Big(\sum_{j=1}^{J} a_j(\mathbf{x})\, y_j + b(\mathbf{x}) \,\Big|\, \mathbf{x} \Big) = \sum_{j=1}^{J} a_j(\mathbf{x})\, E(y_j \mid \mathbf{x}) + b(\mathbf{x})$$

Once we condition on $\mathbf{x}$, any function of $\mathbf{x}$ behaves like a constant: it can be taken outside the expectation.

Law of iterated expectations

Property 2 (LIE). Let $\mathbf{w}$ be a random vector, y a random variable and $\mathbf{x} = h(\mathbf{w})$ a function of $\mathbf{w}$. Then

$$\boxed{E(y \mid \mathbf{x}) = E\big[E(y \mid \mathbf{w}) \mid \mathbf{x}\big]} \tag{4}$$

Each expectation uses a different distribution:

Expectation	Computed with the distribution of…
$E(y \mid \mathbf{x})$	y given $\mathbf{x}$
$E(y \mid \mathbf{w})$	y given $\mathbf{w}$
$E[E(y \mid \mathbf{w}) \mid \mathbf{x}]$	$\mathbf{w}$ given $\mathbf{x}$

The LIE in practice

- In particular, if $\mathbf{x}$ is a subset of $\mathbf{w}$, i.e. $\mathbf{w} = (\mathbf{x}', \mathbf{z}')'$:

$$E(y \mid \mathbf{x}) = E\big[E(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\big]$$

- Often $E(y \mid \mathbf{x}, \mathbf{z})$ is the **structural** conditional expectation, but $\mathbf{z}$ is unobserved. We want the partial effect of x_j on $E(y \mid \mathbf{x}, \mathbf{z})$ holding $\mathbf{z}$ and the other elements of $\mathbf{x}$ fixed.
- We cannot estimate $E(y \mid \mathbf{x}, \mathbf{z})$ directly, but we can estimate $E(y \mid \mathbf{x})$, and the LIE tells us how the two are related.

The unconditional LIE

The simplest statement of the LIE, from Statistics:

$$E(y) = E\big[E(y \mid \mathbf{x})\big]$$

Expectation	Computed with the distribution of…
$E(y \mid \mathbf{x})$	y given $\mathbf{x}$ (conditional)
$E(y)$	y (marginal)
$E[E(y \mid \mathbf{x})]$	$\mathbf{x}$ (marginal)

Example. If 40% of workers are graduates (mean wage 30) and 60% are not (mean wage 20), then $E(wage) = 0.4 \cdot 30 + 0.6 \cdot 20 = 24$: the overall mean is the weighted average of cell means.

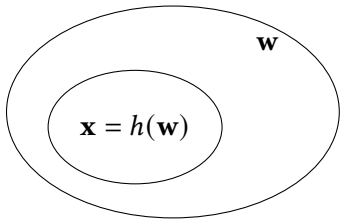
Conditioning on the smaller information set

If $\mathbf{x} = h(\mathbf{w})$, then

$$E(y \mid \mathbf{x}) = E\big[E(y \mid \mathbf{x}) \mid \mathbf{w}\big] \tag{5}$$

Proof. $E(y \mid \mathbf{x})$ is a function of $\mathbf{x}$. The outcome of $\mathbf{x}$ is known when $\mathbf{w}$ is known, so the outcome of $E(y \mid \mathbf{x})$ is known when $\mathbf{w}$ is known. Hence $E(y \mid \mathbf{x}) = E[E(y \mid \mathbf{x}) \mid \mathbf{w}]$. ■

In both (4) and (5) the answer is conditioned on the smaller information set, $\mathbf{x}$.



Knowing $\mathbf{w} \Rightarrow$ knowing $\mathbf{x}$, but not the other way round

Models linear in parameters

- Consider $E(y \mid \mathbf{x}) = \beta_0 + \beta_1 g_1(\mathbf{x}) + \cdots + \beta_M g_M(\mathbf{x})$, where $g_1, \ldots, g_M$ are known functions.
- Define $\mathbf{z} = (g_1(\mathbf{x}), \ldots, g_M(\mathbf{x}))' = (z_1, \ldots, z_M)'$.
- Using the LIE (since $\mathbf{z}$ is a function of $\mathbf{x}$):

$$E(y \mid \mathbf{z}) = \beta_0 + \beta_1 z_1 + \cdots + \beta_M z_M$$

So any model linear in parameters can be rewritten as a conditional expectation that is linear in parameters and in some conditioning variables. Models like (2), with squares and interactions, can be treated as linear models in $\mathbf{z}$.

The best predictor

Property 3. $E(y \mid \mathbf{x})$ is the **best mean square predictor** of y given $\mathbf{x}$:

$$E(y \mid \mathbf{x}) = \arg\min_{c(\mathbf{x})} \; E\left[(y - c(\mathbf{x}))^2\right]$$

Among all functions of $\mathbf{x}$ we could use to predict y, none beats the conditional mean in mean squared error. Section 3 restricts attention to linear functions.

Proof of the best-predictor property

Let $m(\mathbf{x}) \equiv E(y \mid \mathbf{x})$ and $u \equiv y - m(\mathbf{x})$, so that $E(u \mid \mathbf{x}) = 0$.

Claim: For any function $c(\cdot)$, $E[(y - c(\mathbf{x}))^2] \geq E[(y - m(\mathbf{x}))^2]$.

Proof of the best-predictor property

Let $m(\mathbf{x}) \equiv E(y \mid \mathbf{x})$ and $u \equiv y - m(\mathbf{x})$, so that $E(u \mid \mathbf{x}) = 0$.

Claim: For any function $c(\cdot)$, $E[(y - c(\mathbf{x}))^2] \geq E[(y - m(\mathbf{x}))^2]$.

Step 1: Add and subtract $m(\mathbf{x})$, then expand.

$$\begin{aligned}(y - c(\mathbf{x}))^2 &= \big[u + \big(m(\mathbf{x}) - c(\mathbf{x})\big)\big]^2 \\ &= u^2 + 2u\big(m(\mathbf{x}) - c(\mathbf{x})\big) + \big(m(\mathbf{x}) - c(\mathbf{x})\big)^2\end{aligned}$$

Proof of the best-predictor property

Let $m(\mathbf{x}) \equiv E(y \mid \mathbf{x})$ and $u \equiv y - m(\mathbf{x})$, so that $E(u \mid \mathbf{x}) = 0$.

Claim: For any function $c(\cdot)$, $E[(y - c(\mathbf{x}))^2] \geq E[(y - m(\mathbf{x}))^2]$.

Step 1: Add and subtract $m(\mathbf{x})$, then expand.

$$\begin{aligned}(y - c(\mathbf{x}))^2 &= \big[u + \big(m(\mathbf{x}) - c(\mathbf{x})\big)\big]^2 \\ &= u^2 + 2u\big(m(\mathbf{x}) - c(\mathbf{x})\big) + \big(m(\mathbf{x}) - c(\mathbf{x})\big)^2\end{aligned}$$

Step 2: Condition on x. Given $\mathbf{x}$, $m(\mathbf{x}) - c(\mathbf{x})$ is fixed, so the cross term vanishes:

$$\begin{aligned}E\big[(y - c(\mathbf{x}))^2 \mid \mathbf{x}\big] &= E(u^2 \mid \mathbf{x}) + 2\big(m(\mathbf{x}) - c(\mathbf{x})\big)\underbrace{E(u \mid \mathbf{x})}_{=0} + \big(m(\mathbf{x}) - c(\mathbf{x})\big)^2 \\ &= E(u^2 \mid \mathbf{x}) + \big(m(\mathbf{x}) - c(\mathbf{x})\big)^2\end{aligned}$$

Proof of the best-predictor property

Step 3: Minimize for each x. $E(u^2 \mid \mathbf{x})$ does not depend on $c(\cdot)$, and the second term is non-negative and equals zero if and only if $c(\mathbf{x}) = m(\mathbf{x})$.

Proof of the best-predictor property

Step 3: Minimize for each x. $E(u^2 \mid \mathbf{x})$ does not depend on $c(\cdot)$, and the second term is non-negative and equals zero if and only if $c(\mathbf{x}) = m(\mathbf{x})$.

Step 4: Average over x. By the LIE,

$$E\left[(y - c(\mathbf{x}))^2\right] = E(u^2) + E\left[\left(m(\mathbf{x}) - c(\mathbf{x})\right)^2\right] \geq E(u^2) = E\left[(y - m(\mathbf{x}))^2\right] \qquad \blacksquare$$

Error form of the conditional expectation

We can always write

$$\boxed{y = E(y \mid \mathbf{x}) + u, \qquad \text{where } E(u \mid \mathbf{x}) = 0}$$

u is called the **disturbance** or **error term**.

$E(u \mid \mathbf{x}) = 0$ implies

- $E(u) = 0$;
- $E(h(\mathbf{x})\, u) = 0$ for any function h: u is uncorrelated with any function of $\mathbf{x}$.

It does *not* imply that u and $\mathbf{x}$ are independent: $\text{var}(u \mid \mathbf{x})$ can still depend on $\mathbf{x}$ (heteroskedasticity).

The conditional variance

The variance of the conditional distribution of y given $\mathbf{x}$ is the **conditional variance**

$$\operatorname{var}(y \mid \mathbf{x}) = E\left[(y - E(y \mid \mathbf{x}))^2 \mid \mathbf{x}\right] = \sigma^2(\mathbf{x}).$$

Property 1. $\operatorname{var}\left(a(\mathbf{x})\, y + b(\mathbf{x}) \mid \mathbf{x}\right) = a(\mathbf{x})^2 \operatorname{var}(y \mid \mathbf{x})$.

Proof. Let $w \equiv a(\mathbf{x})y + b(\mathbf{x})$. By linearity, $E(w \mid \mathbf{x}) = a(\mathbf{x})E(y \mid \mathbf{x}) + b(\mathbf{x})$, so

$$w - E(w \mid \mathbf{x}) = a(\mathbf{x})\big(y - E(y \mid \mathbf{x})\big).$$

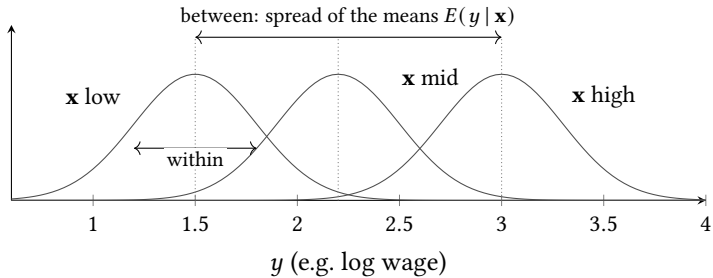
Squaring and taking expectations given $\mathbf{x}$, where $a(\mathbf{x})^2$ is fixed:

$$\operatorname{var}(w \mid \mathbf{x}) = E\left[(w - E(w \mid \mathbf{x}))^2 \mid \mathbf{x}\right] = a(\mathbf{x})^2 E\left[(y - E(y \mid \mathbf{x}))^2 \mid \mathbf{x}\right] = a(\mathbf{x})^2 \operatorname{var}(y \mid \mathbf{x}).$$

Variance decomposition

Property 2 (law of total variance).

$$\operatorname{var}(y) = \underbrace{E\big[\operatorname{var}(y \mid \mathbf{x})\big]}_{\text{within}} + \underbrace{\operatorname{var}\big[E(y \mid \mathbf{x})\big]}_{\text{between}}$$



Proof of the variance decomposition

Using the LIE, $\text{var}(y) = E\left[(y - E(y))^2\right] = E\left[E[(y - E(y))^2 \mid \mathbf{x}]\right]$. (7)

Proof of the variance decomposition

Using the LIE, $\operatorname{var}(y) = E\big[(y - E(y))^2\big] = E\big[E[(y - E(y))^2 \mid \mathbf{x}]\big]$. (7)

Adding and subtracting $E(y \mid \mathbf{x})$:

$$
\begin{aligned}
E[(y - E(y))^2 \mid \mathbf{x}] &= E[(y - E(y \mid \mathbf{x}))^2 \mid \mathbf{x}] + E[(E(y \mid \mathbf{x}) - E(y))^2 \mid \mathbf{x}] \\
&\quad + 2E[(y - E(y \mid \mathbf{x}))(E(y \mid \mathbf{x}) - E(y)) \mid \mathbf{x}] \\
&= \operatorname{var}(y \mid \mathbf{x}) + (E(y \mid \mathbf{x}) - E(y))^2 + 2 \underbrace{E[y - E(y \mid \mathbf{x}) \mid \mathbf{x}]}_{= E(y \mid \mathbf{x}) - E(y \mid \mathbf{x}) = 0} (E(y \mid \mathbf{x}) - E(y))
\end{aligned}
$$

Proof of the variance decomposition

Using the LIE, $\text{var}(y) = E\big[(y - E(y))^2\big] = E\big[E[(y - E(y))^2 \mid \mathbf{x}]\big]$. (7)
Adding and subtracting $E(y \mid \mathbf{x})$:

$$\begin{aligned} E[(y - E(y))^2 \mid \mathbf{x}] &= E[(y - E(y \mid \mathbf{x}))^2 \mid \mathbf{x}] + E[(E(y \mid \mathbf{x}) - E(y))^2 \mid \mathbf{x}] \\ &\quad + 2E[(y - E(y \mid \mathbf{x}))(E(y \mid \mathbf{x}) - E(y)) \mid \mathbf{x}] \\ &= \text{var}(y \mid \mathbf{x}) + (E(y \mid \mathbf{x}) - E(y))^2 + 2 \underbrace{E[y - E(y \mid \mathbf{x}) \mid \mathbf{x}]}_{= E(y \mid \mathbf{x}) - E(y \mid \mathbf{x}) = 0} (E(y \mid \mathbf{x}) - E(y)) \end{aligned}$$

Substituting in (7):

$$\text{var}(y) = E[\text{var}(y \mid \mathbf{x})] + E[(E(y \mid \mathbf{x}) - E(y))^2] = E[\text{var}(y \mid \mathbf{x})] + \text{var}[E(y \mid \mathbf{x})] \quad \blacksquare$$

Consequences of the variance decomposition

- Since $\text{var}(E(y \mid \mathbf{x})) \geq 0$:

$$\text{var}(y) \geq E[\text{var}(y \mid \mathbf{x})]$$

- Moreover, if $\text{var}(y \mid \mathbf{x})$ does not depend on $\mathbf{x}$, the variance of the conditional distribution is no larger than that of the marginal distribution:

$$\text{var}(y) \geq \text{var}(y \mid \mathbf{x})$$

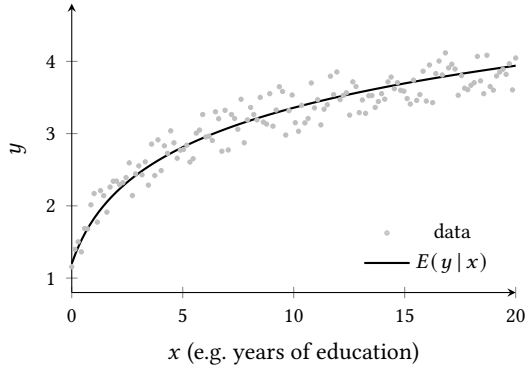
On average, conditioning on information reduces uncertainty. The share $\text{var}[E(y \mid \mathbf{x})]/\text{var}(y)$ is the population analogue of an R^2 for the best predictor.

Summary

1. $E(y \mid \mathbf{x})$ is a function of $\mathbf{x}$ and hence a random variable.
2. Partial effects and elasticities are all derivatives of $E(y \mid \mathbf{x})$ (or its log) holding the rest fixed; only in special cases are they constants.
3. LIE: $E(y \mid \mathbf{x}) = E[E(y \mid \mathbf{w}) \mid \mathbf{x}]$; the smaller information set dominates.
4. $E(y \mid \mathbf{x})$ is the best mean square predictor of y.
5. $y = E(y \mid \mathbf{x}) + u$ with $E(u \mid \mathbf{x}) = 0$, which makes u uncorrelated with any function of $\mathbf{x}$.
6. $\text{var}(y) = E[\text{var}(y \mid \mathbf{x})] + \text{var}[E(y \mid \mathbf{x})]$.

From the best predictor to the best linear predictor

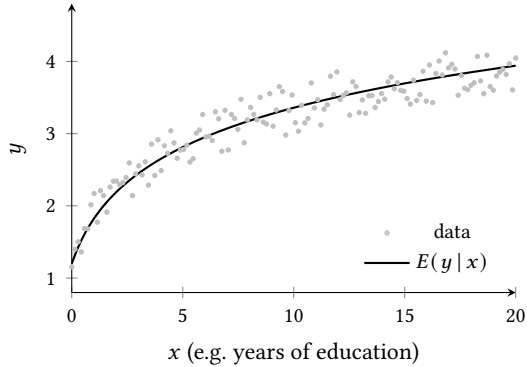
- $E(y \mid \mathbf{x})$ is the best mean square predictor of y given $\mathbf{x}$…



Simulated data.

From the best predictor to the best linear predictor

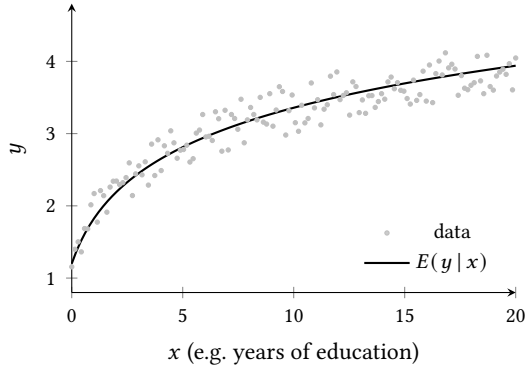
- $E(y \mid \mathbf{x})$ is the best mean square predictor of y given $\mathbf{x}$…
- …but it may be nonlinear and hard to estimate.



Simulated data.

From the best predictor to the best linear predictor

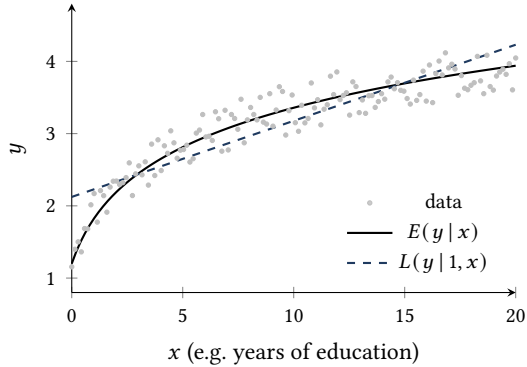
- $E(y \mid \mathbf{x})$ is the best mean square predictor of y given $\mathbf{x}$…
- …but it may be nonlinear and hard to estimate.
- Now we look for the best mean square predictor among linear functions of $\mathbf{x}$.



Simulated data.

From the best predictor to the best linear predictor

- $E(y \mid \mathbf{x})$ is the best mean square predictor of y given $\mathbf{x}$…
- …but it may be nonlinear and hard to estimate.
- Now we look for the best mean square predictor among linear functions of $\mathbf{x}$.
- The **linear projection** is the best linear approximation to the CEF.



Simulated data.

Deriving the linear projection

- The minimisation problem:

$$\min_{b_0,\mathbf{b}} S(b_0,\mathbf{b}) = \min_{b_0,\mathbf{b}} E\left[(y - b_0 - \mathbf{x}'\mathbf{b})^2\right]$$

- Partial derivatives of the objective function:

$$\frac{\partial S(b_0,\mathbf{b})}{\partial b_0} = -2E(y - b_0 - \mathbf{x}'\mathbf{b})$$
$$\frac{\partial S(b_0,\mathbf{b})}{\partial \mathbf{b}} = -2E\big(\mathbf{x}(y - b_0 - \mathbf{x}'\mathbf{b})\big)$$

This is the problem OLS solves, with population expectations in place of sample averages.

Solving the first-order conditions

- The FOC are

$$E(y) = \beta_0 + E(\mathbf{x})'\boldsymbol{\beta} \tag{8}$$
$$E(\mathbf{x}y) = E(\mathbf{x})\beta_0 + E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} \tag{9}$$

Solving the first-order conditions

- The FOC are

$$E(y) = \beta_0 + E(\mathbf{x})'\boldsymbol{\beta} \tag{8}$$
$$E(\mathbf{x}y) = E(\mathbf{x})\beta_0 + E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} \tag{9}$$

- Multiplying (8) by $E(\mathbf{x})$:

$$E(\mathbf{x})E(y) = E(\mathbf{x})\beta_0 + E(\mathbf{x})E(\mathbf{x})'\boldsymbol{\beta} \tag{10}$$

Solving the first-order conditions

- The FOC are

$$E(y) = \beta_0 + E(\mathbf{x})'\boldsymbol{\beta} \quad (8)$$
$$E(\mathbf{x}y) = E(\mathbf{x})\beta_0 + E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} \quad (9)$$

- Multiplying (8) by $E(\mathbf{x})$:

$$E(\mathbf{x})E(y) = E(\mathbf{x})\beta_0 + E(\mathbf{x})E(\mathbf{x})'\boldsymbol{\beta} \quad (10)$$

Subtracting (10) from (9):

$$E(\mathbf{x}y) - E(\mathbf{x})E(y) = \left(E(\mathbf{x}\mathbf{x}') - E(\mathbf{x})E(\mathbf{x})'\right)\boldsymbol{\beta}$$
$$\Leftrightarrow \quad \operatorname{cov}(\mathbf{x}, y) = \operatorname{var}(\mathbf{x})\,\boldsymbol{\beta}$$

Solving the first-order conditions

- The FOC are

$$E(y) = \beta_0 + E(\mathbf{x})'\boldsymbol{\beta} \tag{8}$$
$$E(\mathbf{x}y) = E(\mathbf{x})\beta_0 + E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} \tag{9}$$

- Multiplying (8) by $E(\mathbf{x})$:

$$E(\mathbf{x})E(y) = E(\mathbf{x})\beta_0 + E(\mathbf{x})E(\mathbf{x})'\boldsymbol{\beta} \tag{10}$$

Subtracting (10) from (9):

$$E(\mathbf{x}y) - E(\mathbf{x})E(y) = \left(E(\mathbf{x}\mathbf{x}') - E(\mathbf{x})E(\mathbf{x})'\right)\boldsymbol{\beta}$$
$$\Leftrightarrow \quad \operatorname{cov}(\mathbf{x}, y) = \operatorname{var}(\mathbf{x})\,\boldsymbol{\beta}$$

If $\operatorname{var}(\mathbf{x})$ is non-singular:

$$\boldsymbol{\beta} = \operatorname{var}(\mathbf{x})^{-1}\operatorname{cov}(\mathbf{x}, y),$$
$$\beta_0 = E(y) - E(\mathbf{x})'\boldsymbol{\beta}$$

The linear projection

The **linear projection** (LP), or **best linear predictor** (BLP), of y on $1, \mathbf{x}$ is

$$L(y \mid 1, \mathbf{x}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} \qquad \text{with} \qquad \boldsymbol{\beta} = \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}(\mathbf{x}, y), \quad \beta_0 = E(y) - E(\mathbf{x})'\boldsymbol{\beta}$$

With a scalar x:

$$L(y \mid 1, x) = E(y) - \beta E(x) + \beta x = E(y) + \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} \big(x - E(x)\big)$$

The slope is covariance over variance, and the line passes through $(E(x), E(y))$.

Compact notation

- Include the constant in $\mathbf{w} = (1, \mathbf{x}')'$. The problem becomes

$$\min_{\mathbf{d}} \; S(\mathbf{d}) = \min_{\mathbf{d}} \; E\big[(y - \mathbf{w}'\mathbf{d})^2\big], \qquad \frac{\partial S(\mathbf{d})}{\partial \mathbf{d}} = -2E\big(\mathbf{w}(y - \mathbf{w}'\mathbf{d})\big)$$

- The FOC are $E(\mathbf{w}y) = E(\mathbf{w}\mathbf{w}')\boldsymbol{\delta}$, and so

$$\boldsymbol{\delta} = E(\mathbf{w}\mathbf{w}')^{-1}E(\mathbf{w}y), \qquad \boldsymbol{\delta} = (\beta_0, \boldsymbol{\beta}')'$$

Compare with OLS: $\hat{\boldsymbol{\delta}} = \left(\frac{1}{n}\sum_i \mathbf{w}_i\mathbf{w}_i'\right)^{-1}\frac{1}{n}\sum_i \mathbf{w}_i y_i$. OLS is the **sample analogue** of the LP.

Linear projection without a constant

The LP of y on $\mathbf{x}$ without a constant is

$$L(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\gamma}, \qquad \boldsymbol{\gamma} = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}y).$$

In general $\boldsymbol{\gamma} \neq \boldsymbol{\beta}$.

Without a constant the line is forced through the origin, so it uses raw second moments $E(\mathbf{x}\mathbf{x}')$, $E(\mathbf{x}y)$ instead of variances and covariances. The two coincide only in special cases (e.g. $E(\mathbf{x}) = \mathbf{0}$ and $E(y) = 0$).

The linear projection and the CEF

Property 1. The LP is the best linear approximation to the conditional expectation: it minimises

$$E\left[\left(E(y \mid \mathbf{x}) - b_0 - \mathbf{x}'\mathbf{b}\right)^2\right].$$

Property 2. If the conditional expectation is linear, it coincides with the LP.

So even when the CEF is nonlinear, OLS estimates something meaningful: the best linear approximation to it (the dashed line in the figure that opened this section).

Linearity of the linear projection

Property 3. The LP is a linear operator:

$$L\Big(\sum_{j=1}^{J} a_j y_j \,\Big|\, 1, \mathbf{x}\Big) = \sum_{j=1}^{J} a_j \, L(y_j \mid 1, \mathbf{x})$$

Proof. For each j, $L(y_j \mid 1, \mathbf{x}) = \beta_0^j + \mathbf{x}'\boldsymbol{\beta}^j$, where

$$\boldsymbol{\beta}^j = \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}(\mathbf{x}, y_j), \qquad \beta_0^j = E(y_j) - E(\mathbf{x})'\boldsymbol{\beta}^j.$$

Linearity of the linear projection, proof continued

Write $L\left(\sum_j a_j y_j \mid 1, \mathbf{x}\right) = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}$. (11)

Linearity of the linear projection, proof continued

Write $L\left(\sum_j a_j y_j \mid 1, \mathbf{x}\right) = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}$. (11)

$$\boldsymbol{\gamma} = \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}\left(\mathbf{x}, \sum_j a_j y_j\right) = \sum_j a_j \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}(\mathbf{x}, y_j) = \sum_j a_j \boldsymbol{\beta}^j$$

$$\gamma_0 = E\left(\sum_j a_j y_j\right) - E(\mathbf{x})'\boldsymbol{\gamma} = \sum_j a_j \left(E(y_j) - E(\mathbf{x})'\boldsymbol{\beta}^j\right) = \sum_j a_j \beta_0^j$$

Linearity of the linear projection, proof continued

Write $L\left(\sum_j a_j y_j \mid 1, \mathbf{x}\right) = \gamma_0 + \mathbf{x}'\boldsymbol{\gamma}$. (11)

$$\boldsymbol{\gamma} = \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}\left(\mathbf{x}, \sum_j a_j y_j\right) = \sum_j a_j \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}(\mathbf{x}, y_j) = \sum_j a_j \boldsymbol{\beta}^j$$

$$\gamma_0 = E\left(\sum_j a_j y_j\right) - E(\mathbf{x})'\boldsymbol{\gamma} = \sum_j a_j \left(E(y_j) - E(\mathbf{x})'\boldsymbol{\beta}^j\right) = \sum_j a_j \beta_0^j$$

Substituting γ_0 and $\boldsymbol{\gamma}$ in (11):

$$L\left(\sum_j a_j y_j \,\middle|\, 1, \mathbf{x}\right) = \sum_j a_j \beta_0^j + \mathbf{x}' \sum_j a_j \boldsymbol{\beta}^j = \sum_j a_j \left(\beta_0^j + \mathbf{x}'\boldsymbol{\beta}^j\right) = \sum_j a_j L(y_j \mid 1, \mathbf{x}) \qquad \blacksquare$$

Law of iterated projections

Property 4. Let $\mathbf{x}$ be $k \times 1$ (first element typically 1), $\mathbf{z}$ be $J \times 1$ and y a scalar. Then

$$L(y \mid \mathbf{x}) = L\big[L(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\big]$$

This is the linear analogue of the LIE, and it underlies the omitted variable bias formula.

Proof. $L(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\delta}$ with $\boldsymbol{\delta} = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}y)$, and $L(y \mid \mathbf{x}, \mathbf{z}) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{z}'\boldsymbol{\gamma}$ with

$$\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix} = \begin{pmatrix}E(\mathbf{x}\mathbf{x}') & E(\mathbf{x}\mathbf{z}')\\ E(\mathbf{z}\mathbf{x}') & E(\mathbf{z}\mathbf{z}')\end{pmatrix}^{-1}\begin{pmatrix}E(\mathbf{x}y)\\ E(\mathbf{z}y)\end{pmatrix}$$

Law of iterated projections, proof continued

By linearity of the LP, $L\left[L(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta} + L(\mathbf{z}' \mid \mathbf{x})\,\boldsymbol{\gamma}$.

Law of iterated projections, proof continued

By linearity of the LP, $L\left[L(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta} + L(\mathbf{z}' \mid \mathbf{x})\,\boldsymbol{\gamma}$.
For each element z_j of $\mathbf{z}$: $L(z_j \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\pi}_j$, with $\boldsymbol{\pi}_j = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}z_j)$. Then

$$L(\mathbf{z}' \mid \mathbf{x}) = (\mathbf{x}'\boldsymbol{\pi}_1, \ldots, \mathbf{x}'\boldsymbol{\pi}_J) = \mathbf{x}'(\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_J) = \mathbf{x}'\boldsymbol{\pi}, \qquad \boldsymbol{\pi} = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}\mathbf{z}')$$

Law of iterated projections, proof continued

By linearity of the LP, $L\left[L(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\right] = \mathbf{x}'\boldsymbol{\beta} + L(\mathbf{z}' \mid \mathbf{x})\,\boldsymbol{\gamma}$.
For each element z_j of $\mathbf{z}$: $L(z_j \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\pi}_j$, with $\boldsymbol{\pi}_j = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}z_j)$. Then

$$L(\mathbf{z}' \mid \mathbf{x}) = (\mathbf{x}'\boldsymbol{\pi}_1, \ldots, \mathbf{x}'\boldsymbol{\pi}_J) = \mathbf{x}'(\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_J) = \mathbf{x}'\boldsymbol{\pi}, \qquad \boldsymbol{\pi} = E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}\mathbf{z}')$$

Therefore

$$L\left[L(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\right] = \mathbf{x}'(\boldsymbol{\beta} + \boldsymbol{\pi}\boldsymbol{\gamma})$$

and it remains to show that $\boldsymbol{\beta} + \boldsymbol{\pi}\boldsymbol{\gamma} = \boldsymbol{\delta}$.

Note that $\boldsymbol{\delta} = \boldsymbol{\beta} + \boldsymbol{\pi}\boldsymbol{\gamma}$ is the omitted variables formula: the "short" coefficient equals the "long" one plus (effect of $\mathbf{z}$) $\times$ (projection of $\mathbf{z}$ on $\mathbf{x}$).

Law of iterated projections, last step

$$
\begin{aligned}
\boldsymbol{\beta} + \boldsymbol{\pi}\boldsymbol{\gamma} &= (\mathbf{I}_k : \boldsymbol{\pi}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} \\
&= \left(\mathbf{I}_k : E(\mathbf{x}\mathbf{x}')^{-1} E(\mathbf{x}\mathbf{z}')\right) \begin{pmatrix} E(\mathbf{x}\mathbf{x}') & E(\mathbf{x}\mathbf{z}') \\ E(\mathbf{z}\mathbf{x}') & E(\mathbf{z}\mathbf{z}') \end{pmatrix}^{-1} \begin{pmatrix} E(\mathbf{x}y) \\ E(\mathbf{z}y) \end{pmatrix} \\
&= E(\mathbf{x}\mathbf{x}')^{-1} \left(E(\mathbf{x}\mathbf{x}') : E(\mathbf{x}\mathbf{z}')\right) \begin{pmatrix} E(\mathbf{x}\mathbf{x}') & E(\mathbf{x}\mathbf{z}') \\ E(\mathbf{z}\mathbf{x}') & E(\mathbf{z}\mathbf{z}') \end{pmatrix}^{-1} \begin{pmatrix} E(\mathbf{x}y) \\ E(\mathbf{z}y) \end{pmatrix} \\
&= E(\mathbf{x}\mathbf{x}')^{-1} (\mathbf{I}_k : \mathbf{0}) \begin{pmatrix} E(\mathbf{x}y) \\ E(\mathbf{z}y) \end{pmatrix} = E(\mathbf{x}\mathbf{x}')^{-1} E(\mathbf{x}y) = \boldsymbol{\delta} \qquad \blacksquare
\end{aligned}
$$

The key step: $(E(\mathbf{x}\mathbf{x}') : E(\mathbf{x}\mathbf{z}'))$ is the first block row of the inverted matrix, so their product is $(\mathbf{I}_k : \mathbf{0})$.

Projecting the CEF

Property 5.

$$L(y \mid \mathbf{x}) = L\big(E(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\big)$$

Proof. We know $y = E(y \mid \mathbf{x}, \mathbf{z}) + u$ with $E(u \mid \mathbf{x}, \mathbf{z}) = 0$. By linearity of the LP,

$$L(y \mid \mathbf{x}) = L\big(E(y \mid \mathbf{x}, \mathbf{z}) \mid \mathbf{x}\big) + L(u \mid \mathbf{x})$$

Since $E(u \mid \mathbf{x}, \mathbf{z}) = 0$, the LIE gives $E(u \mid \mathbf{x}) = 0$, hence $E(\mathbf{x}u) = 0$ and $L(u \mid \mathbf{x}) = 0$. ■

Error form of the linear projection

If $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ where $\beta_0, \ldots, \beta_k$ are the LP coefficients of y given 1 and $\mathbf{x}$, then

$$\boxed{E(u) = 0, \qquad E(\mathbf{x}u) = 0}$$

Proof.

$$E(u) = 0 \;\Leftrightarrow\; E(y) = \beta_0 + \beta_1 E(x_1) + \cdots + \beta_k E(x_k)$$
$$E(\mathbf{x}u) = 0 \;\Leftrightarrow\; E(\mathbf{x}y) = E(\mathbf{x})\beta_0 + E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta}$$

and these are precisely the FOC of the LP, equations (8) and (9). ■

CEF error and LP error

	Conditional expectation	Linear projection
Model	$y = E(y \mid \mathbf{x}) + u$	$y = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + u$
Assumption on u	$E(u \mid \mathbf{x}) = 0$	$E(u) = 0,\ E(\mathbf{x}u) = 0$
u uncorrelated with	any function of $\mathbf{x}$	$\mathbf{x}$ (linear functions)
Holds by construction?	yes	yes
Best predictor among	all functions of $\mathbf{x}$	linear functions of $\mathbf{x}$

Summary

1. $L(y \mid 1, \mathbf{x}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta}$ with $\boldsymbol{\beta} = \operatorname{var}(\mathbf{x})^{-1} \operatorname{cov}(\mathbf{x}, y)$: the best linear predictor.
2. It is the best linear approximation to $E(y \mid \mathbf{x})$, and equals it when the CEF is linear.
3. OLS is its sample analogue, so OLS always estimates something well defined.
4. Linearity and the law of iterated projections mirror the CEF properties.
5. The LP error satisfies only $E(u) = 0$ and $E(\mathbf{x}u) = 0$, which is weaker than $E(u \mid \mathbf{x}) = 0$.

Neither object is automatically causal. Making them causal means confronting controls, unobservables, measurement error and simultaneity.