Lecture 1: Regression, identification, and the OLS estimator

Economics 527: Econometric Methods

Vadim Marmer, UBC

Motivation and data

A motivating example: Return to education

- Suppose wages are generated according to:

$$\begin{aligned}\ln \text{Wage} = \beta_1 + \beta_2 \,\text{Education} + \beta_3 \,\text{Gender} + \beta_4 \,\text{Race} \\ + \beta_5 \,\text{Urban} + \beta_6 \,\text{Experience} + U.\end{aligned}$$

- We have a sample of workers and, for each worker, we observe Wage, Education, etc.
- U collects everything besides Education, Gender, etc. that affects wages. Ability is first on that list, and it is unobserved (latent).
- The unknown coefficient β_2 measures the return to education, holding everything else fixed.
- Questions/Issues:
  - Why in this equation β_2 measures the return to education;
  - How to "learn" β_2 from random data;
  - What data randomness means;
  - What has to be assumed about the unobserved U to be able to learn β_2.

Levels and logs

- **Level-level**, $y = \beta x$:

$$\beta = \frac{dy}{dx},$$

  and the proportional change in y per unit of x is β/y, which depends on where y is.

- **Log-level**, $\ln y = \beta x$:

$$\beta = \frac{d \ln y}{dx} = \frac{dy/dx}{y},$$

  so β is the proportional rate of change of y with respect to x. Multiply by 100 for a percentage.

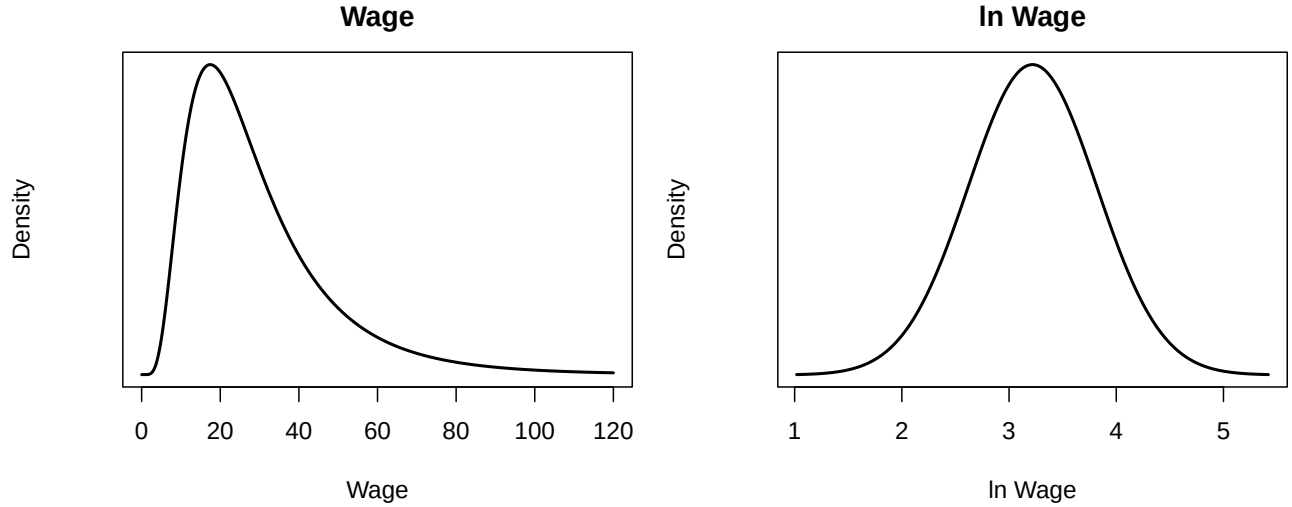
- **Log-log**, $\ln y = \beta \ln x$:

$$\beta = \frac{d \ln y}{d \ln x} = \frac{dy/y}{dx/x},$$

  the elasticity of y with respect to x.

Logs often "work better"

- Wages are bounded below by zero and have a long right tail. Their log is close to symmetric.



- A symmetric dependent variable is not required by anything that follows, but specifications with logs often "work better" in practice.

The data

- We observe a **sample** of n workers:

$$\{(Y_i, X_i) : i = 1, \dots, n\}.$$

- For worker i, $Y_i = \ln \text{Wage}_i$ is a scalar (1×1) and X_i stacks the k regressors, with a 1 in the first position for the intercept:

$$X_i = \begin{pmatrix} 1 \\ \text{Educ}_i \\ \text{Gender}_i \\ \vdots \end{pmatrix} \quad (k \times 1), \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad (k \times 1).$$

- The wage equation for worker i is

$$Y_i = X_i^\top \beta + U_i = \sum_{j=1}^{k} X_{ij}\, \beta_j + U_i,$$

where $X_i^\top \beta = \beta^\top X_i$ is a scalar.

What is random and what is fixed

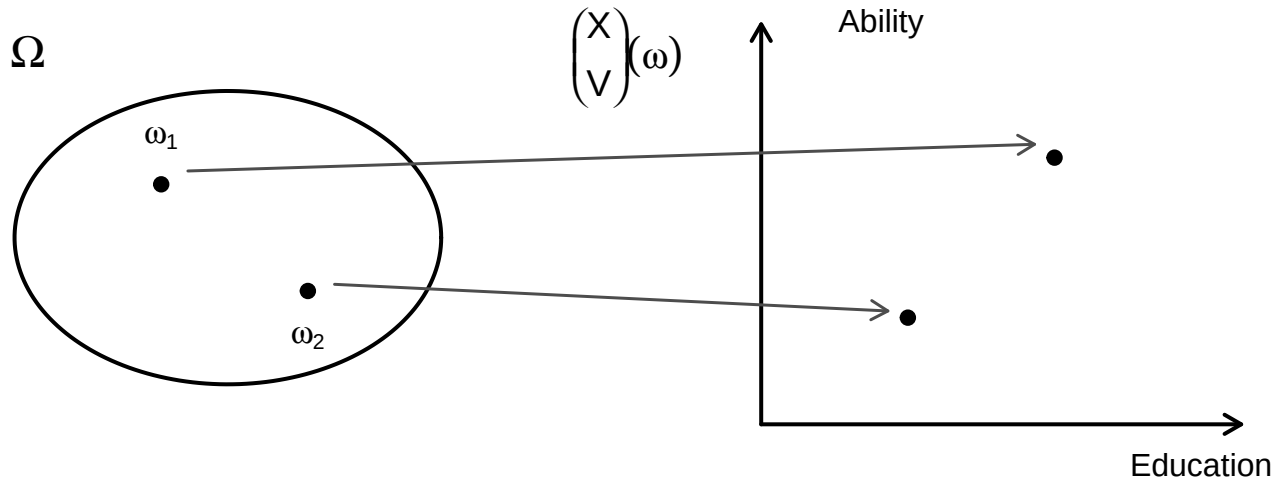
- β is **fixed** and unknown.
- X_i, U_i, and therefore Y_i are **random**.
- The **population** is the distribution that generates $\{(Y_i, X_i) : i = 1, \dots, n\}$, a probability distribution rather than a group of people.
- In a cross-section, each observation is a different worker, so we take the observations to be drawn **independently and identically distributed**:

(Y_i, X_i) are iid, with joint distribution $F_{Y, X_1, \dots, X_k}$.

- Note that independence is between (Y_i, X_i) and (Y_j, X_j) for $i \neq j$, not between Y_i and X_i.
- Under the iid assumption, we can take the population to be the joint distribution of one observation (Y_i, X_i).

Modelling the randomness

- A random experiment has a sample space Ω of outcomes ω. These are possible states of the world.
- Let X denote education and V ability.
- The pair (X, V) is a function from Ω to $\mathbb{R}^2$: for the state of the world ω, $X(\omega)$ is the education and $V(\omega)$ is the ability.



- A probability P assigns to each event A a number $\mathrm{P}(A) \in [0, 1]$. It induces a distribution for (X, V).
- The joint CDF of (X, V) is a statement about sets of outcomes: for all $x, v \in \mathbb{R}$,

$$\begin{aligned} F_{X,V}(x, v) &= \mathrm{P}(X \le x,\, V \le v) \\ &= \mathrm{P}(\{\omega \in \Omega : X(\omega) \le x,\ V(\omega) \le v\}). \end{aligned}$$

Probability

Random variables and the CDF

- A random variable is a map $Y : \Omega \to \mathbb{R}$. For a set $A \subset \mathbb{R}$,

$$\mathrm{P}(Y \in A) = \mathrm{P}(\{\omega : Y(\omega) \in A\}).$$

- Taking $A = (-\infty, y]$ gives the **cumulative distribution function**:

$$F_Y(y) = \mathrm{P}(Y \le y), \qquad y \in \mathbb{R}.$$

- Properties:
  - F_Y is nondecreasing and right-continuous, $F_Y(y) \to 0$ as $y \to -\infty$ and $F_Y(y) \to 1$ as $y \to \infty$;
  - $\mathrm{P}(a < Y \le b) = F_Y(b) - F_Y(a)$;
  - a jump of F_Y at y_1 has size $\mathrm{P}(Y = y_1)$; where F_Y is continuous, $\mathrm{P}(Y = y_1) = 0$.

Continuous random variables and the PDF

- When F_Y is differentiable, the **probability density function** is

$$f_Y(y) = \frac{dF_Y(y)}{dy}, \qquad \text{so} \qquad F_Y(y) = \int_{-\infty}^{y} f_Y(v)\, dv.$$

- $f_Y(y) \ge 0$, $\displaystyle\int_{-\infty}^{\infty} f_Y(y)\, dy = 1$, and for $A \subset \mathbb{R}$,

$$\mathrm{P}(Y \in A) = \int_A f_Y(y)\, dy.$$

- A density is not a probability: $\mathrm{P}(Y = y) = 0$ for every y, and $f_Y(y)$ can exceed one.
- Everything in this course is written for the continuous case. The discrete case replaces integrals with sums.

Expectation

- For a function h,

$$\mathrm{E}\left[h(Y)\right] = \int_{-\infty}^{\infty} h(y)\, f_Y(y)\, dy.$$

- Throughout, every expectation we write down is assumed to exist and to be finite.
- $\mathrm{E}\left[Y\right]$ is the first moment, $\mathrm{E}\left[Y^2\right]$ the second, and

$$\mathrm{Var}\left(Y\right) = \mathrm{E}\left[(Y - \mathrm{E}\left[Y\right])^2\right] = \mathrm{E}\left[Y^2\right] - (\mathrm{E}\left[Y\right])^2.$$

- The mean is the best constant predictor of Y under squared loss:

$$\mathrm{E}\left[(Y - c)^2\right] = \mathrm{E}\left[(Y - \mathrm{E}\left[Y\right])^2\right] + (\mathrm{E}\left[Y\right] - c)^2 \geq \mathrm{Var}\left(Y\right),$$

  with equality only at $c = \mathrm{E}\left[Y\right]$. (The cross term $2(\mathrm{E}\left[Y\right] - c)\,\mathrm{E}\left[Y - \mathrm{E}\left[Y\right]\right]$ is zero.)
- Under absolute loss the minimiser of $\mathrm{E}\left[|Y - c|\right]$ is the median instead. Which loss we use decides which feature of the distribution we end up estimating.

Joint distributions

- For two random variables Y and X on the same Ω:

$$F_{Y,X}(y, x) = \mathrm{P}(Y \leq y,\ X \leq x),$$
$$f_{Y,X}(y, x) = \frac{\partial^2 F_{Y,X}(y, x)}{\partial y\, \partial x}.$$

- $f_{Y,X} \geq 0$, $\displaystyle\iint f_{Y,X}(y, x)\, dy\, dx = 1$, and for $A \subset \mathbb{R}^2$,

$$\mathrm{P}((Y, X) \in A) = \iint_A f_{Y,X}(y, x)\, dy\, dx.$$

- With k regressors the joint CDF of one observation is

$$F_{Y,X_1,\ldots,X_k}(y, x_1, \ldots, x_k) = \mathrm{P}(Y \leq y,\ X_1 \leq x_1,\ \ldots,\ X_k \leq x_k).$$

From the joint density to a marginal

- **Claim.** The marginal density of X is obtained by integrating Y out:

$$f_X(x) = \int_{-\infty}^{\infty} f_{Y,X}(y, x)\, dy.$$

- **Proof.** Start from the marginal CDF and differentiate:

$$\begin{aligned} F_X(x) &= \mathrm{P}(X \leq x) = \mathrm{P}(X \leq x,\ Y < \infty) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{Y,X}(y, u)\, dy\, du, \\ f_X(x) &= \frac{dF_X(x)}{dx} = \frac{d}{dx}\left[\int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{Y,X}(y, u)\, dy\, du\right] \\ &= \int_{-\infty}^{\infty} f_{Y,X}(y, x)\, dy, \end{aligned}$$

  using $\frac{d}{dx}\int_{-\infty}^{x} h(u)\, du = h(x)$.

- The converse fails: knowing f_X and f_Y does **not** determine $f_{Y,X}$. The marginals say nothing about how Y and X move together.

Conditional density

- The density of Y **given** $X = x$, defined where $f_X(x) > 0$:

$$f_{Y|X}(y \mid x) = \frac{f_{Y,X}(y,x)}{f_X(x)}, \qquad y \in \mathbb{R}.$$

- This has the same shape as conditional probability for events, $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$: the joint, rescaled by the marginal density of what we condition on.
- It is a proper density in y:

$$\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\, dy = \frac{1}{f_X(x)} \underbrace{\int_{-\infty}^{\infty} f_{Y,X}(y,x)\, dy}_{= f_X(x)} = 1.$$

Independence

- Y and X are **independent** if conditioning changes nothing:

$$f_{Y|X}(y \mid x) = f_Y(y) \quad \text{for all } y \text{ and all } x \text{ with } f_X(x) > 0.$$

- Equivalently the joint factorises:

$$\begin{aligned} & f_{Y,X}(y,x) = f_Y(y)\, f_X(x) \quad \text{for all } x, y, \\ \text{or equivalently} \qquad & F_{Y,X}(y,x) = F_Y(y)\, F_X(x) \quad \text{for all } x, y. \end{aligned}$$

- Independence is a strong statement: it involves the whole conditional distribution. Regression will need much less.

Expectation of a function of two variables

- For a function h of both arguments,

$$\mathrm{E}\left[h(Y,X)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y,x)\, f_{Y,X}(y,x)\, dy\, dx.$$

- With $h(y,x) = yx$ this is the joint moment $\mathrm{E}\left[YX\right]$. Centring both arguments gives the **covariance**:

$$\begin{aligned} \mathrm{Cov}\left(Y,X\right) &= \mathrm{E}\left[\left(Y - \mathrm{E}\left[Y\right]\right)\left(X - \mathrm{E}\left[X\right]\right)\right] \\ &= \mathrm{E}\left[YX\right] - \mathrm{E}\left[Y\right]\, \mathrm{E}\left[X\right]. \end{aligned}$$

- When $\mathrm{Var}\left(Y\right)$ and $\mathrm{Var}\left(X\right)$ are positive and finite, dividing by the two standard deviations removes the units and gives the correlation,

$$\mathrm{Corr}(Y,X) = \frac{\mathrm{Cov}\left(Y,X\right)}{\sqrt{\mathrm{Var}\left(Y\right)\, \mathrm{Var}\left(X\right)}} \in [-1,1],$$

  where the bound is the Cauchy-Schwarz inequality.
- Y and X are **uncorrelated** if $\mathrm{Cov}\left(Y,X\right) = 0$, equivalently $\mathrm{E}\left[YX\right] = \mathrm{E}\left[Y\right]\, \mathrm{E}\left[X\right]$.
- Uncorrelatedness is a single restriction. It says the best straight-line predictor of Y from X under squared loss has zero slope, and it says nothing about any other way the two variables might be related.

Some implications

- Under independence the expectation of a product of separate functions factorises:

$$\begin{aligned}\mathrm{E}\left[h_1(Y)\,h_2(X)\right] &= \iint h_1(y)\,h_2(x)\;\underbrace{f_{Y,X}(y,x)}_{=f_Y(y)\,f_X(x)}\,dy\,dx \\ &= \int h_1(y)\,f_Y(y)\,dy \cdot \int h_2(x)\,f_X(x)\,dx \\ &= \mathrm{E}\left[h_1(Y)\right]\,\mathrm{E}\left[h_2(X)\right].\end{aligned}$$

- This holds for **every** pair of functions h_1 and h_2, not just for the identity.
- Taking $h_1(y) = y$ and $h_2(x) = x$ gives $\mathrm{E}[YX] = \mathrm{E}[Y]\,\mathrm{E}[X]$, so $\mathrm{Cov}(Y, X) = 0$: independence implies uncorrelatedness.
- The converse fails, and it fails for the reason above: uncorrelatedness is the factorisation for one pair of functions, whereas independence is the factorisation for all of them.

Conditional expectation

Conditional expectation: at a point

- Fix x with $f_X(x) > 0$ and take the mean of the conditional density:

$$\mathrm{E}\left[Y \mid X = x\right] = \int_{-\infty}^{\infty} y\,f_{Y|X}(y \mid x)\,dy.$$

  This is a **number**.
- Repeat for every x with $f_X(x) > 0$. The result is a function of x,

$$x \longmapsto \mathrm{E}\left[Y \mid X = x\right] =: g(x),$$

  the **conditional expectation function**.

Conditional expectation: as a random variable

- Now plug the random variable X into g:

$$\mathrm{E}\left[Y \mid X\right] := g(X) = \int_{-\infty}^{\infty} y\,f_{Y|X}(y \mid X)\,dy.$$

  This is a **random variable**: a different draw of X gives a different value.
- The distinction matters for every proof that follows. $\mathrm{E}[Y \mid X = x]$ is a number indexed by x; $\mathrm{E}[Y \mid X]$ is the random variable obtained by evaluating that function at X.
- The same construction applied to the squared deviation gives the **conditional variance**,

$$\mathrm{Var}\left(Y \mid X\right) = \mathrm{E}\left[\left(Y - \mathrm{E}\left[Y \mid X\right]\right)^2 \,\middle|\, X\right],$$

  again a random variable, and a function of X.
- Since $\mathrm{E}[Y \mid X]$ is random, it has a mean of its own. What is it?

Law of iterated expectations

- **Claim (LIE).**

$$\mathrm{E}\left[\mathrm{E}\left[Y \mid X\right]\right] = \mathrm{E}\left[Y\right].$$

- Averaging the conditional mean over the distribution of X gives back the unconditional mean.
- Read from right to left, it says any expectation can be computed in two stages: first conditional on X, then over X. Most proofs in this course use it that way.

Law of iterated expectations: proof

- Write the outer expectation as an integral against f_X, then unpack g:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[Y \mid X\right]\right] &= \int_{-\infty}^{\infty} \underbrace{\mathrm{E}\left[Y \mid X = x\right]}_{g(x)} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy\right] f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, \underbrace{f_{Y|X}(y \mid x)\, f_X(x)}_{= f_{Y,X}(y,x)}\, dy\, dx \\
&= \int_{-\infty}^{\infty} y\, \underbrace{\left[\int_{-\infty}^{\infty} f_{Y,X}(y,x)\, dx\right]}_{= f_Y(y)}\, dy \\
&= \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \mathrm{E}\left[Y\right].
\end{aligned}$$

- The step from the third line to the fourth changes the order of integration.

Taking out what is known

- Conditional on X, any function of X is a constant and comes out of the expectation:

$$\begin{aligned}
\mathrm{E}\left[Y \cdot h(X) \mid X\right] &= \int_{-\infty}^{\infty} y\, h(X)\, f_{Y|X}(y \mid X)\, dy \\
&= h(X) \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid X)\, dy = h(X)\, \mathrm{E}\left[Y \mid X\right].
\end{aligned}$$

- Two special cases used constantly:

$$\mathrm{E}\left[h(X) \mid X\right] = h(X), \qquad \mathrm{E}\left[c \mid X\right] = c.$$

Question

- Let Y, X, Z be random variables. Which of the following equals $\mathrm{E}\left[\mathrm{E}\left[Y \mid X, Z\right] \mid Z\right]$?
  1. $\mathrm{E}\left[Y\right]$
  2. $\mathrm{E}\left[Y \mid X\right]$
  3. $\mathrm{E}\left[Y \mid Z\right]$
  4. 0

Mean independence

- Y is **mean independent** of X if the conditional mean does not move with the conditioning value:

$$\mathrm{E}\left[Y \mid X = x\right] = c \quad \text{for all } x \text{ with } f_X(x) > 0,$$

  where the constant c does not depend on x.

- **Question.** The constant c does not depend on x. What must it be?

  1. 0
  2. $\mathrm{E}\left[X\right]$
  3. $\mathrm{E}\left[Y\right]$
  4. 42

- Mean independence is weaker than independence:
  - Under mean independence, $\mathrm{E}\left[Y \mid X\right]$ does not depend on X, but $\mathrm{E}\left[Y^2 \mid X\right]$, $\mathrm{Var}\left(Y \mid X\right)$, etc., can change with X.
  - Under independence, the whole conditional distribution is constant in X, so $\mathrm{Var}\left(Y \mid X\right)$, etc., are constant too.

A zero conditional mean

- Suppose

$$\mathrm{E}\left[U \mid X\right] = 0.$$

  By the previous slide this is mean independence of U from X, in the case where the constant is zero.

- Two consequences, by the LIE and by taking out what is known:

$$\mathrm{E}\left[U\right] = \mathrm{E}\left[\underbrace{\mathrm{E}\left[U \mid X\right]}_{=0}\right] = 0,$$

$$\mathrm{E}\left[UX\right] = \mathrm{E}\left[\mathrm{E}\left[UX \mid X\right]\right] = \mathrm{E}\left[X\underbrace{\mathrm{E}\left[U \mid X\right]}_{=0}\right] = 0.$$

  The second line takes X out of the inner expectation, which is allowed because X is known once we condition on it.

- Putting the two together,

$$\mathrm{Cov}\left(U, X\right) = \mathrm{E}\left[UX\right] - \mathrm{E}\left[U\right]\mathrm{E}\left[X\right] = 0 - 0 = 0,$$

  so U and X are uncorrelated.

Three concepts of unrelatedness

- Ordered from strongest to weakest:

$$\text{independence} \implies \mathrm{E}\left[U \mid X\right] = \mathrm{E}\left[U\right] \implies \mathrm{Cov}\left(U, X\right) = 0,$$

  and neither arrow reverses.

- Mean independence allows $\mathrm{Var}\left(U \mid X\right)$ to depend on X; independence does not. A variable can have a conditional mean that never moves with X while its conditional variance does.

- Uncorrelatedness is weaker still: $\mathrm{Cov}\left(U, X\right) = 0$ is a single restriction, and it is compatible with a conditional mean $\mathrm{E}\left[U \mid X\right]$ that varies with X, because $\mathrm{Cov}\left(U, X\right) = \mathrm{Cov}\left(\mathrm{E}\left[U \mid X\right], X\right)$.

The conditional mean minimises mean squared error

- Among all functions m of X, the conditional mean is the best predictor of Y under squared loss. Condition on X first:

$$\begin{aligned}\mathrm{E}\left[(Y - m(X))^2 \mid X\right] = \mathrm{E}\left[\left(Y - \mathrm{E}\left[Y \mid X\right]\right)^2 \mid X\right] \\ + \left(\mathrm{E}\left[Y \mid X\right] - m(X)\right)^2,\end{aligned}$$

  because the cross term is $2\left(\mathrm{E}\left[Y \mid X\right] - m(X)\right)\mathrm{E}\left[Y - \mathrm{E}\left[Y \mid X\right] \mid X\right] = 0$.

- Now average over X with the LIE:

$$\mathrm{E}\left[(Y - m(X))^2\right] \geq \mathrm{E}\left[\left(Y - \mathrm{E}\left[Y \mid X\right]\right)^2\right],$$

  with equality only when $m(X) = \mathrm{E}\left[Y \mid X\right]$ with probability one.

- This is the reason regression is about $\mathrm{E}\left[Y \mid X\right]$ and not some other feature of the conditional distribution. Change the loss and a different feature wins.

The linear regression model

The object of interest

- The relationship between the regressors and the dependent variable is summarised by the conditional expectation function:

$$\mathrm{E}\left[Y_i \mid X_i\right] = \mathrm{E}\left[Y_i \mid X_{i1}, \ldots, X_{ik}\right],$$

for instance

$$\mathrm{E}\left[\ln \mathrm{Wage}_i \mid \mathrm{Education}_i, \mathrm{Experience}_i, \mathrm{Race}_i, \mathrm{Gender}_i, \ldots\right].$$

- The practical problem: with data $\{(Y_i, X_i) : i = 1, \ldots, n\}$, estimate $\mathrm{E}\left[Y_i \mid X_i\right]$.
- Without restrictions this is a function of k arguments to be learned from n points. The **parametric** approach assumes the functional form and leaves a finite number of unknown constants.

The linear regression model

- **Assumption.** The conditional expectation is linear in the parameters:

$$\mathrm{E}\left[Y_i \mid X_i\right] = \beta_1 X_{i1} + \cdots + \beta_k X_{ik} = X_i^\top \beta,$$

where $X_i = (X_{i1}, \ldots, X_{ik})^\top$ and $\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$.

- Write $\mathrm{E}\left[Y_i \mid X_i = x\right] = x^\top \beta = \sum_{l=1}^k \beta_l x_l$. Differentiating in x_j,

$$\frac{\partial \mathrm{E}\left[Y_i \mid X_i = x\right]}{\partial x_j} = \frac{\partial (\beta_1 x_1 + \cdots + \beta_k x_k)}{\partial x_j} = \beta_j.$$

- So β_j is the change in the conditional mean of Y_i per unit change in X_{ij}, holding the other regressors fixed: a **marginal effect**.
- "Linear" refers to β. The regressors can be anything: squares, logs, products, indicator variables.

The error term

- **Define** the error as the deviation from the conditional mean:

$$U_i := Y_i - \mathrm{E}\left[Y_i \mid X_i\right] = Y_i - X_i^\top \beta.$$

- Then, by construction,

$$\begin{aligned} Y_i &= X_i^\top \beta + U_i, \\ \mathrm{E}\left[U_i \mid X_i\right] &= \underbrace{\mathrm{E}\left[Y_i \mid X_i\right]}_{= X_i^\top \beta} - \underbrace{\mathrm{E}\left[X_i^\top \beta \mid X_i\right]}_{= X_i^\top \beta} = 0. \end{aligned}$$

- U_i is not observable: the conditional expectation function is unknown, so the deviation from it is unknown.
- If the pairs (Y_i, X_i) are iid, then so are the U_i, since each is the same function of its own pair.

Two ways to write the same model

- Either start from the conditional mean and define the error,

$$\left.\begin{aligned} \mathrm{E}\left[Y_i \mid X_i\right] &= X_i^\top \beta \\ U_i &:= Y_i - X_i^\top \beta \end{aligned}\right\} \quad \Longrightarrow \quad \mathrm{E}\left[U_i \mid X_i\right] = 0,$$

- or start from the equation and the restriction on the error,

$$\left.\begin{aligned} Y_i &= X_i^\top \beta + U_i \\ \mathrm{E}\left[U_i \mid X_i\right] &= 0 \end{aligned}\right\} \quad \Longrightarrow \quad \mathrm{E}\left[Y_i \mid X_i\right] = X_i^\top \beta.$$

The intercept

- Consider an example with a single regressor $X_i^* \in \mathbb{R}$.
- Suppose $Y_i = \beta^* X_i^* + U_i^*$ with $\mathrm{E}[U_i^* \mid X_i^*] = \alpha \neq 0$.
- We can add and subtract α to get:

$$Y_i = \alpha + \beta^* X_i^* + \underbrace{(U_i^* - \alpha)}_{U_i}, \qquad \mathrm{E}[U_i \mid X_i^*] = 0.$$

- Define vectors $X_i = (1, X_i^*)^\top$ and $\beta = (\alpha, \beta^*)^\top$. Then

$$Y_i = X_i^\top \beta + U_i, \qquad \mathrm{E}[U_i \mid X_i] = \mathrm{E}[U_i \mid 1, X_i^*] = 0,$$

  as conditioning on 1 in addition to X_i^* does not change the conditional expectation.
- Here α is the intercept: a coefficient on a constant regressor (i.e., a regressor equal to 1 for all observations i).
- We include the intercept to absorb constant but non-zero conditional means of the error. This does not change the slope coefficients.
- However, if the conditional mean of the error is not constant, i.e., $\mathrm{E}[U_i^* \mid X_i^*] = \alpha + \gamma X_i^*$, then

$$Y_i = \alpha + (\beta^* + \gamma) X_i^* + \underbrace{(U_i^* - \alpha - \gamma X_i^*)}_{U_i}, \qquad \mathrm{E}[U_i \mid X_i^*] = 0,$$

  and the slope coefficient is affected.

Identification

Identification: the question

- Consider the model $Y_i = X_i^\top \beta + U_i$ with $\mathrm{E}[U_i \mid X_i] = 0$:
  - Y_i, X_i are observed,
  - U_i is unobserved,
  - $\beta \in \mathbb{R}^k$ is unknown.
- Suppose the joint distribution of the observables $F_{Y, X_1, \ldots, X_k}(y, x_1, \ldots, x_k)$ is **known** (we are at the population level). You have no information about U_i beyond $\mathrm{E}[U_i \mid X_i] = 0$.
- Can you recover β?
- If the answer is no with the whole distribution in hand, no amount of data will help, as the joint distribution of the data does not pin down β.

Identification: the definition

- **Definition.** β is **identified** if it is uniquely determined by the joint distribution of the observed variables $(Y_i, X_{i1}, \ldots, X_{ik})$.
- Two things can go wrong:
  - the distribution of the observables is consistent with more than one value of β;
  - it is consistent with none, because the model is wrong.
- We ask: what does $\mathrm{E}[U_i \mid X_i] = 0$ let us compute from the distribution of (Y_i, X_i), and does it pin β down?

What the model gives us

- By $\mathrm{E}[U_i \mid X_i] = 0$ and the LIE, $\mathrm{E}[X_i U_i] = 0$.
- Rewrite U_i in terms of the observables and β:

$$U_i = Y_i - X_i^\top \beta$$

- Obtain the **population normal equations**:

$$\boxed{\mathrm{E}\left[X_i(Y_i - X_i^\top \beta)\right] = 0}.$$

- Multiply out, one step at a time:

$$\begin{aligned} 0 = \mathrm{E}\left[X_i U_i\right] &= \mathrm{E}\left[X_i(Y_i - X_i^\top \beta)\right] \\ &= \mathrm{E}\left[X_i Y_i - X_i X_i^\top \beta\right] \\ &= \mathrm{E}\left[X_i Y_i\right] - \mathrm{E}\left[X_i X_i^\top \beta\right] \\ &= \mathrm{E}\left[X_i Y_i\right] - \mathrm{E}\left[X_i X_i^\top\right]\ \beta, \end{aligned}$$

  the last step because β is a vector of constants and comes out of the expectation.

- Rearranging,

$$\mathrm{E}\left[X_i X_i^\top\right]\ \beta = \mathrm{E}\left[X_i Y_i\right].$$

- Knowns: $\mathrm{E}\left[X_i Y_i\right]$ $(k \times 1)$ and $\mathrm{E}\left[X_i X_i^\top\right]$ $(k \times k)$, both features of the distribution of the observables. Unknown: β.
- If $\mathrm{E}\left[X_i X_i^\top\right]$ is invertible, the system has exactly one solution. Whether it is invertible is the whole question.

The second-moment matrix

- Written out,

$$\mathrm{E}\left[X_i X_i^\top\right] = \mathrm{E}\begin{bmatrix} X_{i1}^2 & X_{i1}X_{i2} & \cdots & X_{i1}X_{ik} \\ X_{i2}X_{i1} & X_{i2}^2 & \cdots & X_{i2}X_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ X_{ik}X_{i1} & X_{ik}X_{i2} & \cdots & X_{ik}^2 \end{bmatrix}.$$

- It is **symmetric**: using $(AB)^\top = B^\top A^\top$,

$$\left(\mathrm{E}\left[X_i X_i^\top\right]\right)^\top = \mathrm{E}\left[(X_i X_i^\top)^\top\right] = \mathrm{E}\left[X_i X_i^\top\right].$$

The second-moment matrix is positive semidefinite

- **Definitions.** A $k \times k$ matrix A is **positive semidefinite** (p.s.d.) if $c^\top A c \geq 0$ for all $c \in \mathbb{R}^k$, and **positive definite** (p.d.) if $c^\top A c > 0$ for all $c \neq 0$.
- Take any $c \in \mathbb{R}^k$ and define the scalar

$$W_i = X_i^\top c = c_1 X_{i1} + \cdots + c_k X_{ik}, \qquad c^\top X_i = (X_i^\top c)^\top = W_i.$$

- Then the quadratic form is the expectation of a square:

$$c^\top \mathrm{E}\left[X_i X_i^\top\right]\ c = \mathrm{E}\left[c^\top X_i X_i^\top c\right] = \mathrm{E}\left[\underbrace{W_i^2}_{\geq 0}\right] \geq 0.$$

- So $\mathrm{E}\left[X_i X_i^\top\right]$ is p.s.d. whatever the distribution of the regressors. Positive **definiteness** is the extra step, and it can fail.

When positive definiteness fails

- Suppose some $\tilde{c} \neq 0$ gives $\tilde{c}^\top \mathrm{E}\left[X_i X_i^\top\right]\ \tilde{c} = 0$. With $\tilde{W}_i = X_i^\top \tilde{c}$,

$$\begin{aligned} 0 = \tilde{c}^\top \mathrm{E}\left[X_i X_i^\top\right]\ \tilde{c} = \mathrm{E}\left[\tilde{W}_i^2\right] &\iff \tilde{W}_i = 0 \text{ with probability 1} \\ &\iff \boxed{\tilde{c}_1 X_{i1} + \cdots + \tilde{c}_k X_{ik} = 0} \end{aligned}$$

- With $\tilde{c} \neq 0$, say $\tilde{c}_1 \neq 0$, one regressor equals, with probability one, an exact linear combination of the others:

$$X_{i1} = -\frac{\tilde{c}_2}{\tilde{c}_1} X_{i2} - \cdots - \frac{\tilde{c}_k}{\tilde{c}_1} X_{ik}.$$

  This is **perfect multicollinearity**.

Positive definiteness and the rank condition

- Putting the two directions together,

$$\begin{aligned} \mathrm{E}\left[X_i X_i^\top\right] \text{ is p.d. } &\iff X_i^\top c = 0 \text{ with probability 1 only if } c = 0 \\ &\iff \operatorname{rank}\left(\mathrm{E}\left[X_i X_i^\top\right]\right) = k. \end{aligned}$$

- The relation has to be exact and linear. A relation $c_1 X_{i1} + c_2 X_{i2} = 0$ holding with probability one for some $(c_1, c_2) \neq (0, 0)$ is ruled out; $X_{i2} = X_{i1}^2$ is not linear and is fine, unless X_{i1} is binary, when $X_{i1}^2 = X_{i1}$.

Identification: the result

- **Assume**
  1. $Y_i = X_i^\top \beta + U_i$ with $\beta \in \mathbb{R}^k$;
  2. $\mathrm{E}\left[U_i \mid X_i\right] = 0$;
  3. No perfect multicollinearity: $\operatorname{rank}\left(\mathrm{E}\left[X_i X_i^\top\right]\right) = k$.
- Then β is identified, and

$$\boxed{\beta = \left(\mathrm{E}\left[X_i X_i^\top\right]\right)^{-1} \mathrm{E}\left[X_i Y_i\right]}$$

- Assumptions 1 and 2 give the population normal equations $\mathrm{E}\left[X_i(Y_i - X_i^\top \beta)\right] = 0$, that is $\mathrm{E}\left[X_i X_i^\top\right] \beta = \mathrm{E}\left[X_i Y_i\right]$; assumption 3 makes the matrix invertible.
- With a single regressor and no intercept ($k = 1$),

$$\mathrm{E}\left[X_i Y_i\right] = \mathrm{E}\left[X_i^2\right] \beta \quad \Longrightarrow \quad \beta = \frac{\mathrm{E}\left[X_i Y_i\right]}{\mathrm{E}\left[X_i^2\right]}.$$

Identification under a weaker assumption

- Replace $\mathrm{E}\left[U_i \mid X_i\right] = 0$ with a weaker condition:

$$\mathrm{E}\left[X_i U_i\right] = 0.$$

- Assume $Y_i = X_i^\top \beta + U_i$ and $\operatorname{rank}\left(\mathrm{E}\left[X_i X_i^\top\right]\right) = k$.
- Then we can proceed as before: $\beta = (\mathrm{E}\left[X_i X_i^\top\right])^{-1} \mathrm{E}\left[X_i Y_i\right]$, so β is still identified.
- What changes is the meaning of β. Without $\mathrm{E}\left[U_i \mid X_i\right] = 0$ we can no longer conclude that $X_i^\top \beta$ is the conditional mean, since in general

$$\mathrm{E}\left[Y_i \mid X_i\right] = X_i^\top \beta + \mathrm{E}\left[U_i \mid X_i\right].$$

Estimation and least squares

From identification to an estimator

- Identification wrote β as a function of population moments:

$$\beta = \left(\mathrm{E}\left[X_i X_i^\top\right]\right)^{-1} \mathrm{E}\left[X_i Y_i\right].$$

- Everything below assumes the sample counterpart of the rank condition, $\operatorname{rank}(X) = k$, so that the inverse exists. We return to it once the matrix notation is in place.
- **Method of moments.** Replace each expectation $\mathrm{E}[\cdot]$ by the sample average $\frac{1}{n}\sum_{i=1}^{n}(\cdot)$:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i = \left(\sum_{i=1}^{n} X_i X_i^\top\right)^{-1} \sum_{i=1}^{n} X_i Y_i.$$

  The two factors of $\frac{1}{n}$ cancel.
- An **estimator** is any function of the sample $\{(Y_i, X_i)\}$. It may not depend on U_i or on β, which are unobserved.

The same estimator from the moment condition

- Alternatively, start from $\mathrm{E}[X_i(Y_i - X_i^\top\beta)] = 0$ and replace the expectation by an average, evaluated at the estimator:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(\underbrace{Y_i - X_i^\top\hat{\beta}}_{=:\,\hat{U}_i}) = 0 \quad\Longrightarrow\quad \sum_{i=1}^{n} X_i X_i^\top\hat{\beta} = \sum_{i=1}^{n} X_i Y_i,$$

  which is the same $\hat{\beta}$.
- $\hat{U}_i = Y_i - X_i^\top\hat{\beta}$ is the **residual**: the sample counterpart of the error $U_i = Y_i - X_i^\top\beta$. Unlike U_i, it is computable.
- $\hat{\beta}$ is built to make the residuals satisfy the sample version of $\mathrm{E}[X_i U_i] = 0$.

The estimator is not the parameter

- Sample averages are random; expectations are fixed numbers. In general

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \neq \mathrm{E}[X_i X_i^\top], \qquad \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \neq \mathrm{E}[X_i Y_i],$$

  and so in general $\hat{\beta} \neq \beta$.
- Flip a fair coin ten times: the sample proportion of heads is rarely one half, although the population proportion is exactly one half.
- If the errors $U_1, \ldots, U_n$ are continuously distributed, $\mathrm{P}(\hat{\beta} = \beta) = 0$.
- $\hat{\beta}$ is a random vector: a different sample gives a different value. It has a distribution, and the next lecture is about where that distribution sits relative to β and how spread out it is.

Sample orthogonality

- Since $\hat{\beta}$ solves $\sum_i X_i(Y_i - X_i^\top\hat{\beta}) = 0$,

$$\sum_{i=1}^{n} X_i\hat{U}_i = 0 \qquad\Longleftrightarrow\qquad \begin{pmatrix}\sum_{i=1}^{n} X_{i1}\hat{U}_i\\ \vdots \\ \sum_{i=1}^{n} X_{ik}\hat{U}_i\end{pmatrix} = \begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}.$$

- One equation per regressor: each regressor is orthogonal to the residual vector,

$$\begin{pmatrix}X_{1j}\\ \vdots\\ X_{nj}\end{pmatrix}^\top \begin{pmatrix}\hat{U}_1\\ \vdots\\ \hat{U}_n\end{pmatrix} = \sum_{i=1}^{n} X_{ij}\,\hat{U}_i = 0, \qquad j = 1, \ldots, k.$$

- If the first regressor is the constant, $X_{i1} = 1$, the first equation says $\sum_{i=1}^{n}\hat{U}_i = 0$: the residuals average to zero.

Matrix notation

- Stack the n observations:

$$\underset{n\times 1}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \underset{n\times 1}{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}, \qquad \underset{n\times k}{X} = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}.$$

- Row i of X is the transpose of the regressor vector X_i; column j of X holds regressor j for all n observations:

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}.$$

- The n equations $Y_i = X_i^\top \beta + U_i$ become one:

$$Y = \begin{pmatrix} X_1^\top \beta \\ \vdots \\ X_n^\top \beta \end{pmatrix} + \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} \qquad \Longleftrightarrow \qquad \underset{n\times 1}{Y} = \underset{n\times k}{X}\ \underset{k\times 1}{\beta} + \underset{n\times 1}{U}.$$

Two identities

- **Claim.**

$$X^\top Y = \sum_{i=1}^{n} X_i Y_i \quad (k\times 1), \qquad X^\top X = \sum_{i=1}^{n} X_i X_i^\top \quad (k\times k).$$

- **Proof of the second.** Transposing the stack of rows gives a row of columns:

$$\begin{aligned} X^\top X &= \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}^\top \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix} \\ &= \underbrace{X_1 X_1^\top}_{k\times k} + \cdots + X_n X_n^\top = \sum_{i=1}^{n} X_i X_i^\top. \end{aligned}$$

  The first identity is the same computation with the second factor X replaced by Y, whose ith row is the scalar Y_i.

The estimator in matrix form

- Substitute the two identities into the method-of-moments formula:

$$\begin{aligned} \hat{\beta} &= \left(\sum_{i=1}^{n} X_i X_i^\top \right)^{-1} \sum_{i=1}^{n} X_i Y_i \\ &= \left(\frac{1}{n} X^\top X \right)^{-1} \left(\frac{1}{n} X^\top Y \right). \end{aligned}$$

- The $\frac{1}{n}$ factors cancel:

$$\boxed{\hat{\beta} = (X^\top X)^{-1} X^\top Y}$$

- $X^\top X$ is $k\times k$ and $X^\top Y$ is $k\times 1$, so $\hat{\beta}$ is $k\times 1$, as it should be.

Residuals in matrix form

- Residual vector and its defining property:

$$\begin{aligned}\hat{U} &= Y - X\hat{\beta},\\ X^\top\hat{U} &= X^\top(Y - X\hat{\beta}) = X^\top Y - X^\top X\underbrace{(X^\top X)^{-1}X^\top Y}_{\hat{\beta}}\\ &= X^\top Y - X^\top Y = 0.\end{aligned}$$

- These are the **sample normal equations**, $X^\top(Y - X\hat{\beta}) = X^\top\hat{U} = 0$, the sample counterpart of $\mathrm{E}\left[X_i(Y_i - X_i^\top\beta)\right] = 0$. They are the same k equations as before:

$$X^\top\hat{U} = \sum_{i=1}^n X_i\hat{U}_i = \begin{pmatrix}\sum_{i=1}^n X_{i1}\hat{U}_i\\ \vdots\\ \sum_{i=1}^n X_{ik}\hat{U}_i\end{pmatrix}.$$

- Population: $\mathrm{E}\left[X_iU_i\right] = 0$. Sample: $\frac{1}{n}\sum_{i=1}^n X_i\hat{U}_i = 0$. The estimator is the value of the coefficient vector that makes the sample mimic the population.

The sample rank condition

- $(X^\top X)^{-1}$ exists exactly when the columns of X are linearly independent:

$$Xc = c_1\begin{pmatrix}X_{11}\\ \vdots\\ X_{n1}\end{pmatrix} + \cdots + c_k\begin{pmatrix}X_{1k}\\ \vdots\\ X_{nk}\end{pmatrix} = 0 \quad \text{only if} \quad c = 0,$$

  that is, $\mathrm{rank}(X) = k$. In particular $n \geq k$: at least as many observations as coefficients.
- If it fails, $\mathrm{rank}(X^\top X) = \mathrm{rank}(X) < k$, and the normal equations have infinitely many solutions.
- This is the sample counterpart of $\mathrm{rank}\left(\mathrm{E}\left[X_iX_i^\top\right]\right) = k$.

Least squares

- Take a candidate value $b \in \mathbb{R}^k$ and measure how badly the fitted values $X_i^\top b$ fit the data by the **sum of squared residuals**:

$$\begin{aligned}S(b) = \sum_{i=1}^n\left(Y_i - X_i^\top b\right)^2 &= \begin{pmatrix}Y_1 - X_1^\top b\\ \vdots\\ Y_n - X_n^\top b\end{pmatrix}^\top\begin{pmatrix}Y_1 - X_1^\top b\\ \vdots\\ Y_n - X_n^\top b\end{pmatrix}\\ &= (Y - Xb)^\top(Y - Xb) = \|Y - Xb\|^2,\end{aligned}$$

  where for $a \in \mathbb{R}^n$, $\|a\|^2 = a^\top a = \sum_{i=1}^n a_i^2$.
- The **ordinary least squares** estimator picks the b that fits best:

$$\hat{\beta}^{\mathrm{OLS}} = \operatorname*{arg\,min}_{b\in\mathbb{R}^k}\sum_{i=1}^n\left(Y_i - X_i^\top b\right)^2 = \operatorname*{arg\,min}_{b\in\mathbb{R}^k}\|Y - Xb\|^2.$$

- **Claim.** $\hat{\beta}^{\mathrm{OLS}} = \hat{\beta} = (X^\top X)^{-1}X^\top Y$: least squares and the method of moments deliver the same estimator.

Proof: completing the square

- Add and subtract $X\hat{\beta}$ inside the criterion, and use $(u+v)^\top(u+v) = u^\top u + v^\top v + 2u^\top v$:

$$\begin{aligned}(Y-Xb)^\top(Y-Xb) &= (\underbrace{Y-X\hat{\beta}}_{\hat{U}} + X(\hat{\beta}-b))^\top(\hat{U}+X(\hat{\beta}-b)) \\ &= \hat{U}^\top\hat{U} + (\hat{\beta}-b)^\top X^\top X(\hat{\beta}-b) \\ &\quad + 2\underbrace{\hat{U}^\top X}_{=(X^\top\hat{U})^\top=0}(\hat{\beta}-b) \\ &= \underbrace{\hat{U}^\top\hat{U}}_{\text{free of } b} + (\hat{\beta}-b)^\top X^\top X(\hat{\beta}-b).\end{aligned}$$

- The cross term vanishes by the normal equations $X^\top\hat{U} = 0$. That is the only place the definition of $\hat{\beta}$ enters.
- So minimising $S(b)$ over b is the same as minimising the quadratic form $(\hat{\beta}-b)^\top X^\top X(\hat{\beta}-b)$.

Proof: the quadratic form

- For any $a \in \mathbb{R}^k$, with $z = Xa \in \mathbb{R}^n$,

$$a^\top X^\top Xa = (Xa)^\top(Xa) = z^\top z = \sum_{i=1}^n z_i^2 \geq 0,$$

  so $X^\top X$ is p.s.d. It is also symmetric, since $(X^\top X)^\top = X^\top X$.
- Under the rank condition, $Xa = 0$ only if $a = 0$, so $\sum_i z_i^2 = 0$ only if $a = 0$: $X^\top X$ is **positive definite**.
- Apply this with $a = \hat{\beta} - b$:

$$S(b) = S(\hat{\beta}) + \underbrace{(\hat{\beta}-b)^\top X^\top X(\hat{\beta}-b)}_{\geq 0,\ \text{and } =0 \text{ iff } b=\hat{\beta}} \geq S(\hat{\beta}).$$

- Hence $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ is the unique minimiser of the sum of squared residuals: it **is** the OLS estimator.

The same result from the first-order condition

- Expand the criterion, $S(b) = Y^\top Y - 2\,b^\top X^\top Y + b^\top X^\top Xb$, and use $\partial(x^\top c)/\partial x = c$ together with $\partial(x^\top Ax)/\partial x = 2Ax$ for symmetric A:

$$\frac{\partial S(b)}{\partial b} = -2X^\top Y + 2X^\top Xb = 0 \quad\Longrightarrow\quad b = (X^\top X)^{-1}X^\top Y.$$

- The Hessian $\partial^2 S/\partial b\,\partial b^\top = 2X^\top X$ is positive definite under the rank condition, so this is a minimum.
- The first-order condition, rewritten as $X^\top(Y - Xb) = 0$, **is** the normal equations. Least squares and the method of moments are two routes to one estimator.

Population least squares

- The population has its own least-squares problem, and β solves it.
- **Claim.** Suppose $Y_i = X_i^\top\beta + U_i$ with $\mathrm{E}[U_i \mid X_i] = 0$, and $\operatorname{rank}(\mathrm{E}[X_iX_i^\top]) = k$. Then

$$\beta = \underset{b\in\mathbb{R}^k}{\arg\min}\ \mathrm{E}\left[(Y_i - X_i^\top b)^2\right],$$

  and the minimiser is unique.
- So OLS is the sample analogue of β in two ways: replace population moments with sample moments, or replace expected squared error with average squared error.

Population least squares: proof

- Condition on X_i, use that the conditional mean minimises conditional mean squared error, then apply the LIE:

$$\begin{aligned}\mathrm{E}\left[(Y_i - X_i^\top b)^2\right] &= \mathrm{E}\left[\,\mathrm{E}\left[(Y_i - X_i^\top b)^2 \mid X_i\right]\,\right] \\ &\geq \mathrm{E}\left[\,\mathrm{E}\left[(Y_i - \underbrace{\mathrm{E}\left[Y_i \mid X_i\right]}_{X_i^\top \beta})^2 \mid X_i\right]\,\right] \\ &= \mathrm{E}\left[(Y_i - X_i^\top \beta)^2\right].\end{aligned}$$

- Equality at some b^* requires $X_i^\top b^* = \mathrm{E}\left[Y_i \mid X_i\right] = X_i^\top \beta$ with probability one, so $X_i^\top (b^* - \beta) = 0$ with probability one, and with no multicollinearity $b^* = \beta$.

Other linear estimators, for contrast

- Nothing forces the weights in the normal equations to be equal across observations. For a known symmetric positive definite $n \times n$ matrix W,

$$\tilde{\beta} = (X^\top W X)^{-1} X^\top W Y$$

  is also a function of the sample, and $W = I_n$ gives back $\hat{\beta}$.

- In general $\tilde{\beta}$ minimises $(Y - Xb)^\top W (Y - Xb)$. With W diagonal, $W = \mathrm{diag}(w_1, \ldots, w_n)$ and $w_i > 0$, this is **weighted least squares**: it minimises $\sum_i w_i (Y_i - X_i^\top b)^2$. Taking W to be the inverse of the variance matrix of the errors gives **generalised least squares**.

- Which of these is a good estimator, and in what sense, is the subject of the next lecture: bias, variance, and the Gauss-Markov theorem.

Summary: the model and identification

- The object of interest is the conditional expectation $\mathrm{E}\left[Y_i \mid X_i\right]$; the linear model assumes $\mathrm{E}\left[Y_i \mid X_i\right] = X_i^\top \beta$, which is the same as $Y_i = X_i^\top \beta + U_i$ with $\mathrm{E}\left[U_i \mid X_i\right] = 0$.

- Probability supplied what was needed on the way: densities and conditional densities, the law of iterated expectations, taking out what is known, and mean independence.

- **Identification.** $\mathrm{E}\left[U_i \mid X_i\right] = 0$ gives $\mathrm{E}\left[X_i U_i\right] = 0$, hence the population normal equations $\mathrm{E}\left[X_i (Y_i - X_i^\top \beta)\right] = 0$, that is $\mathrm{E}\left[X_i X_i^\top\right] \beta = \mathrm{E}\left[X_i Y_i\right]$. If $\mathrm{E}\left[X_i X_i^\top\right]$ is positive definite, which means no exact linear relation among the regressors, then

$$\beta = \left(\mathrm{E}\left[X_i X_i^\top\right]\right)^{-1} \mathrm{E}\left[X_i Y_i\right].$$

Summary: the estimator

- **Estimation.** Under the sample rank condition $\mathrm{rank}(X) = k$, replacing expectations by sample averages gives

$$\hat{\beta} = \left(\sum_{i=1}^n X_i X_i^\top\right)^{-1} \sum_{i=1}^n X_i Y_i = (X^\top X)^{-1} X^\top Y,$$

  which is also the unique minimiser of $\sum_i (Y_i - X_i^\top b)^2$: the OLS estimator.

- $\hat{\beta}$ solves the sample normal equations $X^\top (Y - X\hat{\beta}) = 0$, that is $X^\top \hat{U} = 0$: the sample counterpart of the population normal equations $\mathrm{E}\left[X_i (Y_i - X_i^\top \beta)\right] = 0$.