Linear Regression

Yerlan Kuzbakov

Framing the question and economic model

- **Goal:** Define a clear question of interest and formulate an economic model.
- **Example (Mincer's Model):** $\text{wage} = f(\text{educ}, \text{expe})$.
- **Variables:**
  - wage: hourly wage
  - educ: years of education
  - expe: years of experience

Econometric model (Mincer's equation)

- **Specification:**

$$\text{wage}_i = \beta_1 + \beta_2 \cdot \text{educ}_i + \beta_3 \cdot \text{expe}_i + u_i$$

- **Stochastic term:** u_i captures unobserved factors with assumptions described later.

Cross-sectional data basics

- **Definition:** A sample of individuals, households, firms, cities, countries, etc., taken at a given point in time.
- **Use case:** Widely used in applied microeconomics.
- **Sampling:** Often obtained by random sampling from the population of interest.
- **Random sample:** $\{x_1, \ldots, x_N\}$ is random if the N observations are drawn independently from the same population (i.i.d.).

Example cross-sectional data (Wooldridge, USA 1976)

obsno	wage	educ	expe	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
			...		
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Causal effects and identification

- **Economist's goal:** Infer the causal effect of one variable on another, holding other relevant variables constant.
- **Example (agriculture):** Fertilizer's effect on crop yield; other factors (rainfall, land quality) matter, but randomized experiments can be run.
- **Example (returns to education):** Experience and ability affect earnings; true randomized experiments are typically infeasible.
- **Implication:** Observational data require careful modeling and assumptions for causal inference.

Simple linear regression model and data-generating process

- **Sample:** $\{(y_i, x_i) \mid i = 1, \ldots, N\}$.
- **Model:**

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad \text{with } \mathbb{E}(u_i \mid x_1, \ldots, x_N) = 0.$$

- **Interpretation:**
  - β_1: intercept, β_2: slope (effect of x on y).
  - y_i: dependent variable (regressand), x_i: independent variable (regressor).
  - u_i: error term (disturbance).

Implications of exogeneity assumptions

- **Unconditional mean:** $\mathbb{E}(u_i) = 0$.
- **Orthogonality:** $\mathbb{E}(x_i u_i) = 0$.
- **Uncorrelatedness:** $\operatorname{cov}(x_i, u_i) = 0$.
- **Conditional mean of y:**

$$\mathbb{E}(y_i \mid x_i) = \beta_1 + \beta_2 x_i,$$

which is the population regression function.

Expected causal effect

- **Marginal effect of x on y:**

$$\beta_2 = \frac{\partial \mathbb{E}(y \mid x)}{\partial x}.$$

Mean shift and inclusion of a constant

- **If** the regressors include a constant, we can absorb a nonzero mean in u_i:

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad \mathbb{E}(u_i \mid x_1, \ldots, x_N) = \mu.$$

- **Reparameterization:**

$$y_i = (\beta_1 + \mu) + \beta_2 x_i + (u_i - \mu)$$

so we can write $y_i = \beta_1 + \beta_2 x_i + u_i$ with $\mathbb{E}(u_i \mid x_1, \ldots, x_N) = 0$.

Analogy principle and moment conditions

- **Population moment conditions:**

$$\mathbb{E}(u_i) = 0, \qquad \mathbb{E}(x_i u_i) = 0.$$

- **Using** $u_i = y_i - \beta_1 - \beta_2 x_i$:

$$\mathbb{E}(y_i - \beta_1 - \beta_2 x_i) = 0, \quad \mathbb{E}\big[x_i(y_i - \beta_1 - \beta_2 x_i)\big] = 0.$$

- **Sample counterparts (normal equations):**

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \quad \frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0.$$

OLS estimators from normal equations

- **Means:** $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.
- **First normal equation:** $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$, hence

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$$

- **If** $\sum_{i=1}^{N}(x_i - \bar{x})^2 \neq 0$, then

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \qquad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$$

Fitted values, residuals, and sample regression function

- **Fitted values:** $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.
- **Residuals:** $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$.
- **Sample regression function:** $y = \hat{\beta}_1 + \hat{\beta}_2 x$.

OLS as least squares minimization

- **Residual for hypothetical** (b_1, b_2): $r_i = y_i - b_1 - b_2 x_i$.
- **Objective (SSR):**

$$\mathrm{SSR}(b_1, b_2) = \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2 .$$

- **OLS estimate:**

$$(\hat{\beta}_1, \hat{\beta}_2) = \arg \min_{(b_1, b_2)} \mathrm{SSR}(b_1, b_2).$$

- Squaring imposes heavier penalties on large residuals.

First-order conditions for OLS

- **FOC 1:** $\sum_{i=1}^{N} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \right) = 0.$
- **FOC 2:** $\sum_{i=1}^{N} x_i \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \right) = 0.$

Key properties of OLS

- **Zero-sum residuals:** $\sum_{i=1}^{N} \hat{u}_i = 0$ (by FOC 1).
- **Orthogonality in sample:** $\sum_{i=1}^{N} x_i \hat{u}_i = 0$ (by FOC 2).
- **Regression through the means:** $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$ (by FOC 1).

Summary

- Economic models motivate econometric specifications.
- Cross-sectional data and exogeneity yield population moments.
- OLS solves sample analogs of population conditions via least squares.
- Resulting estimates have intuitive properties and a clear causal interpretation under assumptions.