ECON21030 (S25): Econometrics - Honors

Lecturer: Joseph Hardwick
Notes by: Aden Chen

Thursday 27th March, 2025

Contents

1 Introduction

- The "small bin" problem, dimension reduction, and linearity.
- Given the model $y_i = \beta x_i + \epsilon_i$, ϵ_i is the **error**, and $\hat{\epsilon}_i = y_i - \hat{y}_i$ is the **residual**. The residual is sample-dependent.
-
  - $\min_b \sum |x_i - a - bx_i|$ gives an estimate of the conditional median of y given x. This is called the "quantile regression.'
  - $\min_b \sum |x_i - a - bx_i|^2$ gives the conditional expectation function $\mathrm{E}[Y|X]$. This is called the "ordinary least squares.'

2 Probability

Definition 2.1. A random variable X is **absolutely continuous** if there exists a density function f_X such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, \mathrm{d}t.$$

Remark 2.2. Absolutely continuous distributions assign probability 0 to any finite set of points.

2.1 Expectation

Proposition 2.3.

- E *is linear.*
- *If* $X \leq Y$ *with probability* 1*, then* $\mathrm{E}\, X \leq \mathrm{E}\, Y$.

Theorem 2.4 (Jensen's Inequality)**.** *If* X *is such that* $\mathrm{E}\, X$ *and* $\mathrm{E}\, g(X)$ *exist and* g *is convex, then*

$$g(\mathrm{E}\, X) \leq \mathrm{E}\, g(X)$$

where the inequality is strict if g *is strictly convex and* X *is not constant.*

Proof. From the convexity of g we know $g(x) \geq g(y) + g'(y)(x - y)$ for any x and y. Setting $y = \mu =: \mathrm{E}\, X$ gives

$$g(X) \geq g(\mu) + g'(\mu)(X - \mu), \quad \forall x, y.$$

Taking expectation on both sides gives the desired result. □

Example 2.5. Wages are often modeled using a log-normal distribution: $\log w \sim \mathcal{N}(\mu, \sigma^2)$. Then, $\mathrm{E} \log w = \mu$, but $\mathrm{E}\, w = \mathrm{E}(\exp \log w) \geq e^{\mu}$ (the inequality is strict when $\sigma^2 > 0$). It turns out that $\mathrm{E}\, w = \exp\left(\mu + \sigma^2/2\right)$.

Proposition 2.6 (Properties of Conditional Expectation)**.**

- *Linearity.*
- *(Law of iterated expectation)* $\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y|X))$.
- *(Taking out what is known)* $\mathrm{E}(f(X) + g(X)Y|X) = f(X) + g(X)\,\mathrm{E}(Y|X)$.

Proposition 2.7 (Law of total variance)**.**

$$\mathrm{Var}(Y) = \mathrm{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathrm{E}(Y|X)),$$

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

2.2 Mean Independence

Definition 2.8. Y is **mean independent** of X if $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

Remark 2.9. We have

$$\begin{aligned} X \text{ is independent of } Y &\implies X \text{ is mean independent of } Y \\ &\implies \mathrm{Cov}(X, Y) = 0, \end{aligned}$$

where the second implication follows from the law of iterated expectations:

$$\mathrm{Cov}(X, Y) = \mathrm{E}(X\,\mathrm{E}(Y|X)) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(X\,\mathrm{E}(Y)) - \mathrm{E}(X)\,\mathrm{E}(Y) = 0.$$

The converse of the last implication is not true in general, but true for jointly normal random variables.

2.3 Moments

Definition 2.10. If $\mathrm{E}(X^k)$ exists, then

- $\mathrm{E}(X^k)$ is the ***k*-th moment of *X***.
- $\mathrm{E}[(X - \mathrm{E}\,X)^k]$ is the ***k*-th central moment of *X***. The case $k = 2$ gives the variance of X.

Proposition 2.11 (Existence of Moments)**.** *Suppose* $\mathrm{E}(|X|^k) < \infty$ *for some* $k > 0$. *Then for* $0 < r < k$, $\mathrm{E}(|X|^r) < \infty$.

Proof. First note

$$|X|^r \leq \mathbb{1}_{|X|<1} + |X|^k \mathbb{1}_X \geq 1.$$

Taking expectation on both sides gives

$$\begin{aligned} \mathrm{E}|X|^r &\leq \mathbb{P}(|X| < 1) + \mathrm{E}\left(|X|^k \mathbb{1}_{|X|\geq 1}\right) \\ &\leq \mathbb{P}(|X| < 1) + \mathrm{E}\left(|X|^k\right) < \infty. \end{aligned}$$

□

Remark 2.12. Using the binomial theorem, we can then show that the k-th moment exists if and only if the r-th central moment exists.

2.4 Probability Inequalities

Theorem 2.13 (Chebychev's Inequality)**.** *Suppose X^r is a non-negative integrable random variable for some $r > 0$. Then for any $\delta > 0$, we have*

$$\mathbb{P}(X \geq \delta) \leq \frac{\mathrm{E}(X^r)}{\delta^r}.$$

Proof. Note that $X^r \geq \delta^r \mathbb{1}_{X \geq \delta}$ and take expectations on both sides. □

Remark 2.14. We can bound the probability that X is large using its moments. When $r = 1$, this is called Markov's Inequality.

Lemma 2.15. *$Y = 0$ almost surely if and only if $\mathrm{E}\, Y^2 = 0$.*

Proof. If $\mathrm{E}\, Y^2 = 0$, then $Y^2 = 0$ as. Otherwise, suppose $\mathbb{P}(Y^2 > 0) = \epsilon$ for some $\epsilon > 0$. Write $\{Y^2 > 0\} = \bigcup_n \{Y^2 > n^{-1}\}$. We have

$$0 < \epsilon = \mathbb{P}(Y^2 > 0) \leq \sum_n \mathbb{P}\left(Y^2 > \frac{1}{n}\right),$$

where we used Boole's inequality.[1] There thus exists N such that $\mathbb{P}(Y^2 > N^{-1}) > 0$. We have

$$Y^2 \geq \frac{1}{N} \mathbb{1}\left(Y^2 > \frac{1}{N}\right)$$

and so $\mathrm{E}(Y^2) \geq N^{-1}\mathbb{P}(Y^2 > N^{-1}) > 0$. □

Theorem 2.16 (Cauchy-Schwarz Inequality)**.** *If $\mathrm{E}(X^2)$ and $\mathrm{E}(Y^2)$ exist, then*

$$|\mathrm{E}(XY)|^2 \leq \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$$

with equality if and only if $X = aY$ almost surely for some constant a.

Proof. If $Y = 0$ as the inequality is trivial. If not, $\mathrm{E}(Y^2) > 0$ and we can write

$$\begin{aligned} 0 &\leq \frac{\mathrm{E}\left\{[X\,\mathrm{E}(Y^2) - Y\,\mathrm{E}(XY)]^2\right\}}{\mathrm{E}(Y^2)} \\ &\leq \frac{\mathrm{E}(X^2)\,\mathrm{E}(Y^2) - 2\,\mathrm{E}(XY)^2\,\mathrm{E}(Y^2) + \mathrm{E}(Y^2)\,\mathrm{E}(XY)^2}{\mathrm{E}(Y^2)} \\ &= \mathrm{E}(X^2)\,\mathrm{E}(Y^2) - \mathrm{E}(XY). \end{aligned}$$

We have equality if and only if $X\,\mathrm{E}(Y^2) - Y\,\mathrm{E}(XY) = 0$ as, which holds if and only if $X = aY$ as for some constant a. □

[1] $\mathbb{P}(\cup A_i) \leq \sum \mathbb{P}(A_i)$.

Corollary 2.17. *The correlation is bounded between* -1 *and* 1*, with equality if and only if* $X - \mathrm{E}\, X = b(Y - \mathrm{E}\, Y)$ *for some constant* b*, which holds if and only if* $X = a + bY$ *for some constants* a, b*.*

2.5 Random Vectors

Definition 2.18. If X and Y are random vectors, then

$$\mathrm{Cov}(X, Y) := \mathrm{E}[(X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y)'].$$

Proposition 2.19. *Let* X *be a random vector such that* $\mathrm{Var}(X)$ *exists. If* A *is a constant matrix and* b *a constant vector, then*

$$\mathrm{Var}(AX + b) = A\, \mathrm{Var}(X) A'.$$

2.6 The Binning Estimator

Consider sample $\{Y_i, X_i\}_{i=1}^n$ with X discrete. The **binning estimator** of $\mathrm{E}(Y|X \in B)$ is

$$\hat{\mu}(B) = \frac{\sum Y_i \mathbb{1}(X_i \in B)}{\sum \mathbb{1}(X_i \in B)}.$$

With continuous X we may use a moving bin of the form $x \pm h$. For large sample we can use smaller h.

2.7 Conditional Expectation

Suppose $\mathrm{E}\, X^2 < \infty$ and $\mathrm{E}\, Y_i < \infty$ for each i. Consider the problem of minimizing

$$\mathrm{E}(Y - g(X))^2.$$

The solution is the **best predictor of** $\boldsymbol{Y}$ **under square loss**. That is ,

$$g^* \in \operatorname*{arg\,min}_{g \in L^2(X)} \mathrm{E}(Y - g(X))^2.$$

Proposition 2.20. $g^*(X) = \mathrm{E}(Y|X)$.

Proof.

$$\begin{aligned}\mathrm{E}(Y - g(X))^2 = \mathrm{E}[Y - \mathrm{E}(Y|X)]^2 + \mathrm{E}[\mathrm{E}(Y|X) - g(X)]^2 \\ + 2\,\mathrm{E}[(Y - \mathrm{E}(Y|X))(\mathrm{E}(Y|X) - g(X))],\end{aligned}$$

where the last term is 0 by the law of iterated expectation. □

2.8 Linear Regression

Proposition 2.21. *Note that* $\mathrm{E}(XX')$ *is always positive semidefinite. Moreover, if it is invertible, then it is positive definite.*

Definition 2.22. There is **perfect collinearity** in X if there exists a constant vector $a \neq 0$ such that $a'X = 0$ almost surely.

Proposition 2.23. *Suppose* X *is a* $(k \times 1)$ *random vector and* $\mathrm{E}(XX')$ *exists. Then* $\mathrm{E}(XX')$ *is invertible if and only if there is no perfect collinearity in* X.

Proof. If $X'a = 0$ as, then

$$\mathrm{E}(XX')a = \mathrm{E}(X(X'a)) = \mathrm{E}(X \cdot 0) = 0.$$

So $\mathrm{E}(XX')$ is not full rank and not invertible. If for any $c \in \mathbb{R}^k \setminus \{0\}$ we have $c'X \neq 0$ with positive probability, then

$$c' \,\mathrm{E}(XX')c = \mathrm{E}[(X'c)^2] > 0.$$

Thus $\mathrm{E}(XX')$ is positive definite and in particular invertible. □

We may restrict $L^2(X)$ to a smaller subset

$$H(X) = \{f : f(X) = X'a \text{ for some } a \in \mathbb{R}^k\}.$$

Then, the **best linear predictor** of Y given X is found by solving

$$\min_{b \in \mathbb{R}^k} \mathrm{E}(Y - X'b)^2.$$

Differentiation gives the FOC $2\,\mathrm{E}(XX')b^* - 2\,\mathrm{E}(XY) = 0$. Provided X_j are not perfectly collinear random variables, $\mathrm{E}(XX')$ is full rank, and so

$$b^* = \mathrm{E}(XX')^{-1}\,\mathrm{E}(XY).$$

Define the **prediction error** or **residual** to be $U = Y - X'b^*$. We have

Proposition 2.24. $\mathrm{E}(XU) = 0$.

Proof. $\mathrm{E}(XU) = \mathrm{E}(XY) - \mathrm{E}(XX')b^* = 0$ by the FOC. □

Consider next the problem

$$\min_{b\in\mathbb{R}^k} \mathrm{E}\left(\mathrm{E}(Y|X) - X'b\right)^2 .$$

The solution is the **best linear approximation to $\mathbf{E}(Y|X)$ under square loss**. Write

$$\begin{aligned}\mathrm{E}(\mathrm{E}(Y|X) - X'b)^2 = \mathrm{E}(Y - \mathrm{E}(Y|X))^2 + \mathrm{E}(Y - X'b)^2 \\ - 2\,\mathrm{E}\left[(Y - \mathrm{E}(Y|X))(Y - X'b)\right] .\end{aligned}$$

By the law of iterated expectation, we have

$$\mathrm{E}\left[(Y - \mathrm{E}(Y|X))(Y - X'b)\right] = \mathrm{E}(Y(Y - \mathrm{E}(Y|X))).$$

Thus,

$$\mathrm{E}\left[(\mathrm{E}(Y|X) - X'b)^2\right] = \mathrm{E}(Y - X'b)^2 + \text{constant}.$$

Remark 2.25. Thus, regression can be interpreted as

- the best linear predictor of Y given X, and
- the best linear approximation to $\mathrm{E}(Y|X)$ under square loss.

3 Estimation and Large Sample Theory

3.1 Convergence

Theorem 3.1 (Weak Law of Laege Numbers)**.** *Suppose* $\{X_i\}_{i\geq 1}$ *is an iid sequence of random variables with* $\mathrm{E}(X_i) = \mu$. *Then,* $\overline{X}_n \rightarrow_{\mathrm{p}} \mu$.

Theorem 3.2 (WLLN for Moments)**.** *If* $\mathrm{E}(X_i^k) < \infty$, *then*

$$\frac{1}{n}\sum_i X_i^k \longrightarrow_{\mathrm{p}} \mathrm{E}(X^k).$$

Example 3.3. Let $\{X_i\}_{i\geq 1}$ be an iid sample drawn from F. Define the **empirical distribution** of F by

$$\hat{F}(x) \coloneqq \frac{1}{n}\sum_i \mathbb{1}(X_i < x).$$

WLLN gives $\hat{F}_n(x) \rightarrow_{\mathrm{p}} F(x)$.

Definition 3.4. We say X_n **converges in *r*-th mean** to X for some $r > 0$ if

$$\mathrm{E}(|X_n - X|^r) \longrightarrow 0.$$

Note that Chebyshev gives $\mathbb{P}(|X_n - X| > \epsilon) \leq \mathrm{E}(|X_n - X|^r)/\epsilon^r \rightarrow 0$ and so:

Proposition 3.5. *If* X_n *converges in* r*-th mean to* X, *then* $X_n \rightarrow_{\mathrm{p}} X$.

Proposition 3.6. *Let* X_n *be a sequence of* $(K \times 1)$ *random vectors. Then,*

- $X_n \rightarrow_{\mathrm{p}} X$ *if and only if* $X_{n,i} \rightarrow_{\mathrm{p}} X_i$ *for* $i = 1, \ldots, k$.
- $X_n \rightarrow X$ *in* r*-th mean if and only if* $X_{n,i} \rightarrow X_i$ *in* r*-th mean for* $i = 1, \ldots, k$.

Definition 3.7. A sequence of random variables X_n **converges in distribution** to X ($X_n \xrightarrow{\mathscr{D}} X$) if

$$F_{X_n}(x) \longrightarrow F_X(x)$$

for all x at which F_X is continuous.

Remark 3.8. This is the weakest notion of convergence. Convergence in probability implies convergence in distribution.

Theorem 3.9 (Continuous Mapping Theorem). *Let $g : \mathbb{R}^k \to \mathbb{R}^n$ be continuous on $S \subset \mathbb{R}^k$ with $\mathbb{P}(X \in S) = 1$. Then the following hold:*

(i) if $X_n \to_{\mathrm{p}} X$, then $g(X_n) \to_{\mathrm{p}} g(X)$.

(ii) If $X_n \xrightarrow{\mathscr{D}} X$, then $g(X_n) \xrightarrow{\mathscr{D}} g(X)$.

Remark 3.10. The theorem does not hold for $X_n \to X$ in r-th moment.

Theorem 3.11 (Slutsky's Theorem). *Suppose $X_n \xrightarrow{\mathscr{D}} X$ and $Y_n \to_{\mathrm{p}} c$ for some constant c. Then,*

$$X_n + Y_n \xrightarrow{\mathscr{D}} X + c, \quad X_n Y_n \xrightarrow{\mathscr{D}} Xc, \quad X_n / Y_n \xrightarrow{\mathscr{D}} X/c \text{ provided } c \neq 0.$$

Remark 3.12. It turns out that under our assumption,

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{\mathscr{D}} \begin{pmatrix} X \\ c \end{pmatrix}.$$

We may then apply the continuous mapping theorem.

Proposition 3.13. *Let $A_n \in \mathbb{R}^{P \times K}$ be a sequence of matrices converging in probability to a constant matrix A. Let B_n be a sequence of $(K \times 1)$ random vectors such that $B_n \xrightarrow{\mathscr{D}} \mathcal{N}(\mu, \Sigma)$. Then,*

$$A_n B_n \xrightarrow{\mathscr{D}} A\,\mathcal{N}(\mu, \Sigma) \sim \mathcal{N}(A\mu, A\Sigma A').$$

Proof. Since the columns of A_n, denoted $\mathrm{vec}(A_n)$ converges in probability to $\mathrm{vec}(A)$, a constant vector, we have

$$\begin{pmatrix} B_n \\ \mathrm{vec}(A_n) \end{pmatrix} \xrightarrow{\mathscr{D}} \begin{pmatrix} \mathcal{N}(\mu, \Sigma) \\ \mathrm{vec}(A) \end{pmatrix}.$$

The continuous mapping theorem then gives

$$A_n B_n \xrightarrow{\mathscr{D}} A\,\mathcal{N}(\mu, \Sigma).$$

Since linear transformations of multivariate normal are also multivariate norm, we have

$$A_n B_n \xrightarrow{\mathscr{D}} \mathcal{N}(A\mu, A\Sigma A').$$

□

Theorem 3.14 (Central Limit Theorem)**.** *Let $\{X_i\}_{i\geq 1}$ be an iid sequence of $(K \times 1)$ random vectors with mean μ and finite variance matrix Σ. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \Sigma).$$

Theorem 3.15 (Delta Method)**.** *Let $\{X_n\}_{n\geq 1}$ be a sequence of $(K \times 1)$ random vectors and suppose*

$$n^r(X_n - c) \xrightarrow{\mathscr{D}} X$$

for some $r > 0$ and constant vector c. Let $g : \mathbb{R}^K \to \mathbb{R}_d$ be differentiable at the point c. Then,

$$n^r(g(X_n) - g(c)) \xrightarrow{\mathscr{D}} \mathrm{D}g(c)X.$$

In particular, if $X \sim \mathcal{N}(0, \Sigma)$, then

$$n^r(g(X_n) - g(c)) \xrightarrow{\mathscr{D}} \mathcal{N}(0, \mathrm{D}g(c)\Sigma\mathrm{D}g(c)').$$

4 Ordinary Least Squares Estimation

Theorem 4.1. *Suppose* $\mathrm{E}(y^2) < \infty$ *and* $\mathrm{E}(x_j^2) < \infty$ *for each* $j = 1, \ldots, k$. *The function* $g(x) \coloneqq \mathrm{E}(y|x)$ *is the best predictor of* y *given* x *under square loss. That is,*

$$\mathrm{E}(y|x) \in \arg\min_{g} \mathrm{E}\left[(y - g(x))^2\right].$$

We may interpret the linear model as a best linear approximation to the conditional mean function. We choose b to solve

$$\min_{b \in \mathbb{R}^k} \mathrm{E}\left(\mathrm{E}(y|x) - x'b\right)^2.$$

If $\mathrm{E}(xx')$ is invertible, we have

$$\beta = \mathrm{E}(xx')^{-1}\,\mathrm{E}(xy).$$

Given a sample $\{y_i, x_i\}_{i=1}^n$, we can solve the analogous sample problem

$$\min_{b \in \mathbb{R}^k} \frac{1}{n} \sum_i (y_i - x_i'b)^2.$$

The FOC is

$$\sum_i x_{ij}(y_i - x_i'\hat{\beta}) = 0, \quad 1 \le j \le k.$$

Or, equivalently,

$$\sum_i x_i(y_i - x_i'\hat{\beta}) = 0.$$