ECMA31000: Introduction to Empirical Analysis

Linear Regression II

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- Last time:
  - Interpretations of the linear model
- Today:
  - Causal Interpretation of the linear model
  - Ordinary least squares
  - Estimating subvectors of the best linear predictor

Definitions

$y \in \mathbb{R}, \quad x \in \mathbb{R}^{k+1}, \quad u \in \mathbb{R}.$

- Let (y, x, u) be a random vector such that y and u are scalar random variables and $x \in \mathbb{R}^{k+1}$.
- Assume the first component of x equals 1:

$$x = (x_0, x_1 \ldots, x_k),$$

  where $x_0 = 1$.
- Let $\beta = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ be a constant vector of unknown parameters such that

$$y = x'\beta + u.$$

- β_0 is the *intercept* and the remaining β_j are the *slope* parameters.

Causal Model

- Suppose we assume that y is determined by the equation

  $$y = x'\beta + u,$$

  where x is observed and u is not.
- The *ceteris paribus* effect of x_j on y holding the other elements of x and u constant is β_j.
- We may assume $\mathrm{E}\left(u\right) = 0$ WLOG, by shifting β_0 accordingly, but $\mathrm{E}\left(xu\right), \mathrm{E}\left(u|x\right)$ etc. are not necessarily equal to 0.
- The statement $\mathrm{E}\left(xu\right) = 0$ is therefore an assumption about the joint distribution of (y, x). If this assumption uniquely determines β, it implies that the best linear predictor of y given x also represents the causal effect of x on y.

Causal Model: Example I revisited

- We observe an iid sample of $\{y_i, d_i\}_{i=1}^{n}$ where y_i is the scrap rate of factory i, and

$$d_i = \begin{cases} 1 & \text{if factory } i \text{ receives a job training grant,} \\ 0 & \text{otherwise.} \end{cases}$$

- The observed outcome y_i is a function of the potential outcomes y_{i0}, y_{i1} and treatment:

$$y_i = d_i y_{i1} + (1 - d_i)\, y_{i0}.$$

- WLOG we write

$$\mathrm{E}\,(y_i | d_i) = \beta_0 + \beta_1 d_i.$$

Causal Model: Example I revisited

- We derived

Average effect on factories that recieve grant

$$\beta_1 = \underbrace{\mathrm{E}\left(y_1|d=1\right) - \mathrm{E}\left(y_0|d=1\right)}_{ATT} + \underbrace{\mathrm{E}\left(y_0|d=1\right) - \mathrm{E}\left(y_0|d=0\right)}_{\text{Selection Bias}}.$$

- Suppose that assignment of grants is statistically independent of the factory's potential outcomes y_0, y_1. This means that factories are not applying for/receiving the grant based on y_0. We say assignment to treatment is "randomized".
- The Selection Bias term vanishes, and

$$\beta_1 = \mathrm{E}\left(y_1|d=1\right) - \mathrm{E}\left(y_0|d=1\right) = \mathrm{E}\left(y_1 - y_0\right),$$

where the second equality holds by random assignment.
- β_1 now represents the Average Treatement Effect (or ATE) of receiving a grant in the population of firms.

Causal Model: Example I revisited

$\beta_1 \neq y_{1i} - y_{0i}$.

- It is still not the case that the model

  $$y_i = \beta_0 + \beta_1 d_i + u_i$$

  $y \approx \beta_0 + \beta_{1i} d_i + u_i$

  has a causal interpretation, even though β_1 may be equal to a parameter we are interested in (namely the ATE).
- The reason is that the grants may not impact firms equally, and β_1 represents the average effect.
- If we go one step further, and assume

  $$y_{i0} = \beta_0 + u_i,$$
  $$y_{i1} = \beta_0 + \beta_1 + u_i,$$

  $y_{1i} - y_{0i} = \beta_1$.

  then $\beta_1 = y_{i1} - y_{i0}$ represents an (homogeneous) additive treatment effect and our model has a causal interpretation.

Linear regression when $\mathrm{E}(xu) = 0$

$E(u|x) = 0$.

- Let (y, x, u) be a random vector such that y and u are scalar random variables and $x \in \mathbb{R}^{k+1}$.
- Assume the first component of x equals 1:

$$x = (x_0, x_1 \ldots, x_k),$$

C-S.

$$E(x_i x_j)^2 \leq E(x_i^2)E(x_j^2)$$

So $E(xx')$ exists.

where $x_0 = 1$.

- Let $\beta = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ be a constant vector of unknown parameters such that

$$y = x'\beta + u.$$

$$E\left(\begin{pmatrix} x_0 \\ \vdots \\ x_k \end{pmatrix} (x_0 \ldots x_k)\right)$$

- Suppose $\mathrm{E}(xu) = 0$, justified according to how we interpret the model.
- Suppose also that $\mathrm{E}\left(x_j^2\right) < \infty$ for $1 \leq j \leq k$, so $\mathrm{E}(xx')$ exists.

Linear regression when $\mathrm{E}(xu) = 0$

- There is perfect collinearity in x if there exists a constant vector $a \neq 0$ such that
$$\mathrm{P}\left(a'x = 0\right) = 1.$$
- We assume there is no perfect collinearity in x. This assumption is equivalent to the condition that $\mathrm{E}(xx')$ is invertible.
- Since $\mathrm{E}(xx')$ is positive semidefinite, it is invertible iff it is positive definite.

$$c' E(xx')c = E(c'xx'c) = E[(x'c)^2] \geq 0.$$

Linear regression when $\mathrm{E}(xu) = 0$

$E(X^2) = 0 \implies X = 0$ w.p. 1.

Lemma

Suppose x is a $(K \times 1)$ random vector and $\mathrm{E}(xx')$ exists. Then $\mathrm{E}(xx')$ is invertible iff there is no perfect collinearity in x.

Proof.

If there is perfect collinearity x, then there exists a vector $a \neq 0$ such that $\mathrm{P}(x'a = 0) = 1$. For this vector,

$$\mathrm{E}(xx')\, a = \mathrm{E}(x(x'a)) = \mathrm{E}(x \cdot 0) = 0.$$

Therefore, $\mathrm{E}(xx')$ is not full column rank and so not invertible. Now suppose there is no perfect collinearity in x. For any vector $c \in \mathbb{R}^{k+1} \backslash \{0\}$,

$$c'\mathrm{E}(xx')\, c = \mathrm{E}\left((x'c)^2\right) > 0,$$

since the expectation equals 0 if and only if $\mathrm{P}(x'c = 0) = 1$ which is ruled out by assumption. □

Linear regression when $\mathrm{E}(xu) = 0$

$u = y - x'\beta.$

- Since $\mathrm{E}(xu) = \mathrm{E}(x(y - x'\beta)) = 0$, we obtain
$$\mathrm{E}(xy) = \mathrm{E}(xx')\,\beta.$$
- Since there is no perfect collinearity in x, $\mathrm{E}(xx')$ is invertible, so we can solve for a unique β:
$$\beta = \mathrm{E}(xx')^{-1}\,\mathrm{E}(xy)\,.$$
- In this case, β is point identified, since it is uniquely determined by $\mathrm{E}(xx'), \mathrm{E}(xy)$.

Linear regression when $\mathrm{E}(xu) = 0$

- If $\mathrm{E}(xx')$ is not invertible, there are multiple solutions to

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

- If $\tilde{\beta}$ satisfies $\mathrm{E}(xx')\,\tilde{\beta} = \mathrm{E}(xy)$, then $\mathrm{E}(xx')\left(\tilde{\beta} - \beta\right) = 0$, so

$$\left(\tilde{\beta} - \beta\right)' \mathrm{E}(xx')\left(\tilde{\beta} - \beta\right) = \mathrm{E}\left(\left[x'\left(\tilde{\beta} - \beta\right)\right]^2\right) = 0,$$

  which implies $\mathrm{P}\left(x'\tilde{\beta} = x'\beta\right) = 1$.

- If we interpret $x'\beta$ as a best linear predictor of y, this says there are multiple best predictors.
- If the model is interpreted causally, however, different values of β imply different ceteris paribus effects, holding u and the other components of x fixed.

Questions?

Least Squares Estimation

- Suppose (y, x, u) satisfy

$$y = x'\beta + u; \quad \mathrm{E}\left(xu\right) = 0.$$

- This assumption is equivalent to

$$\beta \in \operatorname*{arg\,min}_{b \in \mathbb{R}^{k+1}} \mathrm{E}\left(y - x'b\right)^2 .$$

- Given an iid sample $\{y_i, x_i\}_{i=1}^n$, the sample analog minimization problem is

$$\min_{b \in \mathbb{R}^{k+1}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i'b\right)^2 .$$

Least Squares Estimation

- A solution to this minimization problem is called an ordinary least squares estimator.
- The solution is unique if $\sum_{i=1}^n x_i x_i'$ is invertible. In that case:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i y_i .$$

- There is no guarantee that $\sum_{i=1}^n x_i x_i'$ is invertible (a necessary condition is that $n \geq k+1$), but if we assume $\mathrm{E}(xx')$ is invertible, the SLLN implies

$E(xx')^{-1}$ exists $\Leftrightarrow$ $\det(E(xx')^{-1}) \neq 0.$

$$\frac{1}{n}\sum_{i=1}^n x_i x_i' \overset{a.s.}{\to} \mathrm{E}(xx') ,$$

so $\frac{1}{n}\sum_{i=1}^n x_i x_i$ is invertible with probability approaching 1.

Least Squares Estimation

$$P\left(\left\| \frac{1}{n}\sum x_i x_i' - E(x_i x_i') \right\| > \varepsilon \right) \to 0.$$

$$\hat{\beta} = \left(\frac{1}{n}\sum x_i x_i' \right)^{-1} \frac{1}{n}\sum x_i y_i .$$

error term

$$y_i = x_i'\beta + u_i$$

$$y_i = x_i'\hat{\beta} + \hat{u}_i$$

Residual.

- The *i*-th fitted value is:

$$\hat{y}_i = x_i'\hat{\beta}_n.$$

- The *i*-th residual is

$$\hat{u}_i = y_i - \hat{y}_i.$$

- The first order condition of the least squares minimization problem implies

$$\sum_{i=1}^{n} x_i \left(y_i - x_i'\hat{\beta}_n \right) = \sum_{i=1}^{n} x_i \hat{u}_i = 0.$$

$$\Uparrow$$

$$\sum (y_i - x_i'\beta)^2 \xrightarrow[\text{FOC}]{} \sum \left(x_i (y_i - x_i'\hat{\beta}) \right) = 0.$$

Matrix Notation

- First write the model for each observation $i = 1, \ldots, n$:

$$y_i = x_i'\beta + u_i; \qquad \mathrm{E}\left(u_i x_i\right) = 0.$$

- Stack the observations into matrices:

$$n \times 1 \; \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} \overset{n}{=} \underbrace{\overset{k+1}{\begin{pmatrix} - & x_1' & - \\ - & x_2' & - \\ - & \vdots & - \\ - & x_n' & - \end{pmatrix}}}_{X} \beta + \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}}_{U}{}^{n \times 1};$$

$$Y = X\beta + U.$$

Matrix Notation

$Y \in \mathbb{R}^n \quad Xb \in \mathbb{R}^n.$

$$Y - Xb = \begin{pmatrix} y_1 - x_1'b \\ \vdots \\ y_n - x_n'b \end{pmatrix}$$

$$(Y-Xb)'(Y-Xb) = (y_1 - x_1'b)^2 + \cdots + (y_n - x_n'b)^2.$$

- Therefore

$$Y = X\beta + U,$$

  where the matrix X is called the design matrix.

- The least squares minimization problem is

$$\min_{b \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} \left(y_i - x_i'b\right)^2 \equiv \min_{b \in \mathbb{R}^{k+1}} (Y - Xb)'(Y - Xb)$$

$$\equiv \min_{b \in \mathbb{R}^{k+1}} \|Y - Xb\|^2,$$

↑ Euclidean norm on $\mathbb{R}^n$.

  where $\|\cdot\|$ denotes the euclidean norm.

Matrix Notation

- Notice that

$$\sum_{i=1}^{n} x_i x_i' = X'X,$$

$$\sum_{i=1}^{n} x_i y_i = X'Y,$$

$\hat{y}_i = x_i' \hat{\beta}_n .$

$\hat{Y} = X \hat{\beta}_n .$

so the OLS estimator is given by

$$\hat{\beta}_n = \left(X'X\right)^{-1} X'Y.$$

- The vector of fitted values is $\hat{Y} = X\hat{\beta}_n$, and the vector of residuals is $\hat{U} = Y - \hat{Y}$.
- The FOC of the least squares problem is equivalently written as $X'\hat{U} = 0$.

$\sum x_i \hat{u}_i = 0 .$

Matrix Notation

$$\sum x_i x_i' = X'X.$$

$$X'X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} - & x_1' & - \\ - & x_2' & - \\ & \vdots & \\ - & x_n' & - \end{pmatrix}$$

$$= \left(\begin{matrix} | \\ x_1 \\ | \end{matrix} \, (x_{10} \; \cdots\cdots \; x_{1k}) \right) + \left(\begin{matrix} | \\ x_2 \\ | \end{matrix} \, (x_{20}, x_{21}, \ldots, x_{2k}) \right) + \cdots$$

$$= x_1 x_1' + x_2 x_2' + \cdots + x_n x_n' = \sum_{i=1}^{n} x_i x_i'.$$

Projection Theorem

- By definition of $\hat{\beta}_n$, the vector $X\hat{\beta}_n$ is the closest (in the euclidean norm) to Y in the set of all vectors

$$\left\{X^{\cancel{\prime}}b : b \in \mathbb{R}^{k+1}\right\} = S\left(X\right),$$

  where $S\left(X\right)$ denotes the span of the columns of X.

- Given a vector $Y \in \mathbb{R}^n$ and a matrix $X \in \mathbb{R}^{n\times(k+1)}$, the projection of Y onto $S\left(X\right)$ is the vector

$$v^* \in \underset{v \in S(X)}{\arg\min} \left\| Y - v \right\|^2 .$$

- We have in fact found v^* already for a full rank matrix X. The following result characterizes v^*.

Projection Theorem

$$\hat{\beta}_n = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$
$$= \underset{\beta}{\operatorname{argmin}} (Y - X\beta)'(Y - X\beta)$$
$$= \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2$$

Theorem

Let $y \in \mathbb{R}^n$ and let S be any nonempty subspace of $\mathbb{R}^n$ equipped with the dot product. There exists a unique point $\hat{y}$ such that $\|y - \hat{y}\|$ is minimized over S. A necessary and sufficient condition for $\hat{y}$ is that $y - \hat{y}$ is orthogonal to every vector in S.

- The closed subspace we consider is the span of the columns of the design matrix X, so $S = S(X)$.
- The projection theorem asserts there is a unique point $Xb \in S(X)$ minimizing $\|Y - Xb\|$.

Projection Theorem

We are projecting Y

Projection of Y. $\hat{Y} \in S(X)$.

- The necessary and sufficient condition for the minimizer is

$$x'\left(Y - \hat{Y}\right) = 0$$

for all $x \in S(X)$. Since the columns of X are a basis of $S(X)$, this condition is equivalent to

$$X'\left(Y - \hat{Y}\right) = 0.$$

$X'(Y - Xb) = 0.$

- $\hat{Y} \in S(X)$ means $\hat{Y} = Xb$ for some $b \in \mathbb{R}^{k+1}$. Thus

$$\begin{aligned} X'Y = X'Xb \implies & b = \left(X'X\right)^{-1} X'Y \\ \implies & \hat{Y} = X\left(X'X\right)^{-1} X'Y. \end{aligned}$$

- That is, the projection of Y onto $S(X)$ is given by $\hat{Y} = P_X Y$, where $P_X = X\left(X'X\right)^{-1} X'$. P_X is called the projection matrix.

Projections

- The projection matrix is symmetric and satisfies

$$\begin{aligned} P_X P_X &= X\left(X'X\right)^{-1} X'X\left(X'X\right)^{-1} X' \\ &= X\left(X'X\right)^{-1} X' \\ &= P_X, \end{aligned}$$

  which reflects the fact that projecting a vector which already lies in $S(X)$ leaves it unchanged.

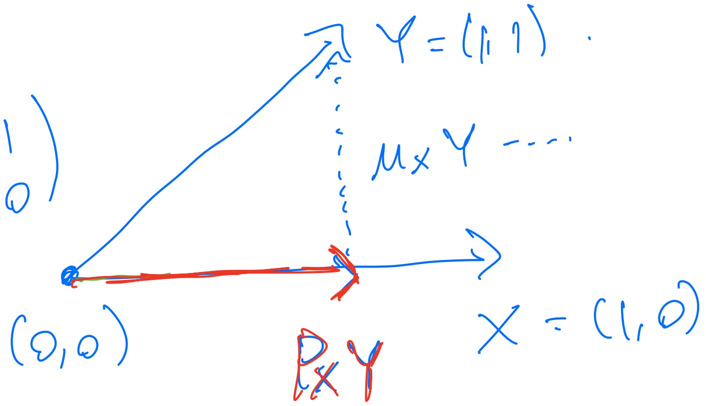
- The residual of the projection of y onto $S(X)$ is

$$\begin{aligned} Y - \hat{Y} &= Y - P_X Y \\ &= \left(I_n - \underbrace{X\left(X'X\right)^{-1} X'}_{P_X}\right) Y. \end{aligned}$$

- Therefore, the matrix $M_X := I_n - X\left(X'X\right)^{-1} X'$ is called the residual maker.

Projections

$Y = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $X = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$



Project Y onto span columns of X.

- $P_X Y$ is the closest point to Y in the span of the columns of X.

- $M_X Y$ is orthogonal to every vector in the span of columns of X.

Properties of Projection Matrices

1. $M_X Y$ in orthogonal complement of X.

2. $Y - M_X Y$ is orthogonal to everything in the orthogonal comp. of X.

Proof (using proj. thm.)

- M_X projects a vector onto the $n-k-1$ dimensional vector space orthogonal to the column space of X. This is called the orthogonal complement of $S(X)$.
- Some properties:
  - $P_X M_X = M_X P_X = 0$, since $P_X (I - P_X) = P_X - P_X P_X = 0$.
  - $P_X X = X (X'X)^{-1} X'X = X$: Projecting X onto its own span does nothing.
  - $M_X X = X - P_X X = 0$: Projecting X onto its orthogonal complement produces 0.
  - For any vector y, $y = P_X y + M_X y = \hat{y} + \hat{u}$.

$$M_X = I_n - P_X \qquad M_X + P_X = I_n$$

$$Y - M_X Y = P_X Y. \qquad P_X y + M_X y = I_n y = y.$$

Questions?

$X(X'X)^{-1}X'Y$.

$v \in$ orthogonal comp of $S(X)$.

$v' \quad X(X'X)^{-1}X'Y$.

So $v'X = 0$.

Partitioned regression

$y = x'\beta + u.$ $\quad y = x_1'\beta_1 + x_2'\beta_2 + u$

x is $k \times 1.$

- Suppose we partition x into $x = (x_1', x_2')'$ and β into $\beta = (\beta_1', \beta_2')'$, each with dimensions (k_1, k_2) and write

x_1 is $k_1 \times 1$

$$y = x_1'\beta_1 + x_2'\beta_2 + u; \quad \mathrm{E}(xu) = 0.$$

x_2 is $k_2 \times 1$

We have $\quad x'\beta = x_0\beta_0 + x_1\beta_1 + \dots + x_k\beta_k.$

$k_1 + k_2 = k.$

$$\beta = \mathrm{E}(xx')^{-1}\mathrm{E}(xy) = \begin{pmatrix} \mathrm{E}(x_1x_1') & \mathrm{E}(x_1x_2') \\ \mathrm{E}(x_2x_1') & \mathrm{E}(x_2x_2') \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{E}(x_1y) \\ \mathrm{E}(x_2y) \end{pmatrix}.$$

- We will not compute this explicitly, but arrive at a formula for β_2 which shows us that the projection coefficient β_2 represents the change in the best linear predictor of y given a change in x_2, while in a sense "holding x_1 fixed".

$xx' = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} (x_1' \; x_2') \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad E(xx') = E\begin{pmatrix} x_1x_1' & x_1x_2' \\ x_2x_1' & x_2x_2' \end{pmatrix}$

Partitioned regression

$$\beta = E(XX')^{-1}E(Xy).$$

$$y = x_1'\tilde{\beta}_1 + \tilde{y}.$$

- Let $\tilde{\beta}_1$ represent the best linear predictor of y given x_1. Define

$$X_{2j} = X_1'\tilde{\gamma}_j + \tilde{X}_{2j}$$

$$\tilde{y} = y - x_1'\tilde{\beta}_1.$$

$$X_2 = \begin{pmatrix} X_{21} \\ \vdots \\ X_{2k_2} \end{pmatrix}$$

- Let $\tilde{\gamma}$ represent the matrix of best linear predictors of x_2 given x_1:

$$\tilde{\gamma} = \begin{pmatrix} \tilde{\gamma}_1' \\ \tilde{\gamma}_2' \\ \vdots \\ \tilde{\gamma}_{k_2}' \end{pmatrix} = \begin{pmatrix} \left[\mathrm{E}\left(x_1x_1'\right)^{-1}\mathrm{E}\left(x_1x_{21}\right)\right]' \\ \left[\mathrm{E}\left(x_1x_1'\right)^{-1}\mathrm{E}\left(x_1x_{22}\right)\right]' \\ \vdots \\ \left[\mathrm{E}\left(x_1x_1'\right)^{-1}\mathrm{E}\left(x_1x_{2k_2}\right)\right]' \end{pmatrix}.$$

- Now define

$$\tilde{x}_2 = x_2 - \tilde{\gamma}x_1.$$

$$\tilde{\gamma}_j = E(X_1X_1')^{-1}E(X_1X_{2j})$$

$$X_2 = \tilde{\gamma}X_1 + \tilde{X}_2$$

Partitioned regression

$y = x'\beta + u.$ $\beta = E(xx')^{-1}E(xy).$

$x'\beta = x'E(xx')^{-1}E(xy$

- Finally, consider the best linear predictor of $\tilde{y}$ given $\tilde{x}_2$. We have

$\tilde{y} = y - x_1'\hat{\beta}_1$ -

$$\tilde{y} = \tilde{x}_2'\bar{\beta}_2 + v; \qquad \mathrm{E}(\tilde{x}_2 v) = 0,$$

where

$y = x_1'\beta_1 + \tilde{y}$ $E(\tilde{x}_2 y) = \cancelto{0}{E(\tilde{x}_2 x_1'\beta_1)} + E(\tilde{x}_2\tilde{y}).$

$x_2 = x_1'\tilde{\gamma} + \tilde{x}_2$

$$\bar{\beta}_2 = \mathrm{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathrm{E}(\tilde{x}_2\tilde{y}).$$
$$= \mathrm{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathrm{E}(\tilde{x}_2 y),$$

$E(\tilde{x}_2 x_1') = 0$

and the last equality follows because $\mathrm{E}(\tilde{x}_2 x_1') = 0$, since $\tilde{x}_2$ is the residual of a linear projection of x_2 onto x_1.

- This shows that $\bar{\beta}_2$ also represents the best linear predictor of y given $\tilde{x}_2$.

Partitioned regression

$$\underset{0}{\underset{\shortparallel}{E(\tilde{x}_2 x_1')}}\beta_1 + E(\tilde{x}_2 x_2' \beta_2) + E(\tilde{x}_2 u) .$$

- We now show that $\bar{\beta}_2 = \beta_2$. Plug y into the expression for $\bar{\beta}_2$:

$$\begin{aligned}
\bar{\beta}_2 &= \mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right)^{-1}\mathrm{E}\left(\tilde{x}_2\left[\overset{y}{x_1'\beta_1 + x_2'\beta_2 + u}\right]\right) \\
&= \mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right)^{-1}\mathrm{E}\left(\tilde{x}_2 x_2'\right)\beta_2 + \mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right)^{-1}\mathrm{E}\left(\tilde{x}_2 u\right) \\
&= \underbrace{\mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right)^{-1}\mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right)}_{I_{k_2}}\beta_2 + \mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right)^{-1}\mathrm{E}\left(\left[x_2 - \tilde{\gamma}x_1\right]u\right) \\
&= \beta_2,
\end{aligned}$$

$$E(x_2 u) = 0.$$

$$E(\tilde{\gamma} x_1 u) = \tilde{\gamma}\,\underset{=0}{E(x_1 u)}.$$

where the third equality follows because

$$\mathrm{E}\left(\tilde{x}_2\tilde{x}_2'\right) = \mathrm{E}\left(\tilde{x}_2\left[x_2 - \tilde{\gamma}x_1\right]'\right) = \mathrm{E}\left(\tilde{x}_2 x_2'\right), \quad \underset{0}{\underset{\shortparallel}{- E(\tilde{x}_2 x_1' \tilde{\gamma}')}}$$

and the fourth holds because

$$\mathrm{E}\left(xu\right) = \mathrm{E}\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} u\right) = 0.$$

Partitioned regression

- In summary, the best predictor of y given x may be decomposed as $BLP(y|x) = x_1'\beta_1 + x_2'\beta_2$, and we can equivalently characterize the subvector β_2 of β as representing the best linear predictor of y given $\tilde{x}_2$.
- Since $\tilde{x}_2$ is the residual of a linear projection of x_2 on x_1, we can think of $x_2'\beta_2$ as the best linear predictor of y given x_2 after "controlling for" x_1.
- x_1 and x_2 are generally correlated, so an increase in x_1 would change our best linear predictor of x_2 by $\tilde{\gamma}\,(\Delta x_1)$. If we don't account for this, we may falsely attribute an increase in y to x_2, when in fact changes in x_1 drive both changes in x_2 and y.
- $\beta_2 x_2$ is generally not the best linear predictor of y given x_2:

Omitted Variables Bias

$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$

- Suppose x_1 and x_2 are scalar random variables, and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

$E(x_1 u) = 0$

$E(x_2 u) = 0$.

where $(\beta_0, \beta_1, \beta_2)$ represents the best linear predictor of y given x_1 and x_2.

- Suppose we omit x_2 and instead specify the regression

$$y = b_0 + b_1 x_1 + v,$$

where (b_0, b_1) represents the best linear predictor of y given x_1.

Omitted Variables Bias

$y = b_0 + b_1 x_1 + v.$

$b_1 = \frac{Cov(y, x_1)}{Var(x_1)}.$

- $b_1 \neq \beta_1$ in general because

$$
\begin{aligned}
b_1 &= \frac{Cov(y, x_1)}{Var(x_1)} \\
&= \frac{\beta_1 Var(x_1) + \beta_2 Cov(x_1, x_2)}{Var(x_1)} \\
&= \beta_1 + \underbrace{\beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)}}_{\text{Omitted Variable Bias}}.
\end{aligned}
$$

$\to cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2, x_1)$

$= \beta_1 Var(x_1) + \beta_2 Cov(x_2, x_1)$

→ Inconsistency in OLS estimator resulting from omitting x_2.

Estimate $y = b_0 + b_1 x_1 + v$ using OLS. $\hat{b}_1 \xrightarrow{p} b_1$

$= \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)}$

Questions?

Frisch-Waugh-Lovell Decomposition

$$Y = X\beta + U.$$
$$= X\hat{\beta} + \hat{U}.$$

Fitted Value ↑ Residual.

- The OLS estimator of a sub-vector of β can be obtained in a similar manner to the population parameter.
- Suppose we observe a sample of $\{y_i, x_i\}_{i=1}^{n}$ which we stack to form the equation

$$Y = X_1\beta_1 + X_2\beta_2 + U.$$

- Let $P_{X_1}, P_{X_2}, M_{X_1}, M_{X_2}$ denote the projection and residual maker matrices of X_1, X_2, respectively.
- Let $\hat{\beta} = \left(\hat{\beta}_1', \hat{\beta}_2'\right)'$ denote the OLS estimator of $\beta = (\beta_1', \beta_2')'$, and $\hat{U}$ the residuals, and suppose we want the OLS estimate of β_2.

Frisch-Waugh-Lovell Decomposition

- Note that

$$Y = \overbrace{X_1\hat{\beta}_1 + X_2\hat{\beta}_2}^{P_X Y} + \overbrace{\hat{U}}^{M_X Y.},$$

so

$$\begin{aligned} M_{X_1} Y &= \overbrace{M_{X_1} X_1}^{=0}\hat{\beta}_1 + M_{X_1} X_2\hat{\beta}_2 + M_{X_1}\hat{U} \\ &= M_{X_1} X_2\hat{\beta}_2 + \hat{U}, \end{aligned}$$

where the last equality follows because

$$M_{X_1}\hat{U} = \hat{U} - X_1\left(X_1'X_1\right)^{-1} X_1'\hat{U} = \hat{U}.$$

$(I - P_{X_1})\hat{U}$

$X_1'\hat{U} = 0$ (Normal equations)

FOC of OLS problem $\sum x_i \hat{u}_i = 0$

$X'\hat{U} = 0.$

Frisch-Waugh-Lovell Decomposition

$X_1'\hat{U} = 0.$

- Next, multiply by X_2' to give

$$\begin{aligned} X_2' M_{X_1} Y &= X_2' M_{X_1} X_2 \hat{\beta}_2 + \overset{=0}{X_2'\hat{U}} \\ &= X_2' M_{X_1} X_2 \hat{\beta}_2, \end{aligned}$$

since the regressors are orthogonal to the residuals ($X'\hat{U} = 0$).

- It follows that

$M_{X_1} \cdot M_{X_1} = M_{X_1}$

$M_{X_1}' = M_{X_1}$

$$\begin{aligned} \hat{\beta}_2 &= \left(X_2' M_{X_1} X_2\right)^{-1} X_2' M_{X_1} Y \\ &= \left[(M_{X_1} X_2)' M_{X_1} X_2\right]^{-1} (M_{X_1} X_2)' M_{X_1} Y \end{aligned}$$

since M_{X_1} is idempotent ($M_{X_1} M_{X_1} = M_{X_1}$).

- Therefore, the OLS estimate is also obtained by regressing the residuals of Y on X_1 on the residuals of X_2 on X_1.

$X_2' M_{X_1} X_2 = X_2' M_{X_1}' M_{X_1} X_2 = (M_{X_1} X_2)' M_{X_1} X_2.$

Finite Sample properties of OLS: Bias

- We have

$$\begin{aligned}\mathrm{E}\left(\hat{\beta}_n|X\right) &= \mathrm{E}\left(\left(X'X\right)^{-1}X'Y|X\right)\\ &= \left(X'X\right)^{-1}X'\mathrm{E}\left(Y|X\right)\\ &= \left(X'X\right)^{-1}X'X\beta\\ &= \beta.\end{aligned}$$

- By the Law of Iterated Expectation:

$$\mathrm{E}\left(\hat{\beta}_n\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{\beta}_n|X\right)\right) = \beta,$$

so $\hat{\beta}_n$ is unbiased.

Finite Sample properties of OLS: Variance

- Suppose in addition that $Var\left(u_i|x_i\right) = \sigma^2$.
- Since the error variance does not depend on the value of x_i, it is called <u>homoskedastic</u>.
- If the error variance depends on x_i, u_i is <u>heteroskedastic</u>.
- Under homoskedasticity,

$$Var\left(U|X\right) = \sigma^2 I_n,$$

because the (i, j) entry is given by

$$Var\left(U|X\right)_{i,j} = \mathrm{E}\left(u_i u_j|X\right) = \begin{cases} \mathrm{E}\left(u_i^2|x_i\right) & i = j \\ \mathrm{E}\left(u_i u_j|x_i, x_j\right) & i \neq j \end{cases}$$

Finite Sample properties of OLS: Variance

- We have $\mathrm{E}\left(u_i^2|x_i\right) = \sigma^2$, and

$$\begin{aligned}\mathrm{E}\left(u_i u_j|x_i, x_j\right) &= \mathrm{E}\left(u_i \mathrm{E}\left(u_j|u_i, x_i, x_j\right)|x_i, x_j\right)\\ &= \mathrm{E}\left(u_i \mathrm{E}\left(u_j|x_j\right)|x_i, x_j\right) = 0.\end{aligned}$$

- Under heteroskedasticity, the off diagonal elements will still be 0 but $\mathrm{E}(u_i^2|x_i) = \sigma\left(x_i\right)^2$, so

$$Var\left(U|X\right) = \begin{pmatrix} \sigma\left(x_1\right)^2 & 0 & 0 & 0 \\ 0 & \sigma\left(x_2\right)^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma\left(x_n\right)^2 \end{pmatrix} := \Omega.$$

Finite Sample properties of OLS: Variance

- We now conclude that

$$\begin{aligned} Var\left(\hat{\beta}_n | X\right) &= Var\left(\beta + \left(X'X\right)^{-1} X'U | X\right) \\ &= \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}. \end{aligned}$$

- Under homoskedasticity, $\Omega = \sigma^2 I_n$, and so the variance becomes

$$\begin{aligned} Var\left(\hat{\beta}_n | X\right) &= \sigma^2 \left(X'X\right)^{-1} X' I_n X \left(X'X\right)^{-1} \\ &= \sigma^2 \left(X'X\right)^{-1}. \end{aligned}$$

Finite Sample properties of OLS: Variance

- Suppose $\mathrm{E}\left(u_i|x_i\right) = 0$ and the error term u_i is homoskedastic.
- Under these assumptions, the OLS estimator is the "best linear unbiased estimator", which means it has the "smallest" variance in the class of linear estimators that are also unbiased conditional on X. This result is known as the *Gauss-Markov Theorem*.
- Our goal is to show that

$$\sigma^2 \left(X'X\right)^{-1}$$

is "smaller" than the conditional variance of any other estimator $\tilde{\beta} = A\left(X\right) Y$ which also satisfies $\mathrm{E}\left(\tilde{\beta}|X\right) = \beta$.

- Precisely: $Var\left(\tilde{\beta}|X\right) - Var\left(\hat{\beta}_n|X\right) = D$ is a positive-semidefinite matrix.

Gauss-Markov Theorem

- In particular, let r be a $(k+1) \times 1$ vector, and let $r'\beta$ denote a linear combination of the β_j.
- The Gauss-Markov theorem implies that $r'\hat{\beta}$ is the Best Linear Unbiased Estimator of $r'\beta$, since

$$\begin{aligned} Var\left(r'\tilde{\beta}|X\right) - Var\left(r'\hat{\beta}|X\right) &= r'Var\left(\tilde{\beta}|X\right)r - r'Var\left(\hat{\beta}|X\right)r \\ &= r'\left[Var\left(\tilde{\beta}|X\right) - Var\left(\hat{\beta}|X\right)\right]r \\ &= r'Dr \geq 0. \end{aligned}$$

- For example, if $r = \left(0, 0 \ldots, 0, \underbrace{1}_{(j+1)st \text{ position}}, 0, \ldots 0\right)$, then $r'\hat{\beta} = \hat{\beta}_j$, and so $Var\left(\hat{\beta}_j|X\right) \leq Var\left(\tilde{\beta}_j|X\right)$.

Gauss-Markov Theorem

- First note that a linear estimator $\tilde{\beta}$ of β satisfies

  $$\tilde{\beta} = A\left(x_1, \ldots, x_n\right) Y = AY,$$

  for some $(k+1) \times N$ matrix $A = A\left(x_1, \ldots, x_n\right)$ depending only on the sample $\{x_i\}_{i=1}^{n}$.
- No matter what the true value of the unknown parameter β is, $\tilde{\beta}$ must also satisfy

  $$\mathrm{E}\left(\tilde{\beta}|X\right) = \beta.$$

- This implies

  $$\mathrm{E}\left(\tilde{\beta}|X\right) = \mathrm{E}\left(AY|X\right) = AX\beta = \beta.$$

Gauss-Markov Theorem

- Since the final equality must hold for any β, it must be the case that $AX = I_{k+1}$.
- Next, we compute the variance of AY conditional on X:

$$\begin{aligned} Var\left(AY|X\right) &= AVar\left(Y|X\right)A' \\ &= AVar\left(U|X\right)A' \\ &= \sigma^2 AA'. \end{aligned}$$

- If $A = \left(X'X\right)^{-1}X'$, we obtain the OLS estimator, with variance $\sigma^2\left(X'X\right)^{-1}$.

Gauss-Markov Theorem

- It remains to show that $Var\left(\hat{\beta}_{OLS}|X\right) \leq Var\left(\tilde{\beta}|X\right)$, where $\leq$ means that
$$\sigma^2 AA' - \sigma^2 \left(X'X\right)^{-1}$$
is a positive semi-definite matrix for any A such that $AX = I_{k+1}$.
- To this end, define
$$C = A - \left(X'X\right)^{-1} X'.$$

Gauss-Markov Theorem

- Note that

$$
\begin{aligned}
AA' - \left(X'X\right)^{-1} &= \left(C + \left(X'X\right)^{-1} X'\right)\left(C + \left(X'X\right)^{-1} X'\right)' \\
&\quad - \left(X'X\right)^{-1} \\
&= CC' + \left(X'X\right)^{-1} X'C' + CX\left(X'X\right)^{-1} \\
&= CC',
\end{aligned}
$$

  where the final equality holds since

$$
CX = AX - \left(X'X\right)^{-1} X'X = I_{k+1} - I_{k+1} = 0.
$$

- The conclusion follows because CC' is always positive semi-definite.

Questions?

Large Sample properties of OLS

- Now drop the assumptions that $\mathrm{E}\left(u_i|x_i\right)=0$ and $Var\left(u_i|x_i\right)=\sigma^2$. Rewrite the model as

$$y_i = x_i'\beta + u_i; \qquad \mathrm{E}\left(u_i x_i\right)=0,$$

  and suppose $\mathrm{E}\left(xx'\right)$ exists and is invertible.

- The OLS estimator is consistent, because

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \overset{a.s.}{\to} \mathrm{E}\left(xx'\right);$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i \overset{a.s.}{\to} \mathrm{E}\left(xy\right),$$

  by the SLLN. As we know, the convergence is joint, so

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} x_i y_i \overset{a.s.}{\to} \mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xy\right)=\beta.$$

Asymptotic Normality of OLS

- Maintain the assumption that $\mathrm{E}\left(xx'\right)$ exists, and assume $Var\left(xu\right) = \mathrm{E}\left(u^2xx'\right)$ exists also. Then:

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right),$$

  where $\Sigma = \mathrm{E}\left(xx'\right)^{-1} Var\left(xu\right) \mathrm{E}\left(xx'\right)^{-1}$.
- This follows because

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i.$$

Asymptotic Normality of OLS

- Since the sequence of vectors $\left\{(y_i, x_i')'\right\}_{i\geq 1}$ is iid, the sequence $\{x_i\left(y_i - x_i'\beta\right)\}_{i\geq 1}$ is iid. Therefore, the CLT implies:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \overset{d}{\to} \mathcal{N}\left(0, Var\left(xu\right)\right).$$

- Applying Slutsky's Theorem gives

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \overset{d}{\to} \mathcal{N}\left(0, \Sigma\right),$$

as desired.

Estimation of Σ

- Deriving asymptotic normality of $\hat{\beta}_n$ will enable us to test hypotheses about the unknown parameter β.
- Since we do not know Σ, we must construct a consistent estimator of it to yield an asymptotic distribution whose quantiles are known so we may conduct tests.
- First, suppose $\mathrm{E}\left(u|x\right) = 0$ and $Var\left(u|x\right) = \sigma^2$. Then

$$\begin{aligned} Var\left(xu\right) &= \mathrm{E}\left(u^2 xx'\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(u^2|x\right) xx'\right) \\ &= \sigma^2 \mathrm{E}\left(xx'\right). \end{aligned}$$

- It follows that

$$\Sigma = \sigma^2 \mathrm{E}\left(xx'\right)^{-1}.$$

Estimation of Σ

- A natural estimator of $\mathrm{E}\left(xx'\right)^{-1}$ is $\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}$.
- It remains to find a consistent estimator of σ^2. We use

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{1}{n}\left\| M_X Y \right\|^2,$$

  since $M_X Y$ is the vector of residuals of a regression of y on x.
- Next, note that $M_X Y = M_X\left(X\beta + U\right) = M_X U$,so

$$\begin{aligned}\left\| M_X Y \right\|^2 &= \left\| M_X U \right\|^2 \\ &= U' M_X' M_X U \\ &= U' M_X U \\ &= U'U - U' P_X U.\end{aligned}$$

Estimation of Σ

- Finally, note that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} u_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i\right)\\ &\overset{a.s.}{\to} \sigma^2 - 0 \cdot \mathrm{E}\left(xx'\right)^{-1} \cdot 0\\ &= \sigma^2,\end{aligned}$$

  by the continuous mapping theorem.

- In summary, a consistent estimator of Σ is given by

$$\hat{\Sigma} = \hat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}.$$

Estimation of Σ

- If we do not assume $\mathrm{E}\left(U|X\right) = 0$ and $Var\left(U|X\right) = \sigma^2 I_n$, Σ does not simplify, and we use

$$\hat{\Sigma} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}.$$

- Proving consistency boils down to showing that

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i' \xrightarrow{p} \mathrm{E}\left(u^2 x x'\right).$$

- First, decompose this quantity as

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 x_i x_i' = \frac{1}{n}\sum_{i=1}^{n} u_i^2 x_i x_i' + \frac{1}{n}\sum_{i=1}^{n} \left(\hat{u}_i^2 - u_i^2\right) x_i x_i'.$$

Estimation of Σ

- We have
$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 x_i x_i' \overset{a.s.}{\rightarrow} \mathrm{E}\left(u^2 x x'\right),$$
so it remains to show that
$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right) x_i x_i' = o_p\left(1\right).$$
- We prove that every element of this matrix is $o_p\left(1\right)$.

Estimation of Σ

- Fix j, k element $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right)x_{i,j}x_{i,k}$, and observe that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right)x_{i,j}x_{i,k}\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\left(\hat{u}_i^2 - u_i^2\right)\right|\left|x_{i,j}x_{i,k}\right|$$
$$\leq \max_{1\leq i\leq n}\left|\hat{u}_i^2 - u_i^2\right|\frac{1}{n}\sum_{i=1}^{n}\left|x_{i,j}x_{i,k}\right|.$$

- Since

$$\frac{1}{n}\sum_{i=1}^{n}\left|x_{i,j}x_{i,k}\right| \overset{a.s.}{\to} \mathrm{E}\left(\left|x_{i,k}x_{i,k}\right|\right),$$

this term is $O_p(1)$. Therefore, it is sufficient to prove

$$\max_{1\leq i\leq n}\left|\hat{u}_i^2 - u_i^2\right| = o_p(1).$$

Estimation of Σ

- To this end, note that

$$\hat{u}_i - u_i = x_i' \left(\beta - \hat{\beta}_n\right),$$

so since $\hat{u}_i = y_i - x_i'\hat{\beta}_n = x_i' \left(\beta - \hat{\beta}_n\right) + u_i$, we have:

$$\begin{aligned}
\left|\hat{u}_i^2 - u_i^2\right| &= \left|x_i' \left(\beta - \hat{\beta}_n\right)(\hat{u}_i + u_i)\right| \\
&= \left|x_i' \left(\beta - \hat{\beta}_n\right)\left(x_i' \left(\beta - \hat{\beta}_n\right) + 2u_i\right)\right| \\
&= \left|\left(x_i' \left(\beta - \hat{\beta}_n\right)\right)^2 + 2u_i x_i' \left(\beta - \hat{\beta}_n\right)\right| \\
&\leq \left|\left(x_i' \left(\beta - \hat{\beta}_n\right)\right)^2\right| + 2\left|u_i x_i' \left(\beta - \hat{\beta}_n\right)\right|.
\end{aligned}$$

Estimation of Σ

- Next, using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\max_{1\leq i\leq n}\left|\hat{u}_i^2-u_i^2\right| &\leq \left\|\beta-\hat{\beta}_n\right\|^2\max_{1\leq i\leq n}\left\|x_i\right\|^2 \\ &\quad +2\left\|\beta-\hat{\beta}_n\right\|\max_{1\leq i\leq n}\left\|x_iu_i\right\| \\ &=\left\|\sqrt{n}\left(\beta-\hat{\beta}_n\right)\right\|^2\frac{\max_{1\leq i\leq n}\left\|x_i\right\|^2}{n} \\ &\quad +2\left\|\sqrt{n}\left(\beta-\hat{\beta}_n\right)\right\|\frac{\max_{1\leq i\leq n}\left\|x_iu_i\right\|}{\sqrt{n}}.\end{aligned}$$

Estimation of Σ

- Since $\sqrt{n}\left(\beta - \hat{\beta}_n\right) = O_p(1)$, we must show:

$$\frac{\max_{1 \leq i \leq n} \|x_i\|^2}{n} = o_p(1);$$

$$\frac{\max_{1 \leq i \leq n} \|x_i u_i\|}{\sqrt{n}} = o_p(1).$$

Lemma

Let $\{Z_i\}_{i \geq 1}$ be a sequence of identically distributed random vectors such that $\mathrm{E}\left(\|Z_i\|^r\right) < \infty$. Then

$$\frac{\max_{1 \leq i \leq n} \|Z_i\|}{n^{1/r}} \xrightarrow{p} 0.$$

Estimation of Σ

Proof.

Fix $\epsilon > 0$ and note that

$$\begin{aligned}
\mathrm{P}\left(\max_{1\leq i\leq n}\|Z_i\| > \epsilon n^{1/r}\right) &= \mathrm{P}\left(\cup_{i=1}^{n}\left\{\|Z_i\|^r > \epsilon^r n\right\}\right) \\
&\leq \sum_{i=1}^{n}\mathrm{P}\left(\|Z_i\|^r > \epsilon^r n\right) \\
&= \sum_{i=1}^{n}\mathrm{P}\left(\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right) > \epsilon^r n\right) \\
&\leq \frac{1}{n\epsilon^r}\sum_{i=1}^{n}\mathrm{E}\left(\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right)\right) \\
&= \frac{1}{\epsilon^r}\mathrm{E}\left(\|Z_i\|^r \mathbf{1}\left(\|Z_i\|^r > \epsilon^r n\right)\right) \\
&\to 0.
\end{aligned}$$

□

Estimation of Σ

- The second equality holds because

$$\{\|Z_i\|^r > \epsilon^r n\} = \{\|Z_i\|^r \mathbf{1}(\|Z_i\|^r > \epsilon^r n) > \epsilon^r n\}.$$

- The second inequality follows by Markov's inequality and the third because the Z_i have identical distribution.
- Convergence to 0 holds because $\mathrm{E}(\|Z_i\|^r) < \infty$.
- Finally, since $\mathrm{E}\left(\|x\|^2\right) < \infty$ and $\mathrm{E}\left(\|ux\|^2\right) < \infty$, we conclude that

$$\max_{1 \le i \le n} \left|\hat{u}_i^2 - u_i^2\right| = o_p(1),$$

so

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{u}_i^2 - u_i^2\right) x_i x_i' = o_p(1).$$

Questions?