OLS Estimators

450A Political Methodology I, handout

Assumptions I–VI refer to the lectures: (I) linearity, (II) random sampling, (III) variation in X (no perfect collinearity), (IV) zero conditional mean, (V) homoskedasticity, (VI) normality of the errors. In the scalar sections, conditioning on X means conditioning on all the sample values $x_1, \ldots, x_n$. The limit theorems used in Section 5 are stated in Section 6.

1 OLS setup

- (X, Y) are jointly distributed, with $Y = \beta_0 + \beta_1 X + u$ and $\mathbb{E}[u \mid X] = 0$
- We observe an i.i.d. sample $\{(x_i, y_i)\}_{i=1}^n$, that is, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- OLS estimators: $\hat{\beta}_0$ and $\hat{\beta}_1$
- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residuals: $\hat{u}_i \equiv y_i - \hat{y}_i$
- $y_i = \hat{y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$ by definition

2 Deriving the OLS estimators

- Objective function: minimize the sum of squared residuals

$$\begin{aligned} S(\tilde{\beta}_0, \tilde{\beta}_1) &= \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 \\ &= \sum_{i=1}^n (y_i^2 - 2y_i\tilde{\beta}_0 - 2y_i\tilde{\beta}_1 x_i + \tilde{\beta}_0^2 + 2\tilde{\beta}_0\tilde{\beta}_1 x_i + \tilde{\beta}_1^2 x_i^2) \end{aligned}$$

- First-order conditions (FOCs): at the minimum $(\hat{\beta}_0, \hat{\beta}_1)$ both partial derivatives are zero,

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_0} = \sum_{i=1}^n (-2y_i + 2\tilde{\beta}_0 + 2\tilde{\beta}_1 x_i) = 0 \tag{1}$$

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_1} = \sum_{i=1}^n (-2y_i x_i + 2\tilde{\beta}_0 x_i + 2\tilde{\beta}_1 x_i^2) = 0 \tag{2}$$

- Solve the first FOC (1):

$$\begin{aligned} 0 &= \sum_{i=1}^{n}(-2y_i + 2\hat{\beta}_0 + 2\hat{\beta}_1 x_i) \\ n\hat{\beta}_0 &= \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ \therefore\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

- Solve the second FOC (2), substituting $\hat{\beta}_0$:

$$\begin{aligned} 0 &= \sum_{i=1}^{n}(-2y_i x_i + 2\hat{\beta}_0 x_i + 2\hat{\beta}_1 x_i^2) \\ \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i \\ \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i \\ \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i \\ \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \hat{\beta}_1 \frac{1}{n} \Big(\sum_{i=1}^{n} x_i \Big)^2 \\ \hat{\beta}_1 \Big(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \Big(\sum_{i=1}^{n} x_i \Big)^2 \Big) &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i \\ \therefore\ \hat{\beta}_1 &= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \big(\sum_{i=1}^{n} x_i \big)^2} \end{aligned}$$

- Show that $\hat{\beta}_1 = \widehat{\mathrm{Cov}}(x_i, y_i) / \widehat{\mathrm{Var}}(x_i)$.

  Expand the numerator and denominator of the deviation form:

$$\begin{aligned} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} &= \frac{\sum_{i=1}^{n}(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})}{\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2)} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\big(\sum_{i=1}^{n} x_i \big)^2}, \end{aligned}$$

  which is the expression for $\hat{\beta}_1$ above. Dividing the numerator and denominator by $n - 1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\widehat{\mathrm{Cov}}(x_i, y_i)}{\widehat{\mathrm{Var}}(x_i)}.$$

3 Properties by construction

- $\sum_{i=1}^{n} \hat{u}_i = 0$

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \text{ which is FOC (1) divided by } -2$$

- $\sum_{i=1}^{n} x_i \hat{u}_i = 0$

$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \text{ which is FOC (2) divided by } -2$$

- $\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) \hat{u}_i = \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i \hat{u}_i = 0$$

- $\bar{\hat{y}} = \bar{y}$

$$\bar{\hat{y}} = \bar{y} \quad \Leftrightarrow \quad \sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i \quad \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \hat{u}_i = 0,$$

  which holds by the first property.

4 Sums of squares

- $\text{SST} = \text{SSE} + \text{SSR}$, where $\text{SST} = \sum_i (y_i - \bar{y})^2$ (total), $\text{SSE} = \sum_i (\hat{y}_i - \bar{y})^2$ (explained) and $\text{SSR} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{u}_i^2$ (residual).

  Add and subtract $\hat{y}_i$ inside the square:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} \big[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \big]^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^{n} \hat{u}_i (\hat{y}_i - \bar{y}) \\
&= \text{SSR} + \text{SSE} + 2 \Big(\sum_{i=1}^{n} \hat{u}_i \hat{y}_i - \bar{y} \sum_{i=1}^{n} \hat{u}_i \Big) \\
&= \text{SSR} + \text{SSE},
\end{aligned}$$

  because $\sum_i \hat{u}_i \hat{y}_i = 0$ and $\sum_i \hat{u}_i = 0$ (Section 3).

5 Statistical properties

- **Unbiasedness**: $\mathbb{E}[\hat{\beta}_0] = \beta_0$ and $\mathbb{E}[\hat{\beta}_1] = \beta_1$ (Assumptions I–IV).

  Using $\sum_i (x_i - \bar{x}) = 0$ twice (so $\sum_i (x_i - \bar{x})\bar{y} = 0$ and $\sum_i (x_i - \bar{x}) x_i = \sum_i (x_i - \bar{x})^2$):

$$\begin{aligned}
\hat{\beta}_1 =& \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
=& \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
=& \frac{\beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
=& \frac{\beta_0 \cdot 0 + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
=& \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1 \mid X] =& \beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \,\middle|\, X \right] \\
=& \beta_1 + \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} \mathbb{E}[u_i \mid X] \\
=& \beta_1 \quad \text{(zero conditional mean and random sampling: } \mathbb{E}[u_i \mid X] = \mathbb{E}[u_i \mid x_i] = 0)
\end{aligned}$$

$$\therefore\ \mathbb{E}[\hat{\beta}_1] = \mathbb{E}\big[\mathbb{E}[\hat{\beta}_1 \mid X]\big] = \beta_1 \quad \text{(law of iterated expectations)}.$$

  For the intercept, averaging $y_i = \beta_0 + \beta_1 x_i + u_i$ over i gives $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$, so

$$\begin{aligned}
\hat{\beta}_0 =& \bar{y} - \hat{\beta}_1 \bar{x} \\
=& \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x} \\
=& \beta_0 + \bar{x}(\beta_1 - \hat{\beta}_1) + \bar{u} \\
\mathbb{E}[\hat{\beta}_0 \mid X] =& \beta_0 + \bar{x}\big(\beta_1 - \mathbb{E}[\hat{\beta}_1 \mid X]\big) + \mathbb{E}[\bar{u} \mid X] = \beta_0 \quad \therefore\ \mathbb{E}[\hat{\beta}_0] = \beta_0.
\end{aligned}$$

- **Sampling distribution** (Assumptions I–VI). Conditional on X,

$$\hat{\beta}_1 \sim N\left(\beta_1, \mathrm{Var}(\hat{\beta}_1 \mid X)\right), \quad \text{where} \quad \mathrm{Var}(\hat{\beta}_1 \mid X) = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

which implies

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1 \mid X)}} \;=\; \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \;\sim\; N(0,1).$$

"Proof": under Assumptions I–VI, given X, $\hat{\beta}_1$ is a linear combination of the independent normal errors $u_i \sim N(0, \sigma_u^2)$, so it is normal, with mean β_1 (shown above) and variance

$$\begin{aligned}
\hat{\beta}_1 =& \beta_1 + \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} u_i \\
\therefore\ \mathrm{Var}(\hat{\beta}_1 \mid X) =& \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} \mathrm{Var}(u_i \mid X) \\
=& \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} \sigma_u^2 = \frac{\sigma_u^2}{\sum_{j=1}^n (x_j - \bar{x})^2},
\end{aligned}$$

where the first step uses independence of the u_i (random sampling) and the second uses homoskedasticity.

- **Consistency**: $\mathrm{plim}_{n\to\infty} \hat{\beta}_0 = \beta_0$ and $\mathrm{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1$ (Assumptions I–IV, with finite $\mathrm{Var}(X) > 0$).

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) u_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} \\
\mathrm{plim}\, \hat{\beta}_1 &= \beta_1 + \frac{\mathrm{plim}\, \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) u_i}{\mathrm{plim}\, \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 + \frac{\mathrm{Cov}(X, u)}{\mathrm{Var}(X)} \quad \text{(LLN and continuous mapping)} \\
&= \beta_1 \quad (\mathrm{Cov}(X, u) = 0 \text{ by zero conditional mean, and } \mathrm{Var}(X) > 0) \\
\mathrm{plim}\, \hat{\beta}_0 &= \mathrm{plim}\, (\bar{y} - \hat{\beta}_1 \bar{x}) = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] = \beta_0.
\end{aligned}$$

- **Asymptotic normality**:

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \overset{\text{approx.}}{\sim} N(0,1),$$

where

$$\widehat{SE}(\hat{\beta}_1) = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad \text{and} \quad \hat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2 \xrightarrow{p} \sigma_u^2.$$

"Proof" (Assumptions I–V, with no normality assumption):

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} u_i \\
\sqrt{n}(\hat{\beta}_1 - \beta_1) &= \frac{\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) u_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} \xrightarrow{d} N\big(0,\ \sigma_u^2/\mathrm{Var}(X)\big),
\end{aligned}$$

since the numerator converges in distribution to $N(0, \sigma_u^2 \mathrm{Var}(X))$ (CLT; replacing $\bar{x}$ by $\mathbb{E}[X]$ in the numerator changes nothing in the limit), the denominator converges in probability to $\mathrm{Var}(X)$ (LLN), and Slutsky's theorem combines them. Finally,

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} = \sqrt{n}(\hat{\beta}_1 - \beta_1) \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}}{\hat{\sigma}_u} \xrightarrow{d} N(0,1),$$

because $\hat{\sigma}_u \xrightarrow{p} \sigma_u$ and $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \xrightarrow{p} \mathrm{Var}(X)$, so by Slutsky's theorem the limit is $N\big(0,\ \frac{\sigma_u^2}{\mathrm{Var}(X)} \cdot \frac{\mathrm{Var}(X)}{\sigma_u^2}\big) = N(0,1)$.

6 Reference: three limit theorems

- **The weak law of large numbers (LLN):** Suppose $\{x_i\}$ is an infinite sequence of i.i.d. random variables with finite expected value $\mathbb{E}[x_i] = \mu$. Then as n goes to infinity, the sample mean $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ converges in probability to μ:

$$\bar{x}_n \xrightarrow{p} \mu \text{ as } n \to \infty.$$

- **The Lindeberg–Lévy central limit theorem (CLT):** Suppose $\{x_i\}$ is an infinite sequence of i.i.d. random variables with finite expected value $\mathbb{E}[x_i] = \mu$ and finite variance $\mathrm{Var}(x_i) = \sigma^2$. Then as n goes to infinity, $\sqrt{n}(\bar{x}_n - \mu)$ converges in distribution to a normal distribution $N(0, \sigma^2)$:

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \text{ as } n \to \infty.$$

- **Slutsky's theorem:** Let $\{x_n\}$ and $\{y_n\}$ be two sequences of random variables (or matrices). If x_n converges in probability to a constant c and y_n converges in distribution to a random variable (or matrix) y, then $x_n y_n \xrightarrow{d} cy$ and $x_n + y_n \xrightarrow{d} c + y$.

7 OLS setup in matrix form

Let $\mathbf{X}$ be the $n \times (k+1)$ data matrix for k regressors (plus the constant), $\mathbf{y}$ the $n \times 1$ vector of the outcome, $\boldsymbol{\beta}$ the $(k+1) \times 1$ vector of true coefficients (the intercept and k slopes), and $\mathbf{u}$ the $n \times 1$ vector of errors:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}_{n\times(k+1)} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}$$

Write $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})$ for the i-th row of $\mathbf{X}$, a $1 \times (k+1)$ row vector. A linear model of n observations and k regressors can then be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

The OLS estimator is the $(k+1) \times 1$ vector of estimated coefficients

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}_{(k+1)\times 1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

Writing this down assumes no perfect collinearity, so that the symmetric $(k+1) \times (k+1)$ matrix $\mathbf{X}'\mathbf{X}$ is invertible. With the linearity assumption as well,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{u}.$$

8 Properties of the OLS estimator

8.1 Unbiasedness

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] &= \mathbb{E}[\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{u} \mid \mathbf{X}] \quad \text{(linearity and no perfect collinearity)} \\
&= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' \, \mathbb{E}[\mathbf{u} \mid \mathbf{X}] \\
&= \boldsymbol{\beta} \quad \text{(zero conditional mean and random sampling: } \mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}) \\
\therefore \; \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\big[\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}]\big] = \boldsymbol{\beta} \quad \text{(law of iterated expectations)}
\end{aligned}$$

So unbiasedness of the OLS estimator needs four assumptions:

1. linearity: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. random sampling of $\{(\mathbf{x}_i, y_i)\}$
3. no perfect collinearity in $\mathbf{X}$ (so $(\mathbf{X}'\mathbf{X})^{-1}$ exists)
4. zero conditional mean: $\mathbb{E}[u_i \mid \mathbf{x}_i] = 0$, which with random sampling gives $\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}$.

8.2 Consistency

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{u} \quad \text{(linearity and no perfect collinearity)} \\
&= \boldsymbol{\beta} + \Big(\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\Big)^{-1} \Big(\sum\nolimits_{i=1}^{n} \mathbf{x}_i' u_i\Big) \\
&= \boldsymbol{\beta} + \Big(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\Big)^{-1} \Big(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{x}_i' u_i\Big)
\end{aligned}$$

Applying the LLN to the two sample means (random sampling):

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i &\xrightarrow{p} \mathbb{E}[\mathbf{x}_i'\mathbf{x}_i] \equiv \mathbf{Q} \quad \text{(a positive definite } (k+1) \times (k+1) \text{ matrix } \mathbf{Q} \text{ exists)} \\
\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i' u_i &\xrightarrow{p} \mathbb{E}[\mathbf{x}_i' u_i] = \mathbf{0} \quad \text{(zero conditional mean)}
\end{aligned}$$

So $\operatorname{plim} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{Q}^{-1}\mathbf{0} = \boldsymbol{\beta}$. Consistency of the OLS estimator therefore needs five assumptions: (1) linearity, (2) random sampling, (3) no perfect collinearity, (4) zero conditional mean, and (5) the second moments of $\mathbf{x}_i$ exist, with $\mathbf{Q}$ positive definite.

Because $\mathbf{Q}$ is positive definite, it is nonsingular, so $\mathbf{Q}^{-1}$ exists. And $\mathbb{E}[\mathbf{x}_i' u_i] = \mathbb{E}\big[\mathbf{x}_i' \,\mathbb{E}[u_i \mid \mathbf{x}_i]\big] = \mathbf{0}$ by the law of iterated expectations.

8.3 Asymptotic normality

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \Big(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\Big)^{-1}\Big[\sqrt{n}\Big(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{x}_i' u_i\Big)\Big]$$

(linearity and no perfect collinearity)

The pieces we need:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_i'\mathbf{x}_i] &= \mathbf{Q} \\
\mathbb{E}[\mathbf{x}_i' u_i] &= \mathbf{0} \quad \text{(zero conditional mean)} \\
\operatorname{Var}(\mathbf{x}_i' u_i) &= \mathbb{E}[\mathbf{x}_i' u_i u_i \mathbf{x}_i] - \mathbb{E}[\mathbf{x}_i' u_i]\,\mathbb{E}[\mathbf{x}_i' u_i]' = \mathbb{E}_x\big[\mathbf{x}_i'\,\mathbb{E}[u_i^2 \mid \mathbf{x}_i]\,\mathbf{x}_i\big] \\
&= \sigma^2\,\mathbb{E}[\mathbf{x}_i'\mathbf{x}_i] \quad \text{(homoskedasticity: } \mathbb{E}[u_i^2 \mid \mathbf{x}_i] = \sigma^2\text{)} \\
&= \sigma^2\mathbf{Q}
\end{aligned}$$

Applying the LLN to the sample mean of $\mathbf{x}_i'\mathbf{x}_i$ (random sampling):

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i \xrightarrow{\;p\;} \mathbf{Q}$$

Applying the CLT to the sample mean of $\mathbf{x}_i' u_i$ (random sampling):

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i' u_i - \mathbf{0}\Big) \xrightarrow{\;d\;} N(\mathbf{0}, \sigma^2\mathbf{Q})$$

Finally, by Slutsky's theorem, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to $\mathbf{Q}^{-1}$ times a $N(\mathbf{0}, \sigma^2\mathbf{Q})$ vector, whose variance is $\mathbf{Q}^{-1}(\sigma^2\mathbf{Q})\mathbf{Q}^{-1} = \sigma^2\mathbf{Q}^{-1}$:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\;d\;} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}).$$

Asymptotic normality of the OLS estimator therefore needs six assumptions: (1) linearity, (2) random sampling, (3) no perfect collinearity, (4) zero conditional mean, (5) the second moments of $\mathbf{x}_i$ exist, with $\mathbf{Q}$ positive definite, and (6) homoskedasticity. Without homoskedasticity, $\hat{\boldsymbol{\beta}}$ is still asymptotically normal, but its variance takes the robust (sandwich) form $\mathbf{Q}^{-1}\mathbb{E}[u_i^2\mathbf{x}_i'\mathbf{x}_i]\mathbf{Q}^{-1}$.

9 The projection matrix and the residual maker

Definitions:

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$: vector of OLS estimated coefficients
- $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$: fitted values
- $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$: residuals
- $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$: projection matrix
- $\mathbf{M} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$: residual maker
- Hence $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ and $\hat{\mathbf{u}} = \mathbf{M}\mathbf{y}$

Properties ($\mathbf{O}$ is a matrix of zeros and $\mathbf{0}$ a vector of zeros):

1. $\mathbf{P}$ is symmetric, $\mathbf{P} = \mathbf{P}'$, and idempotent, $\mathbf{PP} = \mathbf{P}$

$$\mathbf{P}' = \mathbf{X}\big((\mathbf{X}'\mathbf{X})^{-1}\big)'\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}$$

$$\mathbf{PP} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}$$

2. $\mathbf{M}$ is symmetric and idempotent

$$\mathbf{M}' = \mathbf{I} - \mathbf{P}' = \mathbf{M}, \qquad \mathbf{MM} = (\mathbf{I}-\mathbf{P})(\mathbf{I}-\mathbf{P}) = \mathbf{I} - 2\mathbf{P} + \mathbf{PP} = \mathbf{I} - \mathbf{P} = \mathbf{M}$$

3. $\mathbf{MP} = \mathbf{O}$ and $\mathbf{PM} = \mathbf{O}$

$$\mathbf{MP} = (\mathbf{I}-\mathbf{P})\mathbf{P} = \mathbf{P} - \mathbf{PP} = \mathbf{P} - \mathbf{P} = \mathbf{O}$$

$$\mathbf{PM} = \mathbf{P}'\mathbf{M}' = (\mathbf{MP})' = \mathbf{O}' = \mathbf{O}$$

4. $\mathbf{PX} = \mathbf{X}$ and $\mathbf{MX} = \mathbf{O}$

$$\mathbf{PX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}$$

$$\mathbf{MX} = (\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{O}$$

5. $\mathbf{P}\hat{\mathbf{y}} = \hat{\mathbf{y}}$ and $\mathbf{P}\hat{\mathbf{u}} = \mathbf{0}$

$$\mathbf{P}\hat{\mathbf{y}} = \mathbf{PX}\hat{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$$

$$\mathbf{P}\hat{\mathbf{u}} = \mathbf{P}(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{P}\mathbf{y} - \mathbf{P}\hat{\mathbf{y}} = \hat{\mathbf{y}} - \hat{\mathbf{y}} = \mathbf{0}$$

6. $\mathbf{M}\hat{\mathbf{y}} = \mathbf{0}$ and $\mathbf{M}\hat{\mathbf{u}} = \hat{\mathbf{u}}$

$$\mathbf{M}\hat{\mathbf{y}} = (\mathbf{I}-\mathbf{P})\hat{\mathbf{y}} = \hat{\mathbf{y}} - \hat{\mathbf{y}} = \mathbf{0}$$

$$\mathbf{M}\hat{\mathbf{u}} = \mathbf{MM}\mathbf{y} = \mathbf{M}\mathbf{y} = \hat{\mathbf{u}}$$

7. $\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\mathbf{y}$

$$\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\mathbf{P}\mathbf{y} = (\mathbf{P}\mathbf{X})'\mathbf{y} = \mathbf{X}'\mathbf{y}$$

8. $\mathbf{y}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}}$

   The first two equalities hold by definition ($\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and a scalar equals its transpose), so we only need $\mathbf{y}'\hat{\mathbf{y}} = \hat{\mathbf{y}}'\hat{\mathbf{y}}$:

$$\mathbf{y}'\hat{\mathbf{y}} = (\hat{\mathbf{y}} + \hat{\mathbf{u}})'\hat{\mathbf{y}} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{u}}'\hat{\mathbf{y}} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{y}'\mathbf{M}'\mathbf{P}\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{y}'\mathbf{M}\mathbf{P}\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}}$$

9. $\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}}$

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{M}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{P})\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}}$$