1 OLS Estimator: Properties and Inference

We consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i. \tag{1}$$

We obtain the OLS estimator from

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

The first order conditions (FOC) are

$$\begin{aligned} -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0 \quad \left(\sum_{i=1}^{n} \widehat{u}_i = 0\right) \\ -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)x_i &= 0 \quad \left(\sum_{i=1}^{n} \widehat{u}_i x_i = 0\right). \end{aligned}$$

Then, the OLS estimator of β_1 is

$$\widehat{\beta}_1 = \sum_{i=1}^{n} w_i y_i \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

where $w_i = (x_i - \bar{x}) / \sum_{i=1}^{n} (x_i - \bar{x})^2$. Since w_i is a function of x_i, $\widehat{\beta}_1$ is a linear estimator (relationship between y_i and the estimator is linear).

$$\begin{aligned} \widehat{\beta}_1 &= \sum_{i=1}^{n} w_i (\beta_0 + \beta_1 x_i + u_i) \\ &= \beta_0 \sum_{i=1}^{n} w_i + \beta_1 \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} w_i u_i \\ &= \beta_1 + \sum_{i=1}^{n} w_i u_i \end{aligned} \tag{2}$$

because $\sum_{i=1}^{n} w_i = 0$ and $\sum_{i=1}^{n} w_i x_i = 1$ (you can easily show these).

- **Properties**: Under suitable classical linear model assumptions, the OLS estimator shows desirable properties such as unbiasedness, efficiency and consistency.

1. [Unbiasedness]

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_1|X) &= \beta_1 + \sum_{i=1}^{n} \mathbb{E}(w_i u_i|X) \\
&= \beta_1 + \sum_{i=1}^{n} w_i \mathbb{E}(u_i|X) \\
&= 0
\end{aligned}$$

Why do we have the second line? (because w_i is a function of x_i) For the last line, what do we need? (zero conditional mean assumption)

2.

$$\begin{aligned}
var(\widehat{\beta}_1|X) &= var(\sum_{i=1}^{n} w_i u_i|X) \\
&= \sum_{i=1}^{n} var\left(w_i u_i|X\right) + \sum_{i\neq j}^{n} cov\left(w_i u_i, w_j u_j|X\right) \\
&= \sum_{i=1}^{n} \left[\mathbb{E}\left(w_i^2 u_i^2|X\right) - \left\{\mathbb{E}\left(w_i u_i|X\right)\right\}^2\right] + \sum_{i\neq j}^{n} \left[\mathbb{E}\left(w_i w_j u_i u_j|X\right) - \mathbb{E}\left(w_i u_i|X\right)\mathbb{E}\left(w_j u_j|X\right)\right] \\
&= \sum_{i=1}^{n} w_i^2 \mathbb{E}\left(u_i^2|X\right) + \sum_{i\neq j}^{n} w_i w_j \mathbb{E}\left(u_i u_j|X\right) \\
&= \sigma^2 \sum_{i=1}^{n} w_i^2
\end{aligned}$$

Note that $\mathbb{E}\left(w_i u_i|X\right) = w_i \mathbb{E}\left(u_i|X\right) = 0$ if $\mathbb{E}\left(u_i|X\right) = 0$. We have the fourth line because w_i is a function of x_i. For the last line, what do we need? (homoskedasticity, no serial correlation). Since $\mathbb{E}\left(u_i^2|X\right) = var\left(u_i|X\right) + \left\{\mathbb{E}\left(u_i|X\right)\right\}^2$, $var\left(u_i|X\right) = \sigma^2$ for all $i = 1, 2, \cdots, n$. Since $\mathbb{E}\left(u_i u_j|X\right) = cov\left(u_i, u_j|X\right) - \mathbb{E}\left(u_i|X\right)\mathbb{E}\left(u_j|X\right)$, $cov\left(u_i, u_j|X\right) = 0$ for $i \neq j$. Since

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1},$$

$$var(\widehat{\beta}_1|X) = \sigma^2 \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{-1}$$

3. [Efficiency, Gauss-Markov Theorem] We let $\widetilde{\beta}_1 = \sum_{i=1}^{n} k_i y_i$ be an arbitrary linear and unbi-

ased estimator. Note that k_i is a function of x_i. Since

$$\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^{n} k_i \left(\beta_0 + \beta_1 x_i + u_i\right) \\
&= \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i + \sum_{i=1}^{n} k_i u_i,
\end{aligned}$$

we need $\sum_{i=1}^{n} k_i = 0$ and $\sum_{i=1}^{n} k_i x_i = 1$.

$$\begin{aligned}
var(\widetilde{\beta}_1|X) &= var(\beta_1 + \sum_{i=1}^{n} k_i u_i|X) \\
&= \sigma^2 \sum_{i=1}^{n} k_i^2.
\end{aligned}$$

We can show that

$$var(\widetilde{\beta}_1|X) - var(\widehat{\beta}_1|X) \geq 0.$$

This implies that the OLS estimator is BLUE (best linear unbiased estimator). Note that we still need $\mathbb{E}\left(u_i|X\right) = 0$, homoskedasticity and no serial correlation.

4. [Consistency]

$$\begin{aligned}
\widehat{\beta}_1 &= \beta_1 + \sum_{i=1}^{n} w_i u_i = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right) u_i}{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
&\rightarrow {}_p \beta_1 + \frac{\mathbb{E}(x_i u_i)}{Q_{XX}} \\
&\rightarrow {}_p \beta_1
\end{aligned}$$

The second line follows by the LLN (law of large numbers). More specifically,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 &\rightarrow {}_p Q_{XX} \\
\frac{1}{n}\sum_{i=1}^{n} x_i u_i &\rightarrow {}_p \mathbb{E}(x_i u_i) \\
\frac{1}{n}\sum_{i=1}^{n} \bar{x} u_i &\rightarrow {}_p \mathbb{E}(x_i)\mathbb{E}(u_i) = 0
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}(x_i u_i) &= \mathbb{E}\left[\mathbb{E}(x_i u_i|X)\right] \\
&= \mathbb{E}\left[x_i \mathbb{E}(u_i|X)\right]
\end{aligned}$$

by the law of iterated expectations, $\mathbb{E}(x_i u_i) = 0$ if $\mathbb{E}(u_i|X) = 0$. Note that $\mathbb{E}(x_i u_i) =$

$cov\left(x_i, u_i\right)$ and $Q_{XX} = \mathbb{E}\left[\left(x_i - \mathbb{E}\left(x_i\right)\right)^2\right] = var\left(x_i\right).$

- **Inference**: Under all classical linear model assumptions, the OLS estimator is normally distributed and the standard inference procedure is valid.

1. [Unbiased estimator of σ^2] Let
$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2.$$
Then,
$$\mathbb{E}(\widehat{\sigma^2}|X) = \sigma^2.$$
Proof is tedious. Note that we still need $\mathbb{E}\left(u_i|X\right) = 0$, homoskedasticity and no serial correlation.

2. [OLS estimator is normally distributed.] Given all classical linear model assumptions,
$$\widehat{\beta}_1|X \sim N\left(\beta_1, \sigma^2\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^{-1}\right).$$
Since $u_i|X \sim iid\ N\left(0, \sigma^2\right)$ (iid : independent and identically distributed),
$$\sum_{i=1}^{n} w_i u_i|X \sim N\left(0, \sigma^2\sum_{i=1}^{n} w_i^2\right).$$
Note that a linear combination of independent normal random variables is also normally distributed. Therefore,
$$\widehat{\beta}_1|X = \beta_1 + \sum_{i=1}^{n} w_i u_i|X \sim N\left(\beta_1, \sigma^2\sum_{i=1}^{n} w_i^2\right).$$
Note that we need the normality of the error term for the normality of the OLS estimator. If we standardize it, we have (conditional on X)
$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2\sum_{i=1}^{n} w_i^2}} \sim N\left(0, 1\right).$$

3. [Inference] We consider $H_0 : \beta_1 = b$ and $H_1 : \beta_1 \neq b$. Given Assumptions 1-5,
$$t_{\widehat{\beta}_1} = (\widehat{\beta}_1 - b)/\mathrm{se}(\widehat{\beta}_1) \sim t_{n-2}$$

under H_0. Here $\text{se}(\widehat{\beta}_1) = \sqrt{\widehat{\sigma^2} \sum_{i=1}^n w_i^2}$ (standard error of $\widehat{\beta}_1$). Under H_0,

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{\text{se}(\widehat{\beta}_1)} = \frac{\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 \sum_{i=1}^n w_i^2}}}{\frac{\sqrt{\widehat{\sigma^2} \sum_{i=1}^n w_i^2}}{\sqrt{\sigma^2 \sum_{i=1}^n w_i^2}}} = \frac{\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 \sum_{i=1}^n w_i^2}}}{\sqrt{\frac{1}{\sigma^2 (n-2)} \sum_{i=1}^n \hat{u}_i^2}} = \frac{A}{\sqrt{B/(n-2)}}.$$

We can show the following three results. 1) The numerator $A \sim N(0,1)$. 2) The denominator $B \sim \chi^2_{n-2}$. 3) A and B are independent. Consequently, the t-statistic $t_{\widehat{\beta}_1}$ is t_{n-2} distributed (the usual t-test is based on t_{n-2} distribution: the standard inference procedure). Note that we need the normality of $\widehat{\beta}_1$ for this and, therefore, the normality of the error term is required.

2 Asymptotic Distribution of OLS Estimator

- Consistency: Let $\widehat{\theta}_n$ be an estimator of θ based on a sample of size n. Then, $\widehat{\theta}_n$ is a *consistent* estimator of θ if, for every $\epsilon > 0$,

$$\mathbb{P}(|\widehat{\theta}_n - \theta| > \epsilon) \rightarrow 0$$

as $n \rightarrow \infty$. This means θ is the probability limit of $\widehat{\theta}_n$ ($\text{plim}(\widehat{\theta}_n) = \theta$).

$$\widehat{\theta} \rightarrow_p \theta$$

as $n \rightarrow \infty$. Consistency is a minimal requirement for an estimator.

- Law of Large Numbers (LLN): Let $Y_i \sim iid$ for $i = 1, 2, \cdots, n$. (iid : independent and identically distributed)

$$\frac{1}{n} \sum_{i=1}^n Y_i \rightarrow_p E(Y_i)$$

as $n \rightarrow \infty$.

- Central Limit Theorem (CLT): Let $Y_i \sim iid$ for $i = 1, 2, \cdots, n$ with $E(Y_i) = \mu$ and $var(Y_i) = \sigma^2$.

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mu) \rightarrow_d N(0, \sigma^2)$$

as $n \rightarrow \infty$.

- Continuous Mapping Theorem (CMT): Let f be a continuous function.

$$\text{If } \widehat{\theta}_n \rightarrow_p \theta,\ f(\widehat{\theta}_n) \rightarrow_p f(\theta)$$

as $n \rightarrow \infty$.

- Using the CLT and LLN, we can show that the asymptotic distribution of the OLS estimator is normal.

$$\begin{aligned}
\widehat{\beta}_1 &= \beta_1 + \sum_{i=1}^{n} w_i u_i = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,u_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
\sqrt{n}\left(\widehat{\beta}_1-\beta_1\right) &= \left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\bar{x})\,u_i \qquad (3)
\end{aligned}$$

We have

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \to_p \mathbb{E}\left[(x_i-\bar{x})^2\right] = Q_{XX} \qquad (4)$$

by the LLN. Note that

$$\begin{aligned}
\mathbb{E}\left[(x_i-\bar{x})\,u_i\right] &= \mathbb{E}\left[\mathbb{E}((x_i-\bar{x})\,u_i|X)\right] \\
&= \mathbb{E}\left[(x_i-\bar{x})\,\mathbb{E}(u_i|X)\right] = 0
\end{aligned}$$

and

$$\begin{aligned}
var\left[(x_i-\bar{x})\,u_i\right] &= \mathbb{E}\left[(x_i-\bar{x})^2\,u_i^2\right] = \mathbb{E}\left[\mathbb{E}((x_i-\bar{x})^2\,u_i^2|X)\right] \\
&= \mathbb{E}\left[(x_i-\bar{x})^2\,\mathbb{E}(u_i^2|X)\right] \\
&= \sigma^2 Q_{XX}
\end{aligned}$$

by the law of iterated expectations, $\mathbb{E}(u_i|X)=0$ and $var(u_i|X)=\sigma^2$. Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\bar{x})\,u_i \to_d N\left(0,\sigma^2 Q_{XX}\right) \qquad (5)$$

by the CLT. From (3), (4) and (5), we have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta}_1-\beta_1\right) &\to {}_d Q_{XX}^{-1}\; N\left(0,\sigma^2 Q_{XX}\right) \\
\sqrt{n}\left(\widehat{\beta}_1-\beta_1\right) &\to {}_d N\left(0,\sigma^2 Q_{XX}^{-1}\right).
\end{aligned}$$

3 IV Estimation

If x_i is correlated with u_i, it is said to be *endogenous*. If x_i is endogenous, the zero conditional mean assumption, $\mathbb{E}(u_i|X) = 0$, does not hold and, consequently, the OLS estimator becomes inconsistent. If x_i is endogenous, i.e., $cov\left(x_i, u_i\right) \neq 0$,

$$\begin{aligned}
\widehat{\beta}_1 &= \beta_1 + \sum_{i=1}^{n} w_i u_i = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)u_i}{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
&\to {}_p\beta_1 + \frac{\mathbb{E}(x_i u_i)}{Q_{XX}} = \beta_1 + \frac{cov\left(x_i, u_i\right)}{var\left(x_i\right)} \neq \beta_1.
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)u_i &\to {}_p\mathbb{E}\left[\left(x_i - \mathbb{E}\left(x_i\right)\right)\left(u_i - \mathbb{E}\left(u_i\right)\right)\right] = cov\left(x_i, u_i\right) \\
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 &\to {}_p\mathbb{E}\left[\left(x_i - \mathbb{E}\left(x_i\right)\right)^2\right] = var\left(x_i\right).
\end{aligned}$$

The IV estimator is from

$$\begin{aligned}
\textstyle\sum_{i=1}^{n}(y_i - \widehat{\beta}_0^{IV} - \widehat{\beta}_1^{IV} x_i) &= 0 \;\; \left(\textstyle\sum_{i=1}^{n} \widehat{u}_i = 0\right) \\
\textstyle\sum_{i=1}^{n}(y_i - \widehat{\beta}_0^{IV} - \widehat{\beta}_1^{IV} x_i)z_i &= 0 \;\; \left(\textstyle\sum_{i=1}^{n} \widehat{u}_i z_i = 0\right),
\end{aligned}$$

which gives

$$\begin{aligned}
\widehat{\beta}_1^{IV} &= \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} \\
&= \frac{\sum_{i=1}^{n}(z_i - \overline{z})y_i}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} = \frac{\sum_{i=1}^{n}(z_i - \overline{z})\left\{\beta_0 + \beta_1 x_i + u_i\right\}}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \overline{z})u_i}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})}.
\end{aligned}$$

The conditions for an IV are

$$cov\left(z_i, u_i\right) = 0 \quad \text{and} \quad cov\left(z_i, x_i\right) \neq 0. \tag{6}$$

As long as the adopted IV satisfies two conditions in (6), the IV estimator is consistent;

$$\begin{aligned}
\widehat{\beta}_1^{IV} &= \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \overline{z})u_i}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} \\
&\to {}_p\beta_1 + \frac{cov\left(z_i, u_i\right)}{cov\left(z_i, x_i\right)} = \beta_1.
\end{aligned}$$

This explicitly shows why two conditions are necessary for consistency of the estimator. Note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(z_i-\overline{z})u_i &\rightarrow {}_p\mathbb{E}\left[(z_i-\mathbb{E}(z_i))(u_i-\mathbb{E}(u_i))\right]=cov(z_i,u_i) \\
\frac{1}{n}\sum_{i=1}^{n}(z_i-\overline{z})(x_i-\overline{x}) &\rightarrow {}_p\mathbb{E}\left[(z_i-\mathbb{E}(z_i))(x_i-\mathbb{E}(x_i))\right]=cov(z_i,x_i).
\end{aligned}$$

Using LLN and CLT, we can obtain the asymptotic distribution of the IV estimator.

$$\begin{aligned}
\widehat{\beta}_1^{IV} &= \beta_1+\frac{\sum_{i=1}^{n}(z_i-\overline{z})u_i}{\sum_{i=1}^{n}(z_i-\overline{z})(x_i-\overline{x})} \\
\sqrt{n}\left(\widehat{\beta}_1^{IV}-\beta_1\right) &= \left[\frac{1}{n}\sum_{i=1}^{n}(z_i-\overline{z})(x_i-\overline{x})\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i-\overline{z})u_i
\end{aligned}$$

We have

$$\frac{1}{n}\sum_{i=1}^{n}(z_i-\overline{z})(x_i-\overline{x})\rightarrow_p cov(z_i,x_i)=Q_{ZX}$$

by the LLN. Note that

$$\mathbb{E}\left[(z_i-\overline{z})u_i\right]=cov(z_i,u_i)=0$$

and

$$\begin{aligned}
var\left[(z_i-\overline{z})u_i\right] &= \mathbb{E}\left[(z_i-\overline{z})^2u_i^2\right]=\mathbb{E}\left[\mathbb{E}((z_i-\overline{z})^2u_i^2|Z)\right] \\
&= \mathbb{E}\left[(z_i-\overline{z})^2\mathbb{E}(u_i^2|Z)\right] \\
&= \sigma^2 Q_{ZZ}
\end{aligned}$$

by the law of iterated expectations and $var(u_i|Z)=\sigma^2$. Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(z_i-\overline{z})u_i\rightarrow_d N\left(0,\sigma^2Q_{ZZ}\right)$$

by the CLT. We have

$$\begin{aligned}
\sqrt{n}\left(\widehat{\beta}_1^{IV}-\beta_1\right) &\rightarrow {}_dQ_{ZX}^{-1}\ N\left(0,\sigma^2Q_{ZZ}\right) \\
\sqrt{n}\left(\widehat{\beta}_1^{IV}-\beta_1\right) &\rightarrow {}_dN\left(0,\sigma^2Q_{ZX}^{-1}Q_{ZZ}Q_{XZ}^{-1}\right).
\end{aligned}$$