SOC 690S Machine Learning in Causal Inference
Week 1 Supplements

Asympototic OLS Inference

We are interested in the distribution of the sample analog of

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i]$$

$$\text{where } X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix} \in \mathbb{R}^{p \times 1} \text{ and } Y_i \text{ is a } \textit{scalar}$$

Suppose $[Y_i X_i']'$ is *independently and identically distributed* in a sample of size n. The OLS estimator is given by

$$\hat{\beta} = \left[\sum_i X_i X_i' \right]^{-1} \sum_i X_i Y_i$$

Given $Y_i = X_i'\beta + e_i$[1]

$$\hat{\beta} = \left[\sum_i X_i X_i' \right]^{-1} \sum_i X_i \left(X_i'\beta + e_i \right)$$

$$= \beta + \left[\sum_i X_i X_i' \right]^{-1} \sum_i X_i e_i$$

Under standard regularity $E||X_i||^2 < \infty$, $E\left[e_i^2 ||X_i||^2\right] < \infty$, and $E[X_i X_i']$ exists and is positive definite and invertible ($E[X_i X_i'] \succ 0$)

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1}\right)$$

To show this is the case, we express

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_i X_i' X_i \right)^{-1} \left(\frac{1}{n} \sum_i X_i e_i \right)$$

[1]Note here Y_i and X_i are from the sample. However, we use the *population parameter* β to express the relation, which retains the property that $E[X_i e_i] = 0$. This holds as e_i does not have a meaning on its own, but is a derivative of $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$ and $e_i = Y_i - X_i'\beta$.

Here $X_i e_i \in \mathbb{R}^{p\times 1}$. The variance of the sampling distribution of $X_i e_i$, when $n \to \infty$, according to the *Central Limit Theorem*, is given by[2]

$$\sqrt{n}\left(\frac{1}{n}\sum_i X_i e_i\right) \xrightarrow{d} \mathcal{N}(0, E[e_i^2 X_i X_i'])$$

To show this, note the variance of a vector is a variance-covariance matrix

$$\begin{aligned} V(X_i e_i) &= E[(X_i e_i - E[X_i e_i])(X_i e_i - E[X_i e_i])'] \\ &= E[X_i e_i (X_i e_i)'] - E[X_i e_i]E[X_i e_i]' \\ &= E[e_i^2 X_i X_i'] \end{aligned}$$

By *Law of Large Numbers*, we have

$$\frac{1}{n}\sum_i X_i' X_i \xrightarrow{p} E[X_i' X_i]$$

By *Slutsky's Theorem* and the *Continuous Mapping Theorem*[3]

$$\sqrt{n}\left(\frac{1}{n}\sum_i X_i' X_i\right)^{-1}\left(\frac{1}{n}\sum_i X_i e_i\right) \xrightarrow{d} \mathcal{N}\left(0, E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1}\right)$$

The consistent *sample* "sandwich" estimator (Eicker-Huber-White) is then given by

$$\hat{V}(\hat{\beta}) = (X_i' X_i)^{-1}\left(\sum_i^n X_i X_i' \hat{e}_i^2\right)(X_i' X_i)^{-1}$$

This is also known as heteroskedasticity-consistent standard errors. This is, however, not the standard error you get by default from packaged software. Default standard errors are derived under a homoskedasticity assumption $E[e_i^2 | X_i] = \sigma^2$. Given the assumption, we have the "meat"

$$E[e_i^2 X_i X_i'] = E[E[e_i^2 X_i X_i' | X_i]] = \sigma^2 E[X_i X_i']$$

Accordingly,

$$\begin{aligned} E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1} &= \sigma^2 E[X_i X_i']^{-1} E[X_i X_i'] E[X_i X_i']^{-1} \\ &= \sigma^2 E[X_i X_i']^{-1} \end{aligned}$$

[2]We have $\sqrt{n}$ to get a non-degenerate limit distribution. If we do not multiply by $\sqrt{n}$, the variance will converge to 0 at rate $\frac{1}{n}$ as $n \to \infty$.

[3]This expansion may not be obvious if you are not familiar with matrix operation. Suppose $Z \in \mathbb{R}^p$ is a *random* vector and $A \in \mathbb{R}^{p\times p}$ is a *fixed* matrix—the same setup as the two factors we have. By definition, $V(AZ) = E[(AZ - E[AZ])(AZ - E[AZ])'] = E(AZ - AE[Z])(AZ - AE[Z])'] = E[A(Z - E[Z])(Z - E[Z])'A'] = E[AV(Z)A']$.

Delta Method

The heteroskedasticity-consistent standard errors can also be derived from the Delta Method. To introduce the method, suppose we have an estimator $\hat{\theta}$ for a parameter θ. We often want the distribution of a function of the estimator $g(\hat{\theta})$.
Suppose the standard CLT holds,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Take a first-order Taylor expansion of g around θ,

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \approx g'(\theta)\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, (g'(\theta))^2\sigma^2)$$

We now expand the case to the multivariate matrix space. Suppose now $\hat{\theta} \in \mathbb{R}^p$, and the standard CLT holds

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Omega)$$

For a smooth function $g : \mathbb{R}^p \to \mathbb{R}$, with gradient $\nabla g(\theta) \in \mathbb{R}^p$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \nabla g(\theta)'\Omega\nabla g(\theta))$$

If $g : \mathbb{R}^p \to \mathbb{R}^k$, it generalizes to

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, G\Omega G')$$

where $G \in \mathbb{R}^{k \times p}$ is the Jacobian matrix of g at θ.
In the general OLS version, g maps $\mathbb{R}^p \to \mathbb{R}^p$

$$g(\beta) = \begin{bmatrix} E\left[X_{i1}\left(Y_i - \beta_1 X_{i1} - \cdots - \beta_p X_{ip}\right)\right] \\ E\left[X_{i2}\left(Y_i - \beta_1 X_{i1} - \cdots - \beta_p X_{ip}\right)\right] \\ \vdots \\ E\left[X_{ip}\left(Y_i - \beta_1 X_{i1} - \cdots - \beta_p X_{ip}\right)\right] \end{bmatrix}$$

Its Jacobian is

$$\begin{aligned} G(\beta) &= -\begin{bmatrix} E[X_{i1}^2] & E[X_{i1}X_{i2}] & \cdots & E[X_{i1}X_{ip}] \\ E[X_{i2}X_{i1}] & E[X_{i2}^2] & \cdots & E[X_{i2}X_{ip}] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_{ip}X_{i1}] & E[X_{ip}X_{i2}] & \cdots & E[X_{ip}^2] \end{bmatrix} \\ &= -E[X_i X_i'] \end{aligned}$$

For the sample moment (normal equation)

$$g_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta)$$

OLS solves $g_n(\hat{\beta}) = 0$. Expanding g_n at β gives

$$\begin{aligned}
0 = g_n(\beta) &\approx g_n(\beta) + (\hat{\beta} - \beta) \\
\hat{\beta} - \beta &\approx -G(\beta)^{-1} g_n(\beta) \\
\sqrt{n}(\hat{\beta} - \beta) &\approx -G(\beta)^{-1}\sqrt{n} g_n(\beta) \\
&\approx -G(\beta)^{-1}\sqrt{n}\frac{1}{n}\sum_i X_i e_i \\
&\xrightarrow{d} \mathcal{N}(0, G(\beta)^{-1} E[e_i^2 X_i X_i'] G(\beta)^{-1}) \\
&\xrightarrow{d} \mathcal{N}\left(0, E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1}\right)
\end{aligned}$$

What Fails When p/n is Not Small

Theorem 1.2.1 on page 19 of CML establishes the fact that the *prediction* error of OLS as measured by RMSE grows at rate $\sqrt{p/n}$. Indeed, when p/n is not small, the OLS *inference*—the standard errors and confidence intervals for $\hat{\beta}$—also becomes inconsistent. The last chapter of MHE discusses the issue in detail, and here I give the intuition. Remember the sample sandwich estimator

$$\hat{V}(\hat{\beta}) = (X'X)^{-1}\left(\sum_i^n X_i X_i' \hat{e}_i{}^2\right)(X'X)^{-1}$$

is consistent when $p/n \to 0$, as LLN states that

$$\frac{1}{n}\sum_i X_i' X_i \xrightarrow{p} E[X_i' X_i]$$

and CLT ensures

$$\sum_i^n X_i X_i' \hat{e}_i{}^2 \xrightarrow{d} E[e_i^2 X_i X_i']$$

However, when $p/n \to c > 0$, the *operator norm error*[4] no longer vanishes, but grows at

[4]The operator norm error measures the largest possible distortion of a quadratic form caused by replacing the population Gram matrix $E[X_i X_i']$ with the sample Gram $\frac{1}{n} X_i' X_i$.

rate of $\sqrt{p/n}$

$$\left\| \frac{1}{n} X_i' X_i - E[X_i X_i'] \right\|_{\text{op}} = \sup_{\|v\|_2=1} \left| v' \left(\frac{1}{n} X_i' X_i - E[X_i X_i'] \right) v \right| \sim O_p(\sqrt{\frac{p}{n}})$$

The intuition is that each entry of $\frac{1}{n} X_i' X_i$ still satisfies LLN; however there are $p \times p$ entries. Ensuring all of them to be consistent is much harder, and the LLN fails in operator norm.

In this case, using sandwich standard errors always give underestimated true standard errors (or true sampling variability of $\hat{\beta}$). This failure in statistical *inference* in high-dimensional data is another motivation of introducing Machine Learning in statistical modeling and causal inference.