Algorithms for Causal Inference (II)

José Jaén Delgado

In this paper several algorithms and statistical properties for Causal Inference are presented. Python implementations are also provided.

1 Introduction

Causal Inference is imperative for answering questions such as: is it really worth it studying at a private university in terms of salary projection? Does increasing the minimum wage lead to greater youth unemployment? Is rent control an effective policy for driving housing prices down? Does setting a ceiling price result in shortages?

Furthermore, Causal Inference goes beyond Machine Learning (ML) in the sense that interpretable model outputs are available without resorting to Global or Local interpretability techniques such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME). Additionally, precise curve-fitting is mostly not enough to derive causality. This is not to say that ML cannot be used for Causal Inference, in fact, major advancements have been made in the field, but AI is yet to reach the causal layer.

In this paper quantitative methods to estimate causal effects are presented as well as their statistical properties in the form of mathematical proofs. `Python` code can be found in the same GitHub repository, so that results can be reproduced. Different functions will be programmed as to make the interpretation of results easy for users with no statistical background. Also, Directed Acyclic Graphs (DAGs) will be used as to intuitively illustrate our aim with every model.

The paper is structured as follows: Part I covers Finite Sample Theory, Part II focuses on asymptotic properties of the OLS estimator (Large Sample Theory) and Part III combines asymptotic theory with robust methods to deal with endogeneity.

Fumio Hayashi's *Econometrics* has been a vital inspiration for this paper, as well as Judea Pearl's *The Book of Why*.

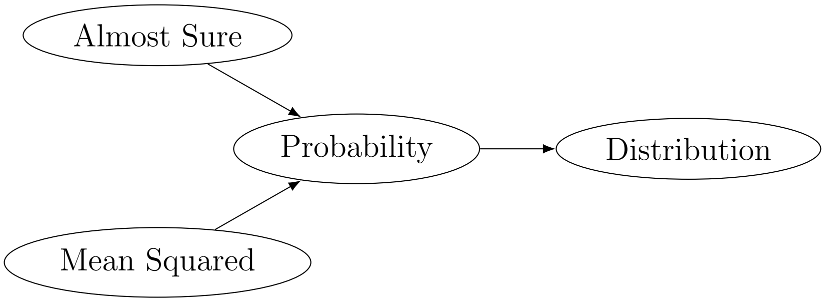
2 Large Sample Theory

An alternative approach to Finite Sample Theory is needed when the joint distribution of $\boldsymbol{\varepsilon}|\mathbf{X}$ is no longer normal. Likewise, when the number of observations in a dataset is sufficiently large, imposing restrictive assumptions on the sample ($\mathbf{y}$, $\mathbf{X}$) is not justified anymore. The stochastic process that generates ($\mathbf{y}$, $\mathbf{X}$) is far more important than the sample data itself for asymptotic theory. In fact, if the Data Generating Process (DGP) is specified, the joint distribution of the finite sample ($\mathbf{y}$, $\mathbf{X}$) can be derived. Thus, assumptions will be made upon $\{y_i, \mathbf{x}_i\}_{i=i}^{n}$ (the DGP).

2.1 Asymptotic Mathematical Tools

We briefly introduce vital mathematical theorems and lemmas that will be used in the following sections.

- Convergence hierarchy: Read as convergence i implies convergence j.



Where almost sure convergence is denoted as $\xrightarrow{a.s}$, mean squared convergence is represented by $\xrightarrow{m.s}$, convergence in probability is $\xrightarrow{p}$ and convergence in distribution is expressed as $\xrightarrow{d}$.

- Slutzky's Theorem: Behavior of the limiting value of random variables.

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x},\ \ \mathbf{y}_n \xrightarrow{p} \boldsymbol{\alpha} \implies \mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x} + \boldsymbol{\alpha}$$

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x},\ \ \mathbf{y}_n \xrightarrow{p} \mathbf{0} \implies \mathbf{y}_n'\mathbf{x}_n \xrightarrow{p} \mathbf{0}$$

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x},\ \ \mathbf{A}_n \xrightarrow{p} \mathbf{A} \implies \mathbf{A}_n\mathbf{x}_n \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$$

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x},\ \ \mathbf{A}_n \xrightarrow{p} \mathbf{A} \implies \mathbf{x}_n\mathbf{A}_n^{-1}\mathbf{x}_n \xrightarrow{d} \mathbf{x}'\mathbf{A}^{-1}\mathbf{x}$$

$$\mathbf{x}_n \xrightarrow{d} \mathbf{x},\ \ \mathbf{y}_n \xrightarrow{p} \mathbf{0} \implies \mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x}$$

2.2 Assumptions

- A.1) Linearity: $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$

This entails that the Linear Regression model to be estimated takes the form:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

- A.2) Ergodic Stationarity: $\{y_i, \mathbf{x}_i\}$ is jointly stationary and ergodic.

$\{y_i, \mathbf{x}_i\}$ is stationary if $\mathrm{E}[\mathbf{x}_i y_i]$ does not depend on i and $\mathrm{Cov}(\mathbf{x}_i y_i, \mathbf{x}_{i-j} y_{i-j}) < \infty$ depends only on j.

Given stationarity, a stochastic process is said to be ergodic it is asymptotically independent:

$$\begin{aligned}&\lim_{n\to\infty} |\mathrm{E}[f(\mathbf{x}_i y_i, \dots, \mathbf{x}_{i+k} y_{i+k}) g(\mathbf{x}_{i+n} y_{i+n}, \dots, \mathbf{x}_{i+n+l} y_{i+n+l})]| \\ &\qquad = |\mathrm{E}[f(\mathbf{x}_i y_i, \dots, \mathbf{x}_{i+k} y_{i+k})]|\,|\mathrm{E}[g(\mathbf{x}_{i+n} y_{i+n}, \dots, \mathbf{x}_{i+n+l} y_{i+n+l})]|\end{aligned}$$

So the joint distribution of $\{y_i, \mathbf{x}_i\}$ remains unchanged and such stochastic process is not too persistent.

- A.3) Exogeneity: $\mathrm{E}[x_{ik}\varepsilon_i] = 0 \ \forall i, k$

$$\begin{aligned}\mathrm{E}[\mathbf{g}_i] = \mathbf{0} \quad (\mathbf{g}_i &:= \mathbf{x}_i\varepsilon_i) \\ \mathrm{E}[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] &= \mathbf{0}\end{aligned}$$

Regressors are orthogonal to the contemporaneous error term.

- A.4) Rank Condition: $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] = \boldsymbol{\Sigma}_{xx} < \infty$

The $K \times K$ matrix $\boldsymbol{\Sigma}_{xx}$ is nonsingular (invertible).

- A.5) $\{\mathbf{g}_i\}$ is a Martingale Difference Sequence (m.d.s):

$$\mathrm{E}[\mathbf{g}_i|\mathbf{g}_{i-1}, \mathbf{g}_{i-2}, \dots, \mathbf{g}_1] = \mathbf{0} \ \ \forall i \geq 2$$

Furthermore:

$$\mathrm{E}[\mathbf{g}_i\mathbf{g}_i'] = \mathrm{E}[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'] < \infty$$

Note that for any given stochastic process, if $\phi(x) < \infty$, then $\{\phi(x)\}$ is said to exist and be finite.

Having presented the assumptions and the asymptotic tools, it is but natural to derive the properties of the OLS estimator:

- Consistency: $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$

$$\begin{aligned}
\mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \text{ (by minimizing } \mathbf{e}'\mathbf{e}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \text{ (by A.1)} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\
&= (n^{-1}\mathbf{X}'\mathbf{X})^{-1}n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \text{ (multiplying and dividing by} \, n)
\end{aligned}$$

Note that $\mathbf{S_{xx}} := n^{-1}(\mathbf{X}'\mathbf{X})$ which can be expressed as $\mathbf{S_{xx}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$

By A.2 $\{y_i, \mathbf{x}_i\}$ is ergodic stationary and thus so is $\{\mathbf{x}_i\mathbf{x}_i'\}$ as it is a function of $\{\mathbf{x}_i\}$

By A.4 $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] = \boldsymbol{\Sigma}_{\mathbf{xx}} < \infty$

Thus, by Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] \implies \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}$$

According to Continuous Mapping Theorem (CMT): $\mathrm{a}(z_n) \xrightarrow{p} \mathrm{a}(z)$ if $z_n \xrightarrow{p} z$ (provided $\mathrm{a}(\cdot)$ is a plausible continuous transformation).

Since $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is nonsingular and thus, invertible, by A.4 we can apply CMT:

$$\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$$

$n^{-1}\mathbf{X}'\mathbf{X}$ is also nonsingular by A.2 (columns would be linearly dependent by pure coincidence). Anyway, $n^{-1}\mathbf{S_{xx}}$ is invertible as the sample size increases since $\boldsymbol{\Sigma}_{\mathbf{xx}}$ is.

$\{\varepsilon_i\}$ is ergodic stationary since by A.1: $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, where $\{y_i, \mathbf{x}_i\}$ is ergodic stationary following A.2.

Also, $\mathrm{E}[\mathbf{x}_i\varepsilon_i] = \mathbf{0}$ by A.3 and $n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i$.

Thus, by Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon \xrightarrow{p} \mathrm{E}[\mathbf{x}_i\varepsilon_i] \;\; (= \mathbf{0})$$

By Slutzky's Theorem, the limiting behavior of the product of some elements is the product of the limits of said components (provided they exist and are finite). Thus:

$$\mathbf{S}_{\mathbf{xx}}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{p} \mathbf{0}$$

This proves $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$ and so the OLS estimator is consistent: $\mathbf{b} - \boldsymbol{\beta} = \mathbf{S}_{\mathbf{xx}}^{-1}\bar{\mathbf{g}} \xrightarrow{p} \mathbf{0}$

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\beta} \tag{1}$$

- Asymptotic Normality: $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1})$

Let us get back to the sampling error, but this time we multiply by $\sqrt{n}$:

$$\begin{aligned}\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) &= \sqrt{n}\mathbf{S}_{\mathbf{xx}}^{-1}\,\bar{\mathbf{g}} \quad (\text{since } \mathbf{g}_i := \mathbf{x}_i\varepsilon_i)\\ &= \mathbf{S}_{\mathbf{xx}}^{-1}\sqrt{n}\,\bar{\mathbf{g}}\end{aligned}$$

Note that $\mathrm{E}[\mathbf{g}_i] = \mathrm{E}[\bar{\mathbf{g}}]$ as $\{\mathbf{x}_i\varepsilon_i\}$ is ergodic stationary and $\mathrm{E}[\mathbf{g}_i] = \mathbf{0}$ by LIE (A.3).

By Ergodic LLN: $\bar{\mathbf{g}} \xrightarrow{p} \mathbf{0}$

By A.5: $\mathbf{S} = \mathrm{E}[\mathbf{g}_i\mathbf{g}_i']$ which is $\mathrm{Avar}(\bar{\mathbf{g}})$ since $\mathrm{E}[\bar{\mathbf{g}}] = \mathbf{0}$

Thus, by Ergodic Stationary Martingale Differences CLT (Ergodic CLT):

$$\sqrt{n}\,\bar{\mathbf{g}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S})$$

Applying Slutzky's Theorem:

$$\mathbf{S}_{\mathbf{xx}}^{-1}\sqrt{n}\,\bar{\mathbf{g}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1})$$

Note that $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1\prime} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$ since it is symmetric.

Consequently:

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}) \tag{2}$$

We have proved the OLS estimator $\mathbf{b}$ is consistent and asymptotically normal.

Let us now focus on consistently estimating $\mathrm{E}[\varepsilon_i^2]$, for which we proceed with the previously introduced estimator S^2. Our objective is then to prove:

$$S^2 := \frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \xrightarrow{p} \mathrm{E}[\varepsilon_i^2]$$

Notice that by multiplying and dividing by the sample size we can rewrite the expression above as:

$$S^2 = \frac{n}{n-k}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right)$$

Where applying properties of the limiting behavior of variables it is easy to see:

$$\lim_{n\to\infty}\frac{n}{n-k} = 1 \implies \operatorname*{plim}_{n\to\infty}\frac{n}{n-k} = 1$$

For the other part of the term note that A.1 (linearity) ensures that residuals can be expressed as:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i - \mathbf{x}_i'\mathbf{b})^2 \quad \text{(since by A.1 } y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta}))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - 2\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta}) + (\mathbf{b} - \boldsymbol{\beta})'\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'(\mathbf{b} - \boldsymbol{\beta})
\end{aligned}$$

Where $\mathbf{b}$ is the K-dimensional vector containing OLS estimators for each population parameter in our linear regression model. Note that A.1 - A.4 along with Ergodic LLN ensure that $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$, so clearly $\mathbf{b}$ - $\boldsymbol{\beta} \xrightarrow{p} 0$

- Since $\mathbf{b}$ - $\boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$, by Continuous Mapping Theorem (CMT): $(\mathbf{b}$ - $\boldsymbol{\beta})' \xrightarrow{p} \mathbf{0}$

  By A.4 (Rank condition) the $K \times K$ full-column rank matrix of moments $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i']$ exists and is finite.

  $\{\mathbf{x}_i\mathbf{x}_i'\}$ is ergodic stationary as it is a function of $\{\mathbf{x}_i\}$, which in turn is an ergodic stationary stochastic process by A.2 (Ergodic Stationary).

  Thus, by Ergodic LLN:

  $$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] < \infty$$

  Consequently, applying Slutzky's Theorem (ST), the last term vanishes, as all limits exist and are finite:

$$(\mathbf{b}-\boldsymbol{\beta})' \ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'(\mathbf{b}-\boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}$$

- Let $\mathbf{g}_i := \mathbf{x}_i\varepsilon_i$ since according to A.5 it is an m.d.s, it follows that $\mathrm{E}[\mathbf{g}_i] = \mathbf{0}$ by LIE.

  $\{\mathbf{g}_i\}$ is ergodic stationary as it is a function of $\{\varepsilon_i, \mathbf{x}_i\}$ which is jointly ergodic stationary by A.2 (ε_i satisfies this condition as it is a linear function of y_i by A.1).

  Thus, by Ergodic LLN:

  $$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i \xrightarrow{p} \mathbf{0}$$

  Consequently, by ST, the middle term also vanishes, as all limits exist and are finite:

  $$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{x}_i(\mathbf{b}-\boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}$$

- Lastly, leveraging the fact that an intercept is introduced ($x_{i1} = 1$, $\forall i$) as virtually all econometric applications, then the second moment of the error term exists and is finite.

  $\{\varepsilon_i\}$ is ergodic stationary as it is a function of $\{y_i\}$ by A.1, which in turn is ergodic stationary by A.2.

  Thus, by Ergodic LLN:

  $$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \xrightarrow{p} \mathrm{E}[\varepsilon_i^2]$$

So clearly, by ST:

$$\frac{n}{n-k}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right) \xrightarrow{p} \mathrm{E}[\varepsilon_i^2] \implies S^2 \xrightarrow{p} \mathrm{E}[\varepsilon_i^2] \tag{3}$$

These results mean that, although using the sample $(\mathbf{y}, \mathbf{X})$ might not yield an unbiased estimator, as the number of observations increases, we do end up with a $\mathbf{b}$ such that it converges to the true value $\boldsymbol{\beta}$. Likewise, it might be that the sampling error is not normally distributed, but only as $n \to \infty$ does the distribution converge to a Gaussian one. Similarly for S^2 and $\mathrm{E}[\varepsilon_i^2]$.

2.3 Hypothesis Testing

As previously shown, no restrictive assumptions on the joint distribution of $\{y_i, \mathbf{x}_i\}$ have to be imposed apart from covariance stationarity and ergodicity. Note that the first one can be forced by carrying out some feature engineering, namely: taking logarithms, calculating first differences, etc.

We now show how the test statistics from Finite Sample Theory converge to a known probability distribution under H_0.

- t-statistic for individual coefficients:

Since $\sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1})$, suppose there exists an estimator $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$ (we will prove it later). Consider the following hypothesis test:

$$\mathrm{H}_0\text{: } \beta_k = \bar{\beta}_k$$

Clearly, $\sqrt{n}(b_k - \bar{\beta}_k) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{Avar}(b_k))$. A natural estimator of the asymptotic variance of the k-th sampling error term is:

$$\widehat{\mathrm{Avar}}(b_k) = \left(\mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk}$$

Since $\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$ by A.2, A.3 & CMT, and we assumed $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$, by ST:

$$\left(\mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk} \xrightarrow{p} \left(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right)_{kk}$$

Thus:

$$t_k = \frac{\sqrt{n}(b_k - \bar{\beta}_k)}{\sqrt{\widehat{\mathrm{Avar}}(b_k)}} \xrightarrow{d} \mathcal{N}(0,1) \tag{4}$$

Note that an heteroskedasticity-consistent standard error can be derived since there is no restriction on homoskedasticity:

$$\mathrm{SE}^*(b_k) = \sqrt{\frac{1}{n}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk}} \tag{5}$$

Which in turn is equivalent to:

$$t_k = \frac{b_k - \bar{\beta}_k}{\sqrt{\frac{1}{n}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk}}} \xrightarrow{d} \mathcal{N}(0,1) \tag{6}$$

• Wald statistic for joint hypothesis testing:

Consider the following hypothesis test:

$$\mathrm{H}_0\colon\ \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Then from Finite Sample Theory:

$$W = n(\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R}\widehat{\mathrm{Avar}}(\mathbf{b})\mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r})$$

Note that by A.2 - A.5:

$$(\mathbf{Rb} - \mathbf{r}) \left[\mathbf{R}\widehat{\mathrm{Avar}}(\mathbf{b})\mathbf{R}\right]^{-1/2} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$\left[\mathbf{R}\widehat{\mathrm{Avar}}(\mathbf{b})\mathbf{R}'\right]^{-1}$ is no more than the squared expression for $\left[\mathbf{R}\widehat{\mathrm{Avar}}(\mathbf{b})\mathbf{R}\right]^{-1/2}$

Similarly, the squared expression of $\mathbf{Rb} - \mathbf{r}$ is $(\mathbf{Rb} - \mathbf{r})'(\mathbf{Rb} - \mathbf{r})$, sharing a common standard deviation.

Consequently:

$$W = n(\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R}\widehat{\mathrm{Avar}}(\mathbf{b})\mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r}) \xrightarrow{d} \chi^2(\#r) \tag{7}$$

So the Large Sample Theory test statistics follow a known distribution under H_0. Note that although the joint distribution of $(\mathbf{y}, \mathbf{X})$ might not coincide with the asymptotic distributions of t_k, W, when the sample size becomes arbitrarily large, it does.

For a sufficiently large dataset $\mathcal{D}$, namely, its rows or observations are of considerable amount, the preceding test statistics are preferred over their Finite Sample counterparts, as less restrictive assumptions are needed to hold.

Note, however, that $\mathrm{E}[\varepsilon_i \mathbf{x}_i] = \mathbf{0}$, that is, regressors should be orthogonal to their contemporaneous error term vector. In addition, $\{\mathbf{g}_i\}$ must be a m.d.s with finite second moments. Although we relaxed some assumptions, this model is not completely flexible.

2.4 Consistent estimation of S

Recall that $\mathbf{S} = \mathrm{E}[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i']$. For simplicity, let us assume $K = 1$, so x_i is a scalar.

$$\begin{aligned}
\varepsilon_i &= y_i - x_i\beta \\
\hat{\varepsilon_i} &= y_i - x_i\hat{\beta} \text{ (as the OLS estimator is consistent)} \\
\hat{\varepsilon_i^2} &= (y_i - x_i\beta)^2 = (x_i\beta + \varepsilon_i - x_i\hat{\beta})^2 = (\varepsilon_i - x_i(\hat{\beta} - \beta))^2 \\
&= \varepsilon_i^2 - 2x_i\varepsilon_i(\hat{\beta} - \beta) + x_i^2(\hat{\beta} - \beta)^2
\end{aligned}$$

Thus:

$$\hat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i^2 x_i^2 - 2x_i^3\varepsilon_i(\hat{\beta} - \beta) + x_i^4(\hat{\beta} - \beta)^2\right)$$

We introduce A.6: $\mathrm{E}[x_i^4] < \infty$, which can be interpreted as large outliers being unlikely.

$\{x_i^4\}$ is ergodic stationary by A.2 as it is a function of $\{x_i\}$

By Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^4 \xrightarrow{p} \mathrm{E}[x_i^4]$$

Thus, by Slutzky's Theorem:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^4(\hat{\beta} - \beta) \xrightarrow{p} 0$$

So the last term of the expression for $\hat{\mathbf{S}}$ above vanishes as the sample size increase. Let us now focus on the middle term, applying *Cauchy-Schwartz Inequality*:

$$\mathrm{E}[|f \cdot h|] \leq \sqrt{\mathrm{E}[f^2] \cdot \mathrm{E}[h^2]}$$

This can be proved leveraging trigonometric properties:

$$\cos(\theta) = \frac{\mathbf{x}'\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||}$$

Note that $|\cos(\theta)| \leq \ 1$ so:

$$\frac{|\mathbf{x}'\mathbf{y}|}{||\mathbf{x}||\,||\mathbf{y}||} \leq 1$$

$$|\mathbf{x}'\mathbf{y}| \leq ||\mathbf{x}||\,||\mathbf{y}||$$

Particularizing for our case:

$$\mathrm{E}[|x_i^3 \cdot \varepsilon_i|] \leq \sqrt{\mathrm{E}[x_i^2\varepsilon_i^2] \cdot \mathrm{E}[x_i^4]}$$

$\mathrm{E}[x_i^2\varepsilon_i^2] < \infty$ by A.5 and $\mathrm{E}[x_i^4] < \infty$ by A.6, thus, $\mathrm{E}[x_i^3\varepsilon_i]$ is bounded by some finite number, entailing that it exists and is finite.

Since $\{x_i^3\varepsilon_i\}$ is a function of $\{x_i\varepsilon_i\}$ and $\{x_i, \varepsilon_i\}$ is ergodic stationary by A.2, $\{x_i^3\varepsilon_i\}$ is an ergodic stationary stochastic process.

Thus, by Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^3\varepsilon_i \xrightarrow{p} \mathrm{E}[x_i^3\varepsilon_i] < \infty$$

By Slutzky's Theorem:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^3\varepsilon_i(\hat{\beta} - \beta) \xrightarrow{p} 0 \quad (\text{since} \quad \hat{\beta} \xrightarrow{p} \beta)$$

Consequently, both, the middle and last term of $\hat{\mathbf{S}}$ vanish. Let us analyze the left hand-side one.

Since $\{\varepsilon_i^2 x_i^2\}$ is ergodic stationary by A.2 and $\mathrm{E}[\varepsilon_i^2 x_i^2] < \infty$ by A.5.

By Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 x_i^2 \xrightarrow{p} \mathrm{E}[\varepsilon_i^2 x_i^2]$$

Thus:

$$\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$$

3 Conditional Homoskedasticity

Previously it was assumed in Finite Sample Theory that $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}_n$. If this is really the case in Large Sample Theory, we can derive the asymptotic distribution of the test statistics t and F that were explained in an earlier chapter.

A new assumption is presented:

- A.7) Conditional Homoskedasticity: $\mathrm{E}[\varepsilon_i|\mathbf{x}_i] = \sigma^2$ (> 0), $\forall i$

Note that this is a stronger assumption than simply imposing unconditional homoskedasticity.

The homoskedasticity-robust expressions derived in previous sections are now reduced to:

$$\begin{aligned}
\mathbf{S} &= \mathrm{E}[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i'] \\
&= \mathrm{E}[\mathrm{E}[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i']|\mathbf{X}] \quad \text{(by LIE)} \\
&= \mathrm{E}[\mathrm{E}[\varepsilon_i^2|\mathbf{X}] \mathbf{x}_i \mathbf{x}_i'] \quad \text{(by linearity of CEF)} \\
&= \mathrm{E}[\sigma^2 \mathbf{x}_i \mathbf{x}_i'] \quad \text{(by A.7)} \\
&= S^2 \mathbf{S}_{\mathbf{xx}} \quad \text{(unbiased estimation)} \\
\mathrm{Avar}(\mathbf{b}) &= \mathbf{\Sigma}_{xx}^{-1} \mathbf{S} \mathbf{\Sigma}_{xx}^{-1} \\
&= \mathbf{\Sigma}_{xx}^{-1} \mathrm{E}[\sigma^2 \mathbf{x}_i \mathbf{x}_i'] \mathbf{\Sigma}_{xx}^{-1} \\
\widehat{\mathrm{Avar}}(\mathbf{b}) &= S^2 \mathbf{S}_{\mathbf{xx}}^{-1} \\
&= n S^2 (\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Note that the t-statistic then becomes:

$$\begin{aligned}
t &= \frac{b_k - \bar{\beta}_k}{\sqrt{\frac{1}{n}\left(\mathbf{S}_{\mathbf{xx}}^{-1} \hat{\mathbf{S}} \mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk}}} \\
&= \frac{b_k - \bar{\beta}_k}{\sqrt{\frac{1}{n}\left(S^2 (\mathbf{X}'\mathbf{X})_{kk}^{-1}\right)}}
\end{aligned} \tag{8}$$

This is numerically identical to the t-statistic previously derived in Finite Sample Theory. Now let us focues on joint hypothesis testing.

Substituting $\widehat{\mathrm{Avar}}(\mathbf{b})$ into the Wald Statistic:

$$\begin{aligned}
\mathbf{W} &= n(\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R} n S^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r}) \\
&= (\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R} S^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r}) \\
&= (\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r})/S^2 \\
&= \#r \underbrace{\frac{\mathbf{w}/\#r}{q/(n-K)}}_{\mathbf{F}} = \frac{SSR_r - SSR_u}{S^2}
\end{aligned} \tag{9}$$

The Wald statistic $\mathbf{W}$ is also numerically identical to its Finite Sample Theory counterpart ($\#r\mathbf{F}$)

Lastly, let us present the population value of $\mathbf{S}$:

$$\begin{aligned}\mathbf{S} &= \mathrm{E}[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'] \\ &= \mathrm{E}[\mathrm{E}[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i']|\mathbf{X}] \quad \text{(by LIE)} \\ &= \mathrm{E}[\mathrm{E}[\varepsilon_i^2|\mathbf{X}]\mathbf{x}_i\mathbf{x}_i'] \quad \text{(by linearity of CEF)} \\ &= \mathrm{E}[\sigma^2\mathbf{x}_i\mathbf{x}_i'] \quad \text{(by A.7)} \\ &= \sigma^2\mathbf{\Sigma}_{\mathbf{xx}}\end{aligned}$$

Notice how A.6 is no longer needed (fourth-moment assumption) and Avar($\mathbf{b}$) takes the following form:

$$\mathrm{Avar}(\mathbf{b}) = \sigma^2\mathbf{\Sigma}_{\mathbf{xx}}^{-1}$$

Conditional Homoskedasticity has to be tested in some way as it is a restrictive assumption. Let us show how starting from the difference between robust $\hat{\mathbf{S}}$ and $S^2\mathbf{S}_{\mathbf{xx}}$:

$$H_0\text{: } \mathrm{E}[\epsilon_i|\mathbf{x}_i] = \sigma^2$$
$$H_1\text{: } \mathrm{E}[\epsilon_i|\mathbf{x}_i] \neq \sigma^2$$

$$\begin{aligned}\hat{\mathbf{S}} - S^2\mathbf{S}_{\mathbf{xx}} &= \frac{1}{n}\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{x}_i' - S^2\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \\ &= \frac{1}{n}\sum_{i=1}^{n}(e_i^2 - S^2)\mathbf{x}_i\mathbf{x}_i'\end{aligned}$$

Which should converge in probability to $\mathbf{0}$ if A.7 holds. Define $\boldsymbol{\psi}_i$ as the unique and stochastic elements of $\mathbf{x}_i\mathbf{x}_i'$ then:

$$n\,\mathbf{c}_n'\hat{\mathbf{B}}\mathbf{c}_n \xrightarrow{d} \chi^2(m)$$

Where $\mathbf{c}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}(e_i^2 - S^2)\boldsymbol{\psi}_i$, $\hat{\mathbf{B}}$ is the estimator for its asymptotic variance and m is the dimension of $\mathbf{c}_n$. Then, by regressing e_i^2 on a constant and $\boldsymbol{\psi}_i$ and calculating R^2, the test for conditional homoskedasticity boils down to:

$$n\,R^2 \xrightarrow{d} \chi^2(m)$$

4 Least Squares Projection

Let us prove the expression for the least squares projection of y on $\mathbf{x}$ when one of the regressors is a constant, i.e: $\widehat{\mathrm{E}^*}[y|\mathbf{x}] = \widehat{\mathrm{E}^*}[y|1, \mathbf{x}] = \mu + \boldsymbol{\gamma}'\tilde{\mathbf{x}}$

Let $\tilde{\mathbf{x}}$ be the vector of stochastic regressors:

$$\mathbf{x} = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}} \end{bmatrix}$$

From the expected values of Normal Equations we get $\mathrm{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} = \mathrm{E}[\mathbf{x}y]$. Consider an optimal estimator such that the forecast error y - $\mathbf{x}'\boldsymbol{\beta}^*$ is orthogonal to $\mathbf{x}$. Assuming $\mathrm{E}[\mathbf{x}\mathbf{x}']$ is nonsingular, then:

$$\boldsymbol{\beta}^* = (\mathrm{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathrm{E}[\mathbf{x}y]$$

$$\mathbf{x}\mathbf{x}' = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}} \end{bmatrix} \begin{bmatrix} 1 & \tilde{\mathbf{x}}' \end{bmatrix} = \begin{bmatrix} 1 & \tilde{\mathbf{x}}' \\ \tilde{\mathbf{x}} & \tilde{\mathbf{x}}\tilde{\mathbf{x}}' \end{bmatrix}$$

$$\mathbf{x}y = \begin{bmatrix} 1 \\ \tilde{\mathbf{x}} \end{bmatrix} y = \begin{bmatrix} y \\ \tilde{\mathbf{x}}y \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \boldsymbol{\gamma} \end{bmatrix}$$

Then, $\mathrm{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} = \mathrm{E}[\mathbf{x}y]$ can be rewritten as:

$$\mathrm{E}\begin{bmatrix} \mu + \boldsymbol{\gamma}\tilde{\mathbf{x}} \\ \tilde{\mathbf{x}}\mu + \tilde{\mathbf{x}}\tilde{\mathbf{x}}'\boldsymbol{\gamma} \end{bmatrix} = \mathrm{E}\begin{bmatrix} y \\ \tilde{\mathbf{x}}y \end{bmatrix}$$

Breaking down the matrix into two equations:

$$\mathrm{E}[\mu] + \mathrm{E}[\boldsymbol{\gamma}\tilde{\mathbf{x}}'] = \mathrm{E}[y] \rightarrow \mu + \boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}'] = \mathrm{E}[y]$$
$$\mathrm{E}[\tilde{\mathbf{x}}\mu] + \mathrm{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'\boldsymbol{\gamma}] = \mathrm{E}[\tilde{\mathbf{x}}y] \rightarrow \mu\mathrm{E}[\tilde{\mathbf{x}}] + \boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'] = \mathrm{E}[\tilde{\mathbf{x}}y]$$

From the first equation: $\mu = \mathrm{E}[y]$ - $\boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}']$

Plugging it into the second one:

$$\begin{aligned}(\mathrm{E}[y]-\boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}'])\mathrm{E}[\tilde{\mathbf{x}}]+\boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'] &= \mathrm{E}[\tilde{\mathbf{x}}y]\\ \mathrm{E}[\tilde{\mathbf{x}}]\mathrm{E}[y]-\boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}]\mathrm{E}[\tilde{\mathbf{x}}']+\boldsymbol{\gamma}\mathrm{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}'] &= \mathrm{E}[\tilde{\mathbf{x}}y]\\ \boldsymbol{\gamma}(\mathrm{E}[\tilde{\mathbf{x}}\tilde{\mathbf{x}}']-\mathrm{E}[\tilde{\mathbf{x}}]\mathrm{E}[\tilde{\mathbf{x}}']) &= \mathrm{E}[\tilde{\mathbf{x}}y]-\mathrm{E}[\tilde{\mathbf{x}}]\mathrm{E}[y]\\ \boldsymbol{\gamma} = \mathrm{V}[\tilde{\mathbf{x}}]^{-1}\mathrm{Cov}(\tilde{\mathbf{x}},y)\end{aligned}$$

Finally:

$$\begin{aligned}\mu &= \mathrm{E}[y]-\mathrm{V}[\tilde{\mathbf{x}}]^{-1}\mathrm{Cov}(\tilde{\mathbf{x}},y)\mathrm{E}[V[\tilde{\mathbf{x}}]^{-1}\mathrm{Cov}(\tilde{\mathbf{x}},y)']\\ &= \mathrm{E}[y]-\boldsymbol{\gamma}'\mathrm{E}[\tilde{\mathbf{x}}]\end{aligned}$$

5 Extra: Chebychev's Weak LLN

- $\lim_{n\to\infty}\mathrm{E}[\bar{z}_n]=\mu$ and $\lim_{n\to\infty}\mathrm{V}[\bar{z}_n]=0$
- $\{z_n\}$ is not necessarily i.i.d

Note that by the definition of convergence in mean square:

$$\lim_{n\to\infty}\mathrm{E}\left[(\bar{z}_n-\mu)^2\right]=0$$

Applying the add-and-subtract strategy:

$$\lim_{n\to\infty}\mathrm{E}\left[(\bar{z}_n-\mathrm{E}[\bar{z}_n]+\mathrm{E}[\bar{z}_n]-\mu)^2\right]=\lim_{n\to\infty}\mathrm{E}\left[((\bar{z}_n-\mathrm{E}[\bar{z}_n])+(\mathrm{E}[\bar{z}_n]-\mu))^2\right]$$
$$\hookrightarrow \quad \lim_{n\to\infty}\mathrm{E}\left[(\bar{z}_n-\mathrm{E}[\bar{z}_n])^2+2(\bar{z}_n-\mathrm{E}[\bar{z}_n])(\mathrm{E}[\bar{z}_n]-\mu)+(\mathrm{E}[\bar{z}_n]-\mu)^2\right]$$

By Slutzky's Theorem the limiting behavior of a product is the product of the limits (if they exist and are finite).

$$\lim_{n\to\infty}\mathrm{E}\left[\mathrm{E}[\bar{z}_n]-\mu\right]=\lim_{n\to\infty}\mathrm{E}[\bar{z}_n]-\mu=0 \qquad \left(\text{since } \lim_{n\to\infty}\mathrm{E}[\bar{z}_n]=\mu\right)$$
$$\lim_{n\to\infty}\mathrm{E}\left[\bar{z}_n-\mathrm{E}[\bar{z}_n]\right]=\lim_{n\to\infty}\mathrm{E}[\bar{z}_n]-\mathrm{E}[\bar{z}_n]=0$$

So the cross-product in the middle side of the equation vanishes by Slutzky's Theorem. Focusing on the LHS:

$$\lim_{n\to\infty}\mathrm{E}\left[(\bar{z}_n-\mathrm{E}[\bar{z}_n])^2\right]=\lim_{n\to\infty}\mathrm{V}[\bar{z}_n]=0$$

$$\lim_{n\to\infty} \mathrm{E}\left[\mathrm{E}[(\bar{z}_n] - \mu)^2\right] = \lim_{n\to\infty} \mathrm{E}[\bar{z}_n]^2 - 2\mathrm{E}[\bar{z}_n]\mu + \mu^2$$

Which cancels out as $\lim_{n\to\infty} \mathrm{E}[\bar{z}_n] = \mu$ and said value is a constant (so its limit is itself). Thus:

$$\lim_{n\to\infty} \mathrm{E}[\bar{z}_n]^2 - 2\mathrm{E}[\bar{z}_n]\mu + \mu^2 = \mu^2 - 2\mu^2 + \mu^2 = 0$$

And so we proved $\bar{z}_n \xrightarrow{m.s} \mu$ as $\lim_{n\to\infty} \mathrm{E}[(\bar{z}_n - \mu)^2] = 0$