Algorithms for Causal Inference (III)

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.

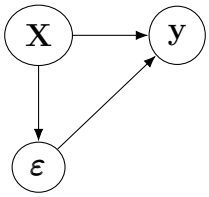
The paper is structured as follows: Part I covers Finite Sample Theory, Part II focuses on asymptotic properties of the OLS estimator (Large Sample Theory) and Part III combines asymptotic theory with robust methods to deal with endogeneity.

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

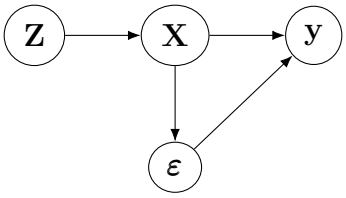
2 Dealing with Endogeneity

So far it has been assumed that regressors are exogenous, i.e $\mathrm{E}[\mathbf{x}_i\varepsilon_i] = \mathbf{0}$, $\forall i$. It is quite difficult that such restrictive assumption holds in many cases. For instance, think of microeconomic data stemming from surverys. Is it really realistic to suppose that households accurately answer income distribution questions? Is GDP perfectly quantified each year? Do we always include all relevant variables in our especifications? Let us be skeptical about it.

In those examples regressors are endogenous in the sense that some variation in $\mathbf{x}_i$ induces bias in sheer OLS estimation. As a consequence, causal effects cannot be correctly estimated. Take a look at the following DAG representing what we have just discussed.



We would either need to close the path $\mathbf{X} \to \boldsymbol{\varepsilon}$ or find a way to isolate $\mathbf{X} \to \mathbf{y}$ so we can disregard OLS bias. Since most of the time it is not up to econometricians to fix measurement errors in datasets or design the sampling strategy to obtain data, we are left with gleaning out the exogenous variation of $\mathbf{X}$ that explains $\mathbf{y}$. Such task can be attained with Instrumental Variables (IVs), a set of exogeneous features that affect $\mathbf{y}$ only through $\mathbf{X}$. Graphically:



Thus, by using the variation of $\mathbf{X}$ correlated with exogenous variables $\mathbf{Z}$ we can estimate the causal effect of $\mathbf{X}$ on $\mathbf{y}$ without needing to worry about endogeneity since $\mathrm{Cov}(\mathbf{Z}, \boldsymbol{\varepsilon}) = \mathbf{0}$. The endogenous variation of $\mathbf{X}$ is ditched out when being instrumented with $\mathbf{Z}$. Note that IVs are not always easy to collect in datasets and their validity mainly rely on Eonomic Theory rather than statistical logic, even though there exist Statistical Inference methods to test it.

Let us present a simple mathematical example to illustrate the IV intuition.

Suppose we seek to estimate the following Linear Regression Models:

$$q_i^D(p_i) = \alpha_0 + \alpha_1 p_i + u_i$$
$$q_i^S(p_i) = \beta_0 + \beta_1 p_i + v_i$$

Where $q_i^D(p_i)$ represents the demand function and $q_i^S(p_i)$ the supply function. Note that we seek to estimate the causal effect of p_i on both quantities. If we were to separately estimate both equations, note that no distinction can be made on the true effect of the OLS coefficient since the same variable is being used exactly in the same fashion in both specifications. Let us prove that p_i is endogenous.

In market equilibrium: $q_i^D = q_i^S$

$$\alpha_0 + \alpha_1 p_i + u_i = \beta_0 + \beta_1 p_i + v_i$$
$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}$$

Then the covariance between price and error term v_i is expressed as:

$$\text{cov}(p_i, v_i) = \text{cov}\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, v_i\right)$$

$$= \text{cov}\left(\frac{v_i - u_i}{\alpha_1 - \beta_1}, v_i\right) \quad \text{(since the first fraction is composed of parameters)}$$

$$= \frac{\text{V}[v_i] - \text{cov}(u_i, v_i)}{\alpha_1 - \beta_1} \quad (\neq 0)$$

Deriving the asymptotic expression of α_1:

$$\begin{aligned}\text{cov}(p_i, q_i) &= \text{cov}(p_i, \alpha_0 + \alpha_1 p_i + u_i) \\ &= \text{cov}(p_i, \alpha_1 p_i + u_i) \quad \text{(since } \alpha_0 \text{ is a constant)} \\ &= \alpha_1 \text{V}[p_i] - \text{V}[u_i] \quad \text{(by the definition of } p_i)\end{aligned}$$

The probability limit of the OLS estimator of price takes the form:

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\alpha_1 \text{V}[p_i] - \text{V}[u_i]}{\text{V}[p_i]} \\ &= \alpha_1 - \text{V}[u_i] \quad (\hat{\alpha}_1 \neq \alpha_1)\end{aligned}$$

So p_i is indeed endogenous and $\hat{\alpha}_1$ is biased. Let us fix this by defining a new model.

3 Model Assumptions

- A.1) Linearity: $\mathrm{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta},$ where $\mathbf{y} \in \Re^n, \mathbf{x}_i \in \Re^L$, $\boldsymbol{\beta} \in \Re^L$.

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

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

- A.2) Instrument Exogeneity: $\exists\, \mathbf{z}_i \;/\, \mathrm{E}[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$, where $\mathbf{z}_i \in \Re^K$

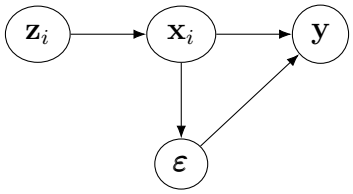
  Define $\mathbf{g}_i \coloneqq \mathbf{z}_i\varepsilon_i$ then $\mathrm{E}[\mathbf{g}_i] = \mathbf{0}$ and so $\mathrm{E}[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$

- A.3) Ergodic Stationarity: $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}$ is jointly ergodic stationary.

- A.4) Instrument Relevance: $\mathrm{E}[\mathbf{z}_i\mathbf{x}_i'] = \boldsymbol{\Sigma}_{\mathbf{zx}}$ $(\neq \mathbf{0})$

  The $K \times L$ matrix $\boldsymbol{\Sigma}_{\mathbf{zx}}$ is of full-column rank and finite.

- A.5) Exclusion Restriction: $\mathbf{z}_i$ affects $\mathbf{y}$ only through the instrumented $\mathbf{x}_i$.



As previously stated, this is a non-testable assumption for continuous variables. Economic Theory has to be resorted to.

- A.6) Asymptotic Normality: $\{\mathbf{g}_i\}$ is a Martingale Difference Sequence (m.d.s):

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

  Furthermore:

$$\mathrm{E}[\mathbf{g}_i\mathbf{g}_i'] = \underbrace{\mathrm{E}[\varepsilon_i^2\mathbf{z}_i\mathbf{z}_i']}_{\mathbf{S}} < \infty$$

Note that three key assumptions ultimately back the validity of $\mathbf{z}_i$: exogeneity (no correlation with the error term), relevance (correlation with $\mathbf{x}_i$) and the Exclusion Restriction. Unfortunately, these assumptions are difficult to test as previously stated, so apart from statistical inference, Economic Theory or domain knowledge must be included to any analysis involving IVs.

4 Method of Moments Estimation (MM)

Since OLS estimation procures biased results, endogeneity must be tackled, so we need to obtain $\mathrm{E}[\mathbf{x}_i\varepsilon_i] = \mathbf{0}$. Since there are some regressors that are correlated with the error term, these have to be instrumented with $\mathbf{z}_i$. Suppose that $K = L$, namely, that there exist as many instrumental variables as endogenous regressors. The instrumentation process is as follows:

$$\begin{aligned}
y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \text{(by A.1)} \\
\mathbf{z}_i y_i &= \mathbf{z}_i\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i\varepsilon_i \\
\mathrm{E}[\mathbf{z}_i y_i] &= \mathrm{E}[\mathbf{z}_i\mathbf{x}_i']\boldsymbol{\beta} + \mathrm{E}[\mathbf{z}_i\varepsilon_i] \\
\mathrm{E}[\mathbf{z}_i y_i] &= \mathrm{E}[\mathbf{z}_i\mathbf{x}_i']\boldsymbol{\beta} \quad \text{(by A.2)} \\
\boldsymbol{\Sigma}_{\mathbf{zy}} &= \boldsymbol{\Sigma}_{\mathbf{zx}}\boldsymbol{\beta} \quad \text{(by A.4)} \\
\boldsymbol{\beta} &= \boldsymbol{\Sigma}_{\mathbf{zx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{zy}}
\end{aligned}$$

By A.3 $\{\mathbf{z}_i, \mathbf{x}_i, y_i\}$ is jointly ergodic stationary, thus, $\{\mathbf{z}_i\mathbf{x}_i'\}$ is also an ergodic stationary stochastic process as it is a function of the former. So by Ergodic LLN:

$$\mathrm{E}[\mathbf{z}_i\mathbf{x}_i'] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'$$

By A.4, $\mathrm{E}[\mathbf{z}_i\mathbf{x}_i']$ is finite and nonsingular, so the sample analogue is also invertible for a sufficiently large number of observations. Then, by Continuous Mapping Theorem:

$$\boldsymbol{\Sigma}_{\mathbf{zx}}^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}$$

Consequently:

$$\mathbf{b} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i \tag{1}$$

Which may also be written as:

$$\mathbf{b} = \mathbf{S}_{\mathbf{zx}}^{-1}\mathbf{S}_{\mathbf{zy}} \tag{2}$$

$\mathrm{E}[\mathbf{g}_n(\tilde{\boldsymbol{\beta}})]$ has to be zero for some $\tilde{\boldsymbol{\beta}}$, which is none other than $\mathbf{b}$. Such weight is the IV estimator for which Method of Moments estimation was used.

Note that

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1} \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{zx}}^{-1}, \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{zy}}$$

Then by Slutzky's Theorem:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{zx}}^{-1}\mathbf{\Sigma}_{\mathbf{zy}} \tag{3}$$

Consequently, $\mathbf{b} \xrightarrow{p} \boldsymbol{\beta}$, and so $\mathbf{b}$ is consistent. Focusing on the asymptotic distribution of the IV estimator, let us start from the sampling error:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\boldsymbol{\beta} + \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{g}_n\end{aligned}$$

Then

$$\begin{aligned}\mathbf{b} &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{g}_n\right)\\ \mathbf{b}-\boldsymbol{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\mathbf{g}_n\\ \sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\sqrt{n}\mathbf{g}_n\end{aligned}$$

$\{\mathbf{z}_i\varepsilon_i\}$ is ergodic stationarity by A.2 as it is a function of $\{\mathbf{z}_i, \varepsilon_i\}$. Then, by Ergodic LLN: $\mathbf{g}_n \xrightarrow{p} \mathrm{E}[\mathbf{g}_i]$. Also, by A.6, $\mathbf{g}_i$ is an m.d.s so by Ergodic CLT: $\sqrt{n}\mathbf{g}_n \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S})$. Then, by Slutzky's Theorem:

$$\sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{zx}}^{-1}\mathbf{S}\mathbf{\Sigma}_{\mathbf{zx}}^{-1}) \tag{4}$$

So the IV estimator is consistent and asymptotically normal under A.1 - A.6.

5 General Method of Moments Estimation (GMM)

Consider the case where $K > L$, that is, when the equation is *overidentified*. We would then have available more instrumental variables and exogeneous features than endogeneous regressors. Notice how the matrix $\mathbf{\Sigma}_{\mathbf{zx}}^{-1}$ is not invertible anymore as its dimension is $K \times L$, and since we imposed K to be larger than L, this results in a nonsymetric matrix. A workaround has to be found to estimate $\boldsymbol{\beta}$. Recall what the objective of Method of Moments estimation truly is: to choose the parameter estimate $\mathbf{b}$ such that the sample moment $\mathbf{g}_n$ is zero. Although it is not feasible anymore to have $\mathbf{g}_n = \mathbf{0}$, we can make it close enough by defining a new objective function. Ne that we can define a distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ as $d(\mathbf{u}, \mathbf{v}) = (\mathbf{u} - \mathbf{v})'\mathbf{W}(\mathbf{u} - \mathbf{v})$, where $\mathbf{W}$ is a weighting matrix. Intuitively, we want the distance between $\mathrm{E}[\mathbf{g}_i]$ and $\mathbf{g}_n$ to be minimal, so the new objective function $J(\tilde{\boldsymbol{\beta}}, \mathbf{W})$ becomes:

$$\mathbf{b}(\hat{\mathbf{W}}) = \arg\min_{\tilde{\boldsymbol{\beta}}} \left\{ n \left(\mathbf{g}_n(\tilde{\boldsymbol{\beta}}) - \mathrm{E}[\mathbf{g}_i] \right)' \mathbf{W} \left(\mathbf{g}_n(\tilde{\boldsymbol{\beta}}) - \mathrm{E}[\mathbf{g}_i] \right) \right\} \tag{5}$$

Where $\tilde{\boldsymbol{\beta}}$ is a running parameter and $\mathbf{W}$ is a $K \times K$ Positive Definite weighting matrix. Although we will prove the underlying reasons in later sections, as of now ignore the sample size multiplying $J(\tilde{\boldsymbol{\beta}}, \mathbf{W})$ and suppose there exists an estimator for $\mathbf{W}$ such that $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$.

Since by A.6 $\{\mathbf{g}_i\}$ is an m.d.s then it follows that $\mathrm{E}[\mathbf{g}_i] = \mathbf{0}$ and so the optimization problem is reduced to:

$$\mathbf{b}(\hat{\mathbf{W}}) = \arg\min_{\tilde{\boldsymbol{\beta}}} \left\{ n \left(\mathbf{\Sigma_{zy}} - \mathbf{\Sigma_{zx}}\tilde{\boldsymbol{\beta}} \right)' \mathbf{W} \left(\mathbf{\Sigma_{zy}} - \mathbf{\Sigma_{zx}}\tilde{\boldsymbol{\beta}} \right) \right\} \tag{6}$$

Developing (6):

$$n \left(\mathbf{\Sigma}'_{\mathbf{zy}}\mathbf{W}\mathbf{\Sigma_{zy}} - \mathbf{\Sigma}'_{\mathbf{zy}}\mathbf{W}\mathbf{\Sigma_{zx}}\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}'\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zy}} + \tilde{\boldsymbol{\beta}}'\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zx}}\tilde{\boldsymbol{\beta}} \right)$$

Differentiating we are left with the following FOC:

$$\begin{aligned} n\,(-2\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zy}} + 2\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zx}}\tilde{\boldsymbol{\beta}} &= \mathbf{0} \\ n\,\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zx}}\tilde{\boldsymbol{\beta}} &= \mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zy}} \\ \boldsymbol{\beta} &= \left(\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zx}}\right)^{-1}\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma_{zy}} \end{aligned} \tag{7}$$

The sample analogue would be:

$$\mathbf{b}(\hat{\mathbf{W}}) = \left(\mathbf{S}'_{\mathbf{zx}}\hat{\mathbf{W}}\mathbf{S}_{\mathbf{zx}}\right)^{-1}\mathbf{S}'_{\mathbf{zx}}\hat{\mathbf{W}}\mathbf{S}_{\mathbf{zy}} \tag{8}$$

We already showed that $\mathbf{S}_{\mathbf{zx}} \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{zx}}$ so by CMT:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)' \xrightarrow{p} \mathbf{\Sigma}'_{\mathbf{zx}}$$

It was assumed that $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$. So by Slutzky's Theorem:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)'\hat{\mathbf{W}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right) \xrightarrow{p} \mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zx}}$$

Furthermore, $\dim(\mathbf{S}'_{\mathbf{zx}}\hat{\mathbf{W}}\mathbf{S}_{\mathbf{zx}}) = L \times L$. Note that $\mathbf{W}$ is a PD matrix and so is $\hat{\mathbf{W}}$ as the sample size increases. Both, $\mathbf{S}'_{\mathbf{zx}}$ and $\mathbf{S}_{\mathbf{zx}}$ are full-column rank matrices as n becomes larger by A.4, consequently by CMT:

$$\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)'\hat{\mathbf{W}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)\right]^{-1} \xrightarrow{p} (\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zx}})^{-1}$$

$\{\mathbf{z}_i y_i\}$ is jointly ergodic stationary since it is a function of $\{\mathbf{z}_i, y_i\}$ which by A.3 is an ergodic stationary stochastic process. Then by Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{zy}}$$

Thus, by Slutzky's Theorem:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)'\hat{\mathbf{W}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i\right) \xrightarrow{p} \mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zy}}$$

Finally, combining all results we get that by Slutzky's Theorem:

$$\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)'\hat{\mathbf{W}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)\right]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}'_i\right)'\hat{\mathbf{W}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i y_i\right) \xrightarrow{p} (\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zx}})^{-1}\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zy}}$$

This implies that $\left(\mathbf{S}'_{\mathbf{zx}}\hat{\mathbf{W}}\mathbf{S}_{\mathbf{zx}}\right)^{-1}\mathbf{S}'_{\mathbf{zx}}\hat{\mathbf{W}}\mathbf{S}_{\mathbf{zy}} \xrightarrow{p} (\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zx}})^{-1}\mathbf{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{\Sigma}_{\mathbf{zy}}$, so the GMM estimator $\mathbf{b}(\hat{\mathbf{W}}) \xrightarrow{p}$ is consistent under A.1 - A.5, namely $\mathbf{b}(\hat{\mathbf{W}}) \xrightarrow{p} \boldsymbol{\beta}$.

For the asymptotic distribution of $\mathbf{b}(\hat{\mathbf{W}})$ we need to derive the sampling error. Instrumenting $\mathbf{x}_i$ with $\mathbf{z}_i$ we obtain:

$$\mathbf{S_{zy}} = \mathbf{S_{zx}}\boldsymbol{\beta} + \mathbf{g}_n \tag{9}$$

Substituting (9) into (8):

$$\mathbf{b}(\hat{\mathbf{W}}) - \boldsymbol{\beta} = \left(\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{S_{zx}}\right)^{-1}\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{g}_n \tag{10}$$

Then

$$\begin{aligned}\sqrt{n}\left(\mathbf{b}(\hat{\mathbf{W}}) - \boldsymbol{\beta}\right) &= \sqrt{n}\left[\left(\left(\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{S_{zx}}\right)^{-1}\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{g}_n\right)\right] \\ &= \left(\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{S_{zx}}\right)^{-1}\mathbf{S'_{zx}}\hat{\mathbf{W}}\sqrt{n}\mathbf{g}_n\end{aligned}$$

By A.3 $\{\mathbf{z}_i\varepsilon_i\}$ is jointly ergodic stationary as it is a function of $\{\mathbf{z}_i, \varepsilon_i\}$. Also, by A.6 $\mathrm{E}[\mathbf{g}_i] = \mathbf{0}$. So by Ergodic LLN:

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

By A.6 $\{g_i\}$ is an m.d.s with $\mathrm{E}[\mathbf{g}_i\mathbf{g}_i'] = \mathbf{S}$ so by Ergodic CLT:

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

By Slutzky's Theorem we get:

$$\sqrt{n}\left(\mathbf{b}(\hat{\mathbf{W}}) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathrm{Avar}\left(\mathbf{b}(\hat{\mathbf{W}})\right)\right) \tag{11}$$

Applying variance properties:

$$\mathrm{Avar}\left(\boldsymbol{\beta}(\mathbf{W})\right) = \left(\boldsymbol{\Sigma}'_{\mathbf{zx}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{zx}}\right)^{-1}\boldsymbol{\Sigma}'_{\mathbf{zx}}\mathbf{W}\mathbf{S}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{zx}}\left(\boldsymbol{\Sigma}'_{\mathbf{zx}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{zx}}\right)^{-1} \tag{12}$$

The sample analogue is:

$$\widehat{\mathrm{Avar}}\left(\mathbf{b}(\hat{\mathbf{W}})\right) = \left(\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{S_{zx}}\right)^{-1}\mathbf{S'_{zx}}\hat{\mathbf{W}}\hat{\mathbf{S}}\hat{\mathbf{W}}\mathbf{S_{zx}}\left(\mathbf{S'_{zx}}\hat{\mathbf{W}}\mathbf{S_{zx}}\right)^{-1} \tag{13}$$

For a correct choice of $(\hat{\mathbf{S}}, \hat{\mathbf{W}})$, by Slutzky's Theorem we get:

$$\widehat{\text{Avar}}\left(\mathbf{b}(\hat{\mathbf{W}})\right) \xrightarrow{p} \text{Avar}(\boldsymbol{\beta}(\mathbf{W}))$$

Consequently the GMM estimator $\mathbf{b}(\hat{\mathbf{W}})$ is asymptotically Gaussian:

$$\sqrt{n}\left(\mathbf{b}(\hat{\mathbf{W}}) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(\boldsymbol{\Sigma}_{\mathbf{zx}}'\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{zx}}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{zx}}'\mathbf{W}\mathbf{S}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{zx}}\left(\boldsymbol{\Sigma}_{\mathbf{zx}}'\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{zx}}\right)^{-1}\right) \tag{14}$$

6 Consistent Estimation of the error variance

For proving the consistency of S^2, it is needed to define a new assumption:

• A.7) Finite Second Moment: $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] = \boldsymbol{\Sigma}_{\mathbf{xx}} < \infty$

Then from the very definition of S^2

$$\begin{aligned}
S^2 &= \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon_i}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\mathbf{b}(\hat{\mathbf{W}})\right)'\left(y_i - \mathbf{x}_i'\mathbf{b}(\hat{\mathbf{W}})\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\mathbf{b}(\hat{\mathbf{W}}) + \varepsilon_i\right)'\left(\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\mathbf{b}(\hat{\mathbf{W}}) + \varepsilon_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right) + \varepsilon_i\right)'\left(\mathbf{x}_i'\left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right) + \varepsilon_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right)'\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right) + 2\left(\boldsymbol{\beta} - \mathbf{b}(\hat{W})\right)'\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i
\end{aligned}$$

For all cases, note that since the GMM estimator is consistent then $\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}}) \xrightarrow{p} \mathbf{0}$. In the middle term, note that by A.3 $\{\mathbf{x}_i\mathbf{x}_i'\}$ is jointly ergodic stationary since it is a function of $\{\mathbf{x}_i\}$. Also, by A.7 $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] < \infty$. By Ergodic LLN:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}(< \infty)$$

So the middle term vanishes as by Slutzky's Theorem and CMT:

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

For the last term, consider the 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:

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

Particularizing for our case:

$$\mathrm{E}[|\mathbf{x}_i\varepsilon_i|] \leq \sqrt{\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'] \cdot \mathrm{E}[\varepsilon_i^2]}$$

$\mathrm{E}[\mathbf{x}_i\mathbf{x}_i']$ is finite by A.7. For $\mathrm{E}[\varepsilon_i^2]$, since we introduce a constant within $(\mathbf{x}_i, \mathbf{z}_i)$ and by A.6 $\mathrm{E}[\mathbf{g}_i\mathbf{g}_i'] < \infty$, then $\mathrm{E}[\varepsilon_i^2]$ is also finite. Consequently, by the Cauchy-Schwartz inequality: $\mathrm{E}[\mathbf{x}_i\varepsilon_i] < \infty$. Thus, by Slutzky's Theorem and CMT:

$$2\left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right)' \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \xrightarrow{p} \mathbf{0}$$

Both the middle and the RHS terms have vanished. We are only left with the LHS expression. Note that by A.3 $\{\varepsilon_i^2\}$ is ergodic stationary since it is a function of $\{y_i, \mathbf{x}_i\}$. Also, we showed that $\mathrm{E}[\varepsilon_i^2]$ is finite, so by Ergodic LLN:

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

Consequently, $S^2 \xrightarrow{p} \mathrm{E}[\varepsilon_i^2]$.

7 Consistent estimation of S

So far all results were dependent on $\hat{\mathbf{S}}$, which has to converge in probability to the true moment $\mathbf{S}$. Recall the definition of the asymptotic variance of $\mathbf{g}_i$:

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

A natural estimator is:

$$\hat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \mathbf{z}_i \mathbf{z}_i'$$

Since $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\mathbf{b}(\hat{\mathbf{W}})$ and $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ by A.1:

$$\begin{aligned}
\hat{\mathbf{S}} &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{x}_i' \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}}) + \varepsilon_i\right)\right)' \mathbf{x}_i' \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}}) + \varepsilon_i\right) \mathbf{z}_i \mathbf{z}_i' \\
&= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i' + \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right)' \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \mathbf{z}_i \mathbf{z}_i' \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right) \\
&+ 2\left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right)' \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{x}_i \mathbf{z}_i \mathbf{z}_i' \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right)
\end{aligned}$$

Note that in the middle term, by A.3 $\{\mathbf{x}_i\mathbf{x}_i'\mathbf{z}_i\mathbf{z}_i'\}$ is jointly ergodic stationary as it is a funtion of $\{\mathbf{z}_i, \mathbf{x}_i\}$. We have also shown that $\mathbf{b}(\hat{\mathbf{W}})$ is a consistent estimator.

• A.8) Finite Fourth Moments: $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'\mathbf{z}_i\mathbf{z}_i'] < \infty$

Thus, by Ergodic LLN, Slutzky's Theorem and CMT:

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

For the last term, note that by Cauchy-Schwartz Inequality:

$$\mathrm{E}[|\varepsilon_i \mathbf{x}_i \mathbf{z}_i \mathbf{z}_i'|] \leq \sqrt{\mathrm{E}[\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i'] \cdot \mathrm{E}[\mathbf{x}_i \mathbf{x}_i' \mathbf{z}_i \mathbf{z}_i']}$$

$\mathrm{E}[\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i']$ is finite by A.6 and $\mathrm{E}[\mathbf{x}_i\mathbf{x}_i'\mathbf{z}_i\mathbf{z}_i']$ is also finite by A.8. Thus, by Slutzky's Theorem and CMT:

$$2\left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right)' \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \mathbf{x}_i \mathbf{z}_i \mathbf{z}_i' \left(\boldsymbol{\beta} - \mathbf{b}(\hat{\mathbf{W}})\right) \xrightarrow{p} \mathbf{0}$$

$\{\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i'\}$ is jointly ergodic stationary by A.3 as it is a function of $\{\varepsilon_i, \mathbf{z}_i\}$. Also, by A.6 $\mathrm{E}[\mathbf{g}_i\mathbf{g}_i'] < \infty$ so applying Ergodic LLN:

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

Consequently, under A.1 - A.8 it is found that $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$