Algorithms for Causal Inference

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: Section II covers Finite Sample Theory, Section III focuses on asymptotic properties of the OLS estimator (Large Sample Theory) and Section IV 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 Finite Sample Theory

2.1 Notation

Before proceeding with Causal Inference methods for finite samples, it is convenient to introduce the notation to be used.

Let $\mathbf{y} \in \Re^n$ denote the target vector o dependent variable vector and n be the total number of observations. Let $\mathbf{x}_i \in \Re^K$ represent the feature vectors or regressor vectors (where virtually in all applications $x_{i1} = 1 \, \forall i$). If stacked together, feature vectors $\mathbf{x}_i$ become feature matrix or data matrix $\mathbf{X} \in \Re^{n \times K}$.

The goal of applying Causal Inference is to derive causal effects of all K features contained in $\{\mathbf{x}_i\}_{i=1}^n$ on the corresponding regressand y_i. Such task will be carried out by estimating the following model:

$$\mathbf{y} = \mathrm{E}[\mathbf{y}|\mathbf{X}] + \boldsymbol{\varepsilon} \tag{1}$$

Where $\mathrm{E}[\mathbf{y}|\mathbf{X}]$ is the Conditional Expectation Function (CEF) and $\boldsymbol{\varepsilon} \in \Re^n$ is the vector of error terms (unobservable to the econometrician). Clearly, an estimation method for the CEF has to be proposed. Let us focus on the cases where $y_i \in \Re \, \forall i$, so that the dependent variable is continuous, that is, we concentrate solely on regression problems, rather than classification ones (where y_i takes discrete values). In such cases, the preferred mapping function of ($\mathbf{y}$, $\mathbf{X}$) is a linear one. The linear mapping between the target variable and regressors is facilitated by a vector of weights or vector of parameters $\boldsymbol{\beta} \in \Re^K$. Graphically, using a DAG:

$$\mathbf{X} \xrightarrow{\boldsymbol{\beta}} \mathbf{y}$$

In future DAGs, $\boldsymbol{\beta}$ will be taken for granted. It is important to notice that $\mathbf{X}$ is an $n \times K$ matrix and $\mathbf{y}$ is an n-dimensional vector, so actually there are K lines pointing from each column in $\mathbf{X}$ to $\mathbf{y}$.

Since we are considering a Finite Sample perspective, restrictive assumptions on the sample ($\mathbf{y}$, $\mathbf{X}$) have to be imposed. By finite sample, we mean a small-medium sized dataset $\mathcal{D}$. Sometimes, especially years ago, it is not possible to work with Big Data, ultimately working with smaller datasets. This is the case for certain microeconometric data like household surveys.

In the next page we present a set of assumptions on ($\mathbf{y}$,$\mathbf{X}$) that defines our model.

2.2 Assumptions

- A.1) Linearity: $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- A.2) Exogeneity: $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$

By the Law of Iterated Expectations (LIE) we can further derive:

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}]] &= \mathrm{E}[\boldsymbol{\varepsilon}] \quad (= \mathbf{0}) \\
\mathrm{E}[\varepsilon_i x_{jk}] &= \mathrm{E}[\mathrm{E}[\varepsilon_i x_{jk}|x_{jk}]] \\
&= \mathrm{E}[\mathrm{E}[\varepsilon_i|x_{jk}]x_{jk}] \quad \text{(by linearity of CEF)} \\
&= 0 \\
\mathrm{cov}(\varepsilon_i, x_{jk}) &= \mathrm{E}[\varepsilon_i x_{jk}] - \mathrm{E}[\varepsilon_i]\mathrm{E}[x_{jk}] \\
&= \mathrm{E}[\varepsilon_i x_{jk}] \quad (\text{since } \mathrm{E}[\varepsilon_i] = 0) \\
&= 0 \quad (\text{since we showed } \mathrm{E}[\varepsilon_i x_{jk}] = 0)
\end{aligned}$$

Meaning that the regressors are orthogonal to the error term (regressors are exogeneous).

- A.3) Rank Condition: $\Pr(\mathrm{rank}(\mathbf{X}) = K) = 1$

So $\mathbf{X}$ is of full-column rank, i.e, it does not contain any linear combination of $\mathbf{x}_i$. Since the rank of any matrix $\mathbf{A}$ is defined as $\mathrm{rank}(\mathbf{A}) = \min\{n, K\}$, this means that $n \geq K$. An Algorithm to overcome the failure of A.3 will be proposed at the end of the Finite Sample Theory section.

- A.4) Homoskedasticity: $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}_K$

In terms of the variance: $\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] - \underbrace{\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}]}_{\mathbf{0}} = \sigma^2\mathbf{I}_K$

Error terms are also uncorrelated ($i \neq j$):

$$\begin{aligned}
\mathrm{E}[\varepsilon_i\varepsilon_j|\mathbf{X}] &= \mathrm{E}[\varepsilon_j \mathrm{E}[\varepsilon_i|\mathbf{X}, \varepsilon_j]|\mathbf{X}] \\
&= \mathrm{E}[\varepsilon_j|\mathbf{X}]\mathrm{E}[\mathrm{E}[\varepsilon_i|\mathbf{X}, \varepsilon_j]|\mathbf{X}] \\
&= 0
\end{aligned}$$

So why are these assumptions restrictive? It well may be that $\mathrm{E}[\mathbf{y}|\mathbf{X}]$ is nonlinear, like in the classification tasks described above where $y_i \in \{0, 1\}$. Furthermore, $\mathbf{x}_i$ is normally not exogeneous, even in random samples, i.e: $\mathrm{cov}(\mathbf{x}_i, \varepsilon_i) \neq 0$.

Feature matrix $\mathbf{X}$ might be sparse or $K >> n$, yielding a singular cross-product matrix $\mathbf{X}'\mathbf{X}$. In this last case, ML algorithms such as Gradient Descent easily overcome failure of A.3, as well as computing the Moore–Penrose inverse (virtually what every statistical software does in practice, such as `scikit-learn`). We have yet to impose two additional assumptions on $(\mathbf{y}, \mathbf{X})$, but these will be introduced in following subsections for the sake of clarity.

Note how the Linear Regression model for Causal Inference is pretty rigid for finite samples.

2.3 Estimation Algorithm

We can proceed to estimate $\boldsymbol{\beta}$. Firstly, it is necessary to define an objective or cost function $J(\boldsymbol{\theta})$, composed of the sum of loss functions $L(\boldsymbol{\theta}_i)$ so $J(\boldsymbol{\theta}) = \sum_i L(\boldsymbol{\theta}_i)$. Since we need to resort to computer software to carry out computations, an efficient way to code the estimation algorithm has to be proposed. This is where **vectorization** comes to play. Rather than calculating $L(\boldsymbol{\theta}_i)$ for each observation and then summing it all together, we can operate with vectors and matrices to directly optimize $J(\boldsymbol{\theta})$. The mathematical operations needed for that are listed some paragraphs below.

A reasonable $J(\boldsymbol{\theta})$ needs to be adequate to the regression problem we are facing. Let predictions be denoted as $\hat{\mathbf{y}} \in \mathfrak{R}^n$, where $\hat{y}_i \coloneqq x_i'\mathbf{b}$ and $\mathbf{b}$ is the OLS estimator. Penalizing prediction errors denoted by $(y_i - \hat{y}_i)$ is but a natural way to go. Minimizing the squared distance of such expression (residuals) presents desirable properties to be expounded in the next pages. Formally, we seek to minimize:

$$J(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}\right)' \left(y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}\right) \tag{2}$$

$J(\tilde{\boldsymbol{\beta}})$ is referred to as the Sum of Squared Residuals (SSR). The computationally efficient (vectorized) version can be expressed in terms of matrices:

$$J(\tilde{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \tag{3}$$

Where $\tilde{\boldsymbol{\beta}}$ denotes a running parameter (not yet the final weights vector).

The final weights or coefficients vector is defined as:

$$\mathbf{b} = \arg\min_{\tilde{\boldsymbol{\beta}}} \left\{ (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \right\} \tag{4}$$

To derive the optimal solution for the Least Squares problem, we need to take into account the following Linear Algebra properties:

$$\text{Property I: } \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}' \qquad \text{Property II: } \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

Note that:

$$\begin{aligned}
(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) &= (\mathbf{y}' - \tilde{\boldsymbol{\beta}}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}
\end{aligned}$$

Thus

$$\frac{\partial S\tilde{S}R(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the derivative to $\mathbf{0}$:

$$\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{5}$$

By A.3, since $\mathbf{X}$ is of full-column rank, $\mathbf{X}'\mathbf{X}$ is Positive Definite and thus invertible (nonsingular). We obtain the following closed-form solution:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{6}$$

Plugging $\mathbf{b}$ into (V) and defining the residual vector $\mathbf{e} := \mathbf{y} - \mathbf{X}\mathbf{b}$:

$$\begin{aligned}
\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) &= \mathbf{0} \\
\mathbf{X}'\mathbf{e} &= \mathbf{0}
\end{aligned} \tag{7}$$

The last identity is known as the Normal Equations and will be crucial for following proofs. It is the sample manifestation of the orthogonality of the feature matrix with respect to $\boldsymbol{\varepsilon}$, namely: $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$.

In the next subsection finite sample properties for $\mathbf{b}$ are derived.

2.4 Finite Sample Properties of the OLS estimator

- 1) Unbiasedness: $\mathrm{E}[\mathbf{b}|\mathbf{X}] = \boldsymbol{\beta}$

$$
\begin{aligned}
\mathrm{E}[\mathbf{b}|\mathbf{X}] &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \quad \text{(minimizing SSR)} \\
&= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\mathbf{X}] \quad \text{(by A.1)} \\
&= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_K}\,\mathrm{E}[\boldsymbol{\beta}|\mathbf{X}] + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \quad \text{(by linearity of CEF)} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \quad \text{(since } \boldsymbol{\beta} \text{ is a parameter)} \\
&= \boldsymbol{\beta} \quad \text{(by A.2: } \mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0})
\end{aligned}
$$

Note that A.3 ensures $\exists(\mathbf{X}'\mathbf{X})^{-1}$ and a frequentist approach is taken since no prior distribution is imposed upon $\boldsymbol{\beta}$, ultimately lacking a probability distribution. Bayesian Inference procures a posterior distribution over the weight vector $\boldsymbol{\beta}$ through Maximum A Posteriori (MAP) estimation, allowing to quantify uncertainty in model outputs. We will include a Bayesian framework in the future.

- 2) Gauss-Markov Theorem (Efficiency): $\nexists\,\hat{\boldsymbol{\beta}}$ s.t $\mathrm{V}[\mathbf{b}|\mathbf{X}] \geq \mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]$

Firstly, let us derive the conditional variance of $\mathbf{b}$ on $\mathbf{X}$:

$$
\begin{aligned}
\mathrm{V}[\mathbf{b}|\mathbf{X}] &= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \\
&= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\mathbf{X}] \\
&= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_K}\,\mathrm{V}[\boldsymbol{\beta}|\mathbf{X}]\,\underbrace{\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{I}_K} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(since } \boldsymbol{\beta} \text{ is a parameter} \implies \mathrm{V}[\boldsymbol{\beta}|\mathbf{X}] = \mathbf{0}) \\
&= \sigma^2\,\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_K}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by A.4 : } \mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}] = \sigma^2\mathbf{I}_K) \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

The Gauss-Markov Theorem states that $\mathbf{b}$ is efficient (lowest variance) in the class of linear unbiased estimators. Indeed, as we will prove, there does not exist any unbiased linear estimator $\hat{\boldsymbol{\beta}}$ such that it exhibits a lower variance than the OLS estimator $\mathbf{b}$ under the proposed assumptions for Finite Sample Theory (A.1 - A.4).

Not surprinsingly, it is not difficult to find a linear estimator $\hat{\boldsymbol{\beta}}$ such that it performs better than $\mathbf{b}$ in predictive terms, mainly because A.1 - A.4 are unlikely to hold. Additionally, $\mathbf{b}$ tends to overfit the sample or training data $\mathcal{D}$.

Let $\hat{\boldsymbol{\beta}}$ be an unbiased estimator for $\boldsymbol{\beta}$ and linear in $\mathbf{y}$ such that:

$$\hat{\beta} = \mathbf{C}\mathbf{y}$$

Where $\mathbf{C}$ is a function of $\mathbf{X}$. Let $\mathbf{D} = \mathbf{C}$ - $\mathbf{A}$, where $\mathbf{A} \coloneqq (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
Then $\hat{\boldsymbol{\beta}} = (\mathbf{D} + \mathbf{A})\mathbf{y}$

$$\begin{aligned}
\mathrm{E}[\hat{\beta}|\mathbf{X}] &= \mathrm{E}[(\mathbf{D} + \mathbf{A})\mathbf{y}|\mathbf{X}] \quad \text{(by definition of } \mathbf{C}\text{)} \\
&= \mathrm{E}[\mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y}|\mathbf{X}] \\
&= \mathrm{E}[\mathbf{D}(\mathbf{X}\beta + \varepsilon) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon)|\mathbf{X}] \\
&= \mathrm{E}[\mathbf{D}\mathbf{X}\beta + \mathbf{D}\varepsilon + \beta + \mathbf{A}\varepsilon|\mathbf{X}] \\
&= \mathbf{D}\mathbf{X}\beta + \beta + \mathbf{D}\mathrm{E}[\varepsilon|\mathbf{X}] + \mathbf{A}\mathrm{E}[\varepsilon|\mathbf{X}] \\
&= \mathbf{D}\mathbf{X}\beta + \beta
\end{aligned}$$

As $\hat{\boldsymbol{\beta}}$ is unbiased by construction: $\mathrm{E}[\hat{\beta}|\mathbf{X}] = \beta$

$$\begin{aligned}
\mathrm{E}[\hat{\beta}|\mathbf{X}] &= \mathbf{D}\mathbf{X}\beta + \beta \\
\beta &= \mathbf{D}\mathbf{X}\beta + \beta \\
\mathbf{D}\mathbf{X}\beta = \mathbf{0} &\implies \mathbf{D}\mathbf{X} = \mathbf{0}
\end{aligned}$$

Deriving the sampling error:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{D} + \mathbf{A})\mathbf{y} \\
&= \mathbf{D}\mathbf{X}\beta + \mathbf{D}\varepsilon + \mathbf{A}\mathbf{y} \\
&= \mathbf{D}\varepsilon + \mathbf{A}(\mathbf{X}\beta + \varepsilon) \quad \text{(since } \mathbf{D}\mathbf{X} = \mathbf{0}\text{)} \\
&= \mathbf{D}\varepsilon + \beta + \mathbf{A}\varepsilon \quad \text{(since } \mathbf{A}\mathbf{y} = \beta + \mathbf{A}\varepsilon\text{)} \\
\hat{\beta} - \beta &= (\mathbf{D} + \mathbf{A})\varepsilon
\end{aligned}$$

Then:

$$\begin{aligned}
\mathrm{V}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] &= \mathrm{V}[(\mathbf{D} + \mathbf{A})\mathbf{y}|\mathbf{X}] \\
\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \mathrm{V}[(\mathbf{D} + \mathbf{A})\mathbf{y}|\mathbf{X}] \quad \text{(since } \boldsymbol{\beta} \text{ is a parameter)} \\
&= (\mathbf{D} + \mathbf{A})\mathrm{V}[\varepsilon|\mathbf{X}](\mathbf{D} + \mathbf{A})' \\
&= \sigma^2[(\mathbf{D} + \mathbf{A})(\mathbf{D} + \mathbf{A})'] \quad \text{(by A.4)} \\
&= \sigma^2[\mathbf{D}\mathbf{D}' + \mathbf{A}\mathbf{A}'] \\
&= \sigma^2\mathbf{D}\mathbf{D}' + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

The step that was skipped is reproduced below:

$$\begin{aligned}\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \sigma^2[(\mathbf{D}+\mathbf{A})(\mathbf{D}+\mathbf{A})'] \\ &= \sigma^2[\mathbf{DD}'+\mathbf{DA}'+\mathbf{AD}'+\mathbf{AA}']\end{aligned}$$

Note that

$$\begin{aligned}\mathbf{DA}'+\mathbf{AD}' &= 2\mathbf{DA}' \quad (\text{since } (\mathbf{AB})' = \mathbf{B}'\mathbf{A}') \\ &= 2\mathbf{DX}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{0} \quad (\text{since } \mathbf{DX} = \mathbf{0})\end{aligned}$$

Clearly:

$$\sigma^2[\mathbf{DD}'+\mathbf{DA}'+\mathbf{AD}'+\mathbf{AA}'] = \sigma^2[\mathbf{DD}'+\mathbf{AA}']$$

Consequently:

$$\mathrm{V}[\hat{\boldsymbol{\beta}}] = \sigma^2\mathbf{DD}' + \underbrace{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}}_{\mathrm{V}[\mathbf{b}|\mathbf{X}]}$$

Since **DD**' is Positive Semidefinite by construction, it entails $\mathbf{DD} \geq \mathbf{0}$. Thus:

$$\begin{aligned}\sigma^2\mathbf{DD}' + \sigma^2(\mathbf{X}'\mathbf{X})^{-1} &\geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ \mathrm{V}[\hat{\boldsymbol{\beta}}] &\geq \mathrm{V}[\mathbf{b}|\mathbf{X}]\end{aligned}$$

Indeed, the OLS estimator **b** exhibits the lowest variance within the class of linear unbiased estimators. We have just proved the Gauss-Markov Theorem.

- 3) Unbiased estimator for σ^2: Notice how σ^2 has been present throughout the entirety of the Gauss-Markov Theorem proof. It is actually of little practicality to deal with a parameter in our expressions, as we want to estimate a Linear Regression Model. Since it is not observable, we have to estimate σ^2 as well, for which we propose S^2, defined as

$$S^2 = \frac{SSR}{n-K}$$
$$S^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

We introduce two special matrices that are crucial for the following proof:

$$\begin{aligned}
\mathbf{P} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad \text{(projection matrix)} \\
\mathbf{M} &= (\mathbf{I}_n - \mathbf{P}) \quad \text{(annihilator matrix)}
\end{aligned}$$

Note that both matrices satisfy some desirable properties: symmetry and idempotence. Below is the proof:

$$\begin{aligned}
\mathbf{P}' &= (\mathbf{X}(\mathbf{X}'\mathbf{X}^{-1})\mathbf{X})' \\
&= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{P}} \quad \text{(since } \ (\mathbf{AB})' = \mathbf{B}'\mathbf{A}') \\
\mathbf{P}^2 &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\underbrace{\mathbf{X}'][\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{I}_k}\mathbf{X}'] \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
\mathbf{M}' &= (\mathbf{I}_n - \mathbf{P})' \\
&= \mathbf{I}_n' - \mathbf{P}' = \mathbf{I}_n - \mathbf{P} \quad \text{(since } \mathbf{P} \text{ and } \mathbf{I}_n \text{ are symmetric)} \\
\mathbf{M}^2 &= (\mathbf{I}_n - \mathbf{P})'(\mathbf{I}_n - \mathbf{P}) \\
&= \mathbf{I}_n - \mathbf{P} - \mathbf{P} + \mathbf{P}^2 \\
&= \mathbf{I}_n - \mathbf{P} \quad \text{(since } \mathbf{P} \text{ is idempotent and } \mathbf{AI} = \mathbf{A})
\end{aligned}$$

The vector of residuals can be expressed in terms of $\mathbf{M}$ and $\mathbf{y}$:

$$\begin{aligned}
\mathbf{e} &= \mathbf{y} - \mathbf{Xb} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underbrace{(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})}_{\mathbf{y}} \\
&= \mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\
&= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\
&= (\mathbf{I}_n - \mathbf{P})\boldsymbol{\varepsilon} \quad (= \mathbf{M}\boldsymbol{\varepsilon})
\end{aligned}$$

Note that SSR or $\mathbf{e}'\mathbf{e}$ can be written as:

$$\begin{aligned}
SSR &= \mathbf{e}'\mathbf{e} \\
&= (\mathbf{M}\boldsymbol{\varepsilon})'\mathbf{M}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \quad \text{(since } \mathbf{M} \text{ is idempotent)}
\end{aligned}$$

Consequently:

$$\begin{aligned}
\mathrm{E}[S^2|\mathbf{X}] &= \mathrm{E}\left[\frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{n-k}\middle|\mathbf{X}\right] \\
&= \frac{1}{n-k}\mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j\middle|\mathbf{X}\right] \\
&= \frac{1}{n-k}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\mathrm{E}\left[\varepsilon_i\varepsilon_j|\mathbf{X}\right] \\
&= \frac{1}{n-k}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ii}\sigma^2 \quad \text{(by A.4)} \\
&= \frac{\sigma^2}{n-k}\mathrm{trace}(\mathbf{M})
\end{aligned}$$

Only for $i = j$ does $\mathrm{E}[\varepsilon_i\varepsilon_j] = \sigma^2$ then $m_{ij} \neq 0$ in the same case. Consequently it becomes m_{ii} (or m_{jj}) which includes only the diagonal elements of the matrix $\mathbf{M}$. Since we are summing the diagonal of a matrix, the trace operator kicks in.

$$\begin{aligned}
\mathrm{trace}(\mathbf{M}) &= \mathrm{trace}(\mathbf{I}_n - \mathbf{P}) \quad \text{(by definition of } \mathbf{M}\text{)} \\
&= n - \mathrm{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \quad \text{(by definition of } \mathbf{P}\text{)} \\
&= n - \mathrm{trace}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) \quad (\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}(\mathbf{BA})) \\
&= n - k \quad (\text{since } (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}_k)
\end{aligned}$$

Thus:

$$\begin{aligned}
\mathrm{E}[S^2|\mathbf{X}] &= \frac{\sigma^2}{n-k}\mathrm{trace}(\mathbf{M}) \\
&= \frac{\sigma^2}{n-k}\, n-k \\
&= \sigma^2
\end{aligned}$$

So far it has been proved that $\mathbf{b}$ is unbiased and efficient under A.1 - A.4. Likewise, an unbiased estimator S^2 has been proposed for σ^2, for which a `Python` implementation can be found in the same GitHub repository as this paper's.

These are the most important Finite Sample properties of $\mathbf{b}$ before proceeding to statistical tests. We are now ready to explain how to carry out hypothesis testing for some coefficients of the vector $\mathbf{b}$.

2.5 Hypothesis Testing in Finite Sample Theory