Algorithms for Causal Inference (I)

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 Finite Sample Theory

2.1 Notation

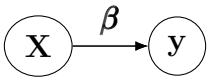
Before proceeding with Causal Inference methods for finite samples, it is convenient to introduce the notation to be used.

Let $\mathbf{y} \in \mathfrak{R}^n$ denote the target vector o dependent variable vector and n be the total number of observations. Let $\mathbf{x}_i \in \mathfrak{R}^K$ 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 $\mathbf{X} \in \mathfrak{R}^{n \times K}$.

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 $\boldsymbol{\varepsilon} \in \mathfrak{R}^n$ 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 $y_i \in \mathfrak{R} \, \forall i$, 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 $\boldsymbol{\beta} \in \mathfrak{R}^K$. Graphically, using a DAG:



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: $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}_n$

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

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 (5) 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 $\mathbf{X}$:

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

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:

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

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 $\hat{\boldsymbol{\beta}} = (\mathbf{D} + \mathbf{A})\mathbf{y}$

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

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

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

Deriving the sampling error:

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

Then:

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

The step that was skipped is reproduced below:

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

Note that

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

Clearly:

$$\sigma^2[\mathbf{D}\mathbf{D}'+\mathbf{D}\mathbf{A}'+\mathbf{A}\mathbf{D}'+\mathbf{A}\mathbf{A}'] = \sigma^2[\mathbf{D}\mathbf{D}'+\mathbf{A}\mathbf{A}']$$

Consequently:

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

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

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

Indeed, the OLS estimator $\mathbf{b}$ 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

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

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:

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

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

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

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:

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

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

Since estimating a Linear Regression model with OLS procures a vector of estimators, it is in the interest of the researcher to test several restrictions. A common value to test is 0, which boils down to checking whether a particular feature is trivial in explaining $\mathbf{y}$ or not. Mathematically, we test if the true parameter value $\boldsymbol{\beta} \in \bar{\boldsymbol{\beta}}_0$, where the latter term represents the null hypothesis space.

Denote $\mathbf{b} - \bar{\boldsymbol{\beta}}_0$ as the sampling error, we need to construct a test statistics whose probability distribution is known under H_0. In Finite Sample Theory, it is needed to impose normaly on $\boldsymbol{\varepsilon}|\mathbf{X}$, otherwise the joint distribution of $(\boldsymbol{\varepsilon}, \mathbf{X})$ has to be specified.

Since we are following a frequentist approach, no prior beliefs are taken into account, and the *ratio decidendi* gets down to analyzing the p-value, i.e: the probability of finding test values at least as extreme as the observed ones. Confidence Intervals (CI), will be defined for a $\alpha = 5\%$ significance level, meaning that if p-value < 0.05, the data provided enough evidence against H_0, so it is rejected. Otherwise, we simply fail to reject the null hypothesis. This is cumbersome as no probability distribution is being defined for H_0. In Bayesian Statistics the probability of rejecting H_0 can be exactly derived by defining Credible Intervals (to be covered in future work).

Assuming $\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be problematic. Nevertheless, error measurement is the main objective of the normal distribution, and feature engineering in the form of nonlinear transformations can be carried out on the regressors so as to approximate the distribution of $\boldsymbol{\varepsilon}|\mathbf{X}$ to a Gaussian one. Thus, we introduce:

- A.5) Normality: $\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Note that A.2 implies $\boldsymbol{\mu} = \mathbf{0}$ and A.4 entails $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$ so: $\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

Notice how neither the mean nor the variance of $\boldsymbol{\varepsilon}|\mathbf{X}$ depend on $\mathbf{X}$. This means that the marginal distribution of $\boldsymbol{\varepsilon}$ is also normal, concretely: $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

Let us now derive the sampling error for the OLS estimator:

$$\begin{aligned} \mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \end{aligned}$$

Thus the sampling error is expressed as:

$$\mathbf{b} - \boldsymbol{\beta} = \mathbf{A}\boldsymbol{\varepsilon} \quad (\mathbf{A} \coloneqq (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \tag{8}$$

Conditioning on $\mathbf{X}$, the moments of the sampling error are:

$$\begin{aligned}
\mathrm{E}[\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] &= \mathbf{A}\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \quad \text{(by linearity of CEF)} \\
&= \mathbf{0} \quad \text{(by A.2)} \\
\mathrm{V}[\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] &= \mathbf{A}\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\mathbf{A}' \\
&= \sigma^2\mathbf{A}\mathbf{A}' \quad \text{(by A.4)} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\underbrace{\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{I}_k} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Since the sampling error depends on $\boldsymbol{\varepsilon}$ and by A.5 $\boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$:

$$\mathbf{b} - \boldsymbol{\beta}|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{9}$$

We firstly derive the test statistic for individual restrictions on $\boldsymbol{\beta} \in \bar{\boldsymbol{\beta}}_0$ and then proceed to the multivariate case.

Note how the variance of the sampling error conditional on $\mathbf{X}$ depends on the feature matrix. Consequently, the marginal distribution of $\mathbf{b} - \boldsymbol{\beta}|\mathbf{X}$ is not necessarily Gaussian. Decades ago it did not use to be of much practicality to deal with test statistics that depended on the sample ($\mathbf{y}$, $\mathbf{X}$) like the case above. Nowadays, it is common to resort to statistical software that supports non-standardized distributions (`scipy.stats.norm()` in `Python` or `pnorm()` in `R`, for instance). Anyway, it is not difficult to standardize $\mathbf{b} - \boldsymbol{\beta}|\mathbf{X}$ so as to obtain a non-sample dependent distribution. Note that $\mathbf{X}$ is only present in the variance, so dividing the sampling error by its standard deviation yields:

$$\frac{\mathbf{b} - \boldsymbol{\beta}}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}}}\bigg|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k) \tag{10}$$

Suppose we seek to test the following restriction:

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

Then, under the truth of H_0 Equation (10) turns into:

$$z_k \coloneqq \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2\left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}}\bigg|\mathbf{X} \sim \mathcal{N}(0, 1) \tag{11}$$

If the parameter σ^2 is known, then the test statistic z_k follows a standard normal distribution. Since this is not often the case, we estimate the variance of the sampling error with S^2 (which was proved to be unbiased). Thus, we work with:

$$\left.\frac{b_k - \bar{\beta}_k}{\sqrt{S^2 \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}}\right| \mathbf{X} \tag{12}$$

Which no longer follows a standard normal distribution since $S^2 \neq \sigma^2$ and the former is a random variable (not a parameter). However, we can reach a known probability distribution:

$$\begin{aligned}
\frac{b_k - \bar{\beta}_k}{\sqrt{S^2 \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}} &= \frac{b_k - \bar{\beta}_k}{\sqrt{S^2 \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}} \sqrt{\frac{\sigma^2}{\sigma^2}} \\
&= \underbrace{\frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}}}_{z_k} \sqrt{\frac{\sigma^2}{S^2}} \\
&= \frac{z_k}{\sqrt{S^2/\sigma^2}}
\end{aligned}$$

Recall the definition of S^2 and plug it into the denominator:

$$\begin{aligned}
\sqrt{\frac{\mathbf{e}'\mathbf{e}/(n-K)}{\sigma^2}} &= \sqrt{\frac{\mathbf{e}'\mathbf{e}}{(n-K)\sigma^2}} \\
&= \sqrt{\frac{q}{n-K}} \quad (q \coloneqq \mathbf{e}'\mathbf{e}/\sigma^2)
\end{aligned}$$

It was already shown that $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$, so:

$$\begin{aligned}
q &= \frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2} \\
&= \frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{M}\frac{\boldsymbol{\varepsilon}}{\sigma}
\end{aligned}$$

By A.5: $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ so $\frac{\boldsymbol{\varepsilon}}{\sigma} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.

Since $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j$, then q is a sum of squared standard normal variables, the very definition of a chi-squared random variable. Note that $\text{rank}(\mathbf{M}) = \text{trace}(\mathbf{M})$ since it is an idempotent matrix. Consequently, q has $n - K$ degrees of freedom: $q|\mathbf{X} \sim \chi^2(n-K)$.

The only stochastic component of z_k is b_k since $\bar{\beta}_k$ is selected by the econometrician and σ^2 is a parameter. Let us analyze whether it is correlated with $\mathbf{e}$, the other random component of q:

$$\mathrm{Cov}(\mathbf{b}, \mathbf{e}|\mathbf{X}) = \mathrm{E}\left[(\mathbf{b} - \mathrm{E}[\mathbf{b}|\mathbf{X}])\,(\mathbf{e} - \mathrm{E}[\mathbf{e}|\mathbf{X}])'\,|\mathbf{X}\right]$$

$$\begin{aligned}
\mathbf{b} - \mathrm{E}[\mathbf{b}|\mathbf{X}] &= \mathbf{b} - \boldsymbol{\beta} \quad \text{(since } \mathbf{b} \text{ is unbiased)} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \quad \text{(by A.1)} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\
&= \mathbf{A}\boldsymbol{\varepsilon} \quad (\mathbf{A} \coloneqq (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')
\end{aligned}$$

$$\begin{aligned}
\mathbf{e} - \mathrm{E}[\mathbf{e}|\mathbf{X}] &= \mathbf{b} - \mathrm{E}[\mathbf{y} - \mathbf{X}\mathbf{b}|\mathbf{X}] \\
&= \mathbf{e} - \mathrm{E}[\mathbf{y}|\mathbf{X}] + \mathrm{E}[\mathbf{X}\mathbf{b}|\mathbf{X}] \\
&= \mathbf{e} - \mathrm{E}[\mathbf{X}, \boldsymbol{\beta} + \boldsymbol{\varepsilon}|\mathbf{X}] + \mathbf{X}\mathrm{E}[\mathbf{b}|\mathbf{X}] \\
&= \mathbf{e} - \mathbf{X}\boldsymbol{\beta} + \mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] + \mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{e} \quad \text{(by A.2)}
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b}, \mathbf{e}|\mathbf{X}) &= \mathrm{E}[\mathbf{A}\boldsymbol{\varepsilon}\mathbf{e}'|\mathbf{X}] \\
&= \mathrm{E}[\mathbf{A}\boldsymbol{\varepsilon}(\mathbf{M}\boldsymbol{\varepsilon})'|\mathbf{X}] \\
&= \mathrm{E}[\mathbf{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}|\mathbf{X}] \quad \text{(since } \mathbf{M} \text{ is symmetric)} \\
&= \mathbf{A}\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{M} \\
&= \sigma^2\mathbf{A}\mathbf{M} = \sigma^2\mathbf{M}\mathbf{A}' \\
&= \mathbf{0}
\end{aligned}$$

This last equality holds since:

$$\begin{aligned}
\mathbf{M}\mathbf{A}' &= \mathbf{M}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{I}_n - \mathbf{P})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{0} \quad (\text{since } \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{I}_k)
\end{aligned}$$

Since $(\mathbf{b},\ \mathbf{e}|\mathbf{X})$ is jointly normally distributed and $\mathrm{Cov}(\mathbf{b}, \mathbf{e}|\mathbf{X}) = \mathbf{0}$, then $\mathbf{b}$ is uncorrelated with $\mathbf{e}$, implying $\mathrm{cov}(z_k, q|\mathbf{X}) = 0$.

Combining $q|\mathbf{X} \sim \chi^2(n-K)$ and $\text{cov}(z_k, q|\mathbf{X}) = 0$, then:

$$t_k := \frac{b_k - \bar{\beta}_k}{\sqrt{S^2 \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}} \Bigg| \mathbf{X} \sim t_{n-K} \tag{13}$$

So for individually testing regression coefficients, the test-statistic t_k should be used, which follows a Student's t-distribution with $n-K$ degrees of freedom.

What if we would like to test several restrictions on $\boldsymbol{\beta}$? Individually testing coefficient values with the test statistic above would increase α, so all restrictions should be tested at once. This is called joint hypothesis testing and we can leverage matrix representations to carry it out.

Let $\mathbf{R} \in \mathfrak{R}^{\#r \times K}$ denote the restriction matrix, representing the weights affected by the null hypothesis space $\boldsymbol{\beta}_0$ and the relevant mathematical operations involved. Note that $\#r$ equals the total number of restrictions to test. Let $\mathbf{r} \in \mathfrak{R}^{\#r}$ be the vector of hypothesized coefficient values. Formally, we can test:

$$\text{H}_0\text{: } \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Just like before, we start by deriving the sampling error under the truth of H_0:

$$\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})|\mathbf{X} = \mathbf{R}\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}$$

Now we compute the conditional expected value and variance:

$$\begin{aligned}
\text{E}[\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})|\mathbf{X}] &= \text{E}[\mathbf{R}\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \mathbf{R}\mathbf{A}\text{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \quad \text{(by linearity of CEF)} \\
&= \mathbf{0} \quad \text{(by A.2)} \\
\text{V}[\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})|\mathbf{X}] &= \text{V}[\mathbf{R}\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \mathbf{R}\mathbf{A}\text{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\mathbf{A}'\mathbf{R}' \\
&= \sigma^2\mathbf{R}\mathbf{A}\mathbf{A}'\mathbf{R}' \quad \text{(by A.4)} \\
&= \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'
\end{aligned}$$

Notice how neither $\mathbf{R}$ nor $\mathbf{r}$ are stochastic. It is the researcher who defines the maintained hypothesis H_0. Thus, it is not restrictive to assume $\text{rank}(\mathbf{R}) = \#r$, that is, $\mathbf{R}$ is of full-row rank. Indeed, no linear combinations should be present in $\mathbf{R}$, otherwise inference cannot be carried out.

Thus, since $\mathbf{R}$ is not random and we concluded that $\mathbf{b} - \boldsymbol{\beta}|\mathbf{X}$ follows a Gaussian distribution:

$$\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') \tag{14}$$

Standardizing the random variable (divide by its standard deviation):

$$\mathbf{R}(\mathbf{b} - \boldsymbol{\beta})\sigma^2 \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1/2} \Big| \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$$

It is not common to proceed with this expression as σ^2 is usually not known and the squared term is preferred. Thus, the test statistic becomes:

$$\frac{(\mathbf{b} - \boldsymbol{\beta})'\mathbf{R}' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \mathbf{R}(\mathbf{b} - \boldsymbol{\beta})}{S^2} \Bigg| \mathbf{X}$$

Note that $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ under H_0 so:

$$\frac{(\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r})}{S^2} \Bigg| \mathbf{X}$$

Following the previous strategy of multiplying and dividing by σ^2 and recalling the definition of S^2 we obtain:

$$\frac{(\mathbf{Rb} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r})/\#r}{S^2} \Bigg| \mathbf{X} = \frac{\mathbf{w}/\#r}{q/(n-K)} \Bigg| \mathbf{X}$$

Where $\mathbf{w} \coloneqq (\mathbf{Rb} - \mathbf{r})'\sigma^2 \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r})$.

Let $\mathbf{v} \coloneqq \mathbf{Rb} - \mathbf{r}$, then:

$$\mathbf{w} = \mathbf{v}'\mathrm{V}[\mathbf{v}|\mathbf{X}]\mathbf{v}$$

Since $\mathbf{R}$ is of full-row rank by construction, $\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is nonsingular. Note that by A.3, $\mathbf{X}$ is of full-column rank, so it follows that $(\mathbf{X}'\mathbf{X})$ is Positive Definite and thus nonsingular. Adding a full-row rank matrix $\mathbf{R}$ to $(\mathbf{X}'\mathbf{X})^{-1}$ does not alter its nonsingularity. Also, note that $\mathtt{dim}(\mathbf{X}'\mathbf{X}) = K \times K$ and $\mathtt{dim}(\mathbf{R}) = \#r \times K$ so $\mathtt{dim}(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') = \#r \times \#r$. Furthermore, $\sqrt{V[\mathbf{v}|\mathbf{X}]}\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$.

Consequently, $\mathbf{w} \sim \chi^2(\#r)$.

Recall that $q \sim \chi^2(n-K)$. Lastly, the only stochastic element of $\mathbf{w}$ is $\mathbf{b}$, whereas the non-deterministic element of q is $\mathbf{e}$. We already proved $\text{Cov}(\mathbf{b}, \mathbf{e}|\mathbf{X}) = \mathbf{0}$ so $(\mathbf{w}, q)$ is jointly uncorrelated conditional on $\mathbf{X}$. Thus:

$$F = \frac{\mathbf{w}/\#r}{q/(n-K)}\bigg|\mathbf{X} \sim F(\#r, n-K) \tag{15}$$

Equation (15) represents the F-statistic used for joint hypothesis testing. It follows a F-distribution, defined as the ratio of two uncorrelated chi-squared random variables divided by their respective degrees of freedom.

The reader can refer to our GitHub repository for some Python implementations of these test-statistics.

2.6 Generalized Least Squares (GLS)

Failure of A.4 entails that the values of the off-diagonal of $\text{E}[\boldsymbol{\varepsilon}|\mathbf{X}]$ are non-zero, and thus each $\varepsilon_i\varepsilon_j$ in said matrix would not generally be a linear function of $\mathbf{X}$. Likewise, hypothesis testing is no longer valid, since the t-statistic does not follow a Student's t-distribution anymore (same applies to the F-statistic). Furthermore, $\mathbf{b}$ is not BLUE since the Gauss-Markov Theorem does not hold without A.3. Note that $\mathbf{b}$ is still unbiased though, since A.4 was not relevant to derive unbiasedness. Most of the time it is wise to relax A.4, since it is not likely that the variance of the error term is proportional to the identity matrix $\mathbf{I}_n$.

Suppose that for any $\sigma^2 > 0$, $\exists \mathbf{V}(\mathbf{X}) \in \mathfrak{R}^{n\times n}$ s.t $\text{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{V}(\mathbf{X})$, where $\mathbf{V}(\mathbf{X})$ is nonsingular and known. Thus, $\exists \mathbf{C} \in \mathfrak{R}^{n\times n}$ s.t $\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}$ (note that $\mathbf{V}(\mathbf{X}) = \mathbf{V}$).

Let the assumption below be denoted as A.6, a new Regression model is formed as following: $\mathbf{C}\mathbf{y} = \mathbf{C}\text{E}[\mathbf{y}|\mathbf{X}] + \mathbf{C}\boldsymbol{\varepsilon}$

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \tag{16}$$

So A.1 still holds! Let us see if A.2 is still satisfied:

$$\begin{aligned}
\text{E}[\tilde{\boldsymbol{\varepsilon}}|\mathbf{X}] &= \text{E}[\mathbf{C}\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \mathbf{C}\text{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \mathbf{0}
\end{aligned}$$

The equivalent of A.2 is then proved to hold: $\mathrm{E}[\tilde{\boldsymbol{\varepsilon}}|\mathbf{X}] = \mathbf{0}$ since $\mathbf{C}$ is a function of feature matrix $\mathbf{X}$.

Since $\mathbf{V}$ is nonsingular and a function of $\mathbf{C}$, the latter is full rank. By A.3, $\mathbf{X}$ is of full-column rank so $\mathrm{rank}(\mathbf{CX}) = \mathrm{rank}(\mathbf{X}) = \mathrm{K}$. A.3 is also valid, let's check A.4:

$$\begin{aligned}
\mathrm{E}[\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}'|\mathbf{X}] &= \mathrm{E}[\mathbf{C}\boldsymbol{\varepsilon}(\mathbf{C}\boldsymbol{\varepsilon})'|\mathbf{X}] \\
&= \mathrm{E}[\mathbf{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{C}'|\mathbf{X}] \\
&= \mathbf{C}\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{C}' \quad (\text{since } \mathbf{C} \text{ is a function of } \mathbf{X}) \\
&= \sigma^2\mathbf{C}\mathbf{V}\mathbf{C}' \quad (\text{by A.6}) \\
&= \sigma^2 \quad (\text{since } \mathbf{V} = \mathbf{C}^{-1}(\mathbf{C}')^{-1})
\end{aligned}$$

Finally, since $\mathbf{C}\boldsymbol{\varepsilon}$ is a linear transformation, A.5 is likewise satisfied, so $\tilde{\boldsymbol{\varepsilon}}|\mathbf{X}$ is jointly normally distributed. All the results we derived earlier hold for the new Linear Regression Model. As this is a generalization of the classical Regression Model, it is referred to as Generalized Least Squares (GLS). The GLS estimator can be derived as following:

$$\hat{\boldsymbol{\beta}}^{GLS} = \arg\min_{\tilde{\boldsymbol{\beta}}} \left\{ (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \right\} \tag{17}$$

Where $\tilde{\boldsymbol{\beta}}$ is a running parameter. Note that since A.3 is satisfied, we can obtain a closed-form solution.

Note that

$$\begin{aligned}
(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) &= (\mathbf{y}' - \tilde{\boldsymbol{\beta}}'\mathbf{X}')(\mathbf{V}^{-1}\mathbf{y} - \mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}}) \\
&= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} \\
&= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - 2\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}}
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{\partial S\tilde{S}R(\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}} &= -2\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{0} \\
&= \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
\hat{\boldsymbol{\beta}}^{GLS} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}
\end{aligned} \tag{18}$$

Alternatively, starting from the OLS estimator expression for the new Regression Model:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{GLS} &= [(\mathbf{CX})'\mathbf{CX}]^{-1}(\mathbf{CX})'\mathbf{Cy} \\
&= (\mathbf{X}'\mathbf{C}'\mathbf{CX})^{-1}\mathbf{X}'\mathbf{C}'\mathbf{Cy} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad \text{(by A.1 and since } \mathbf{V}^{-1} = \mathbf{C}'\mathbf{C})
\end{aligned}
$$

We now prove that in the absence of conditional homoskedasticity $\hat{\boldsymbol{\beta}}^{GLS}$ is more efficient than $\mathbf{b}$

$$
\begin{aligned}
\mathrm{V}[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}] &= \mathrm{V}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}|\mathbf{X}] \\
&= \mathrm{V}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\mathbf{X}] \quad \text{(by A.1)} \\
&= \mathrm{V}[\boldsymbol{\beta} + \tilde{\mathbf{A}}\boldsymbol{\varepsilon}|\mathbf{X}] \quad (\tilde{\mathbf{A}} \coloneqq (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}) \\
&= \tilde{\mathbf{A}}\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\tilde{\mathbf{A}}' \quad \text{(since } \boldsymbol{\beta} \text{ is a parameter)} \\
&= \sigma^2\tilde{\mathbf{A}}\mathbf{V}\tilde{\mathbf{A}}' \quad \text{(by A.6)} \\
&= \sigma^2 \underbrace{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}}_{\tilde{\mathbf{A}}} \mathbf{V} \underbrace{\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}}_{\tilde{\mathbf{A}}'} \\
&= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{V}[\mathbf{b}|\mathbf{X}] &= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \\
&= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\mathbf{X}] \quad \text{(by A.1)} \\
&= \mathrm{V}[\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] \quad (\mathbf{A} \coloneqq (\mathbf{X}'\mathbf{X}))^{-1}\mathbf{X}') \\
&= \mathbf{A}\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\mathbf{A}' \quad \text{(since } \boldsymbol{\beta} \text{ is a parameter)} \\
&= \sigma^2\mathbf{A}\mathbf{V}\mathbf{A}' \quad \text{(by A.6)} \\
&= \sigma^2 \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{A}} \mathbf{V} \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{A}'}
\end{aligned}
$$

Since $\hat{\boldsymbol{\beta}}^{GLS}$ is unbiased, linear and resilient to homoskedasticity, i.e, satisfies the Gauss-Marvov Theorem, then:

$$
\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \geq \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \implies \mathrm{V}[\mathbf{b}|\mathbf{X}] \geq \mathrm{V}[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}] \tag{19}
$$

2.7 Other relevant results

We prove that the OLS estimator $\mathbf{b}$ truly minimizes SSR. Let $\tilde{\boldsymbol{\beta}}$ be any hypothetical linear estimator of $\boldsymbol{\beta}$, then:

$$\begin{aligned}
SSR_{\tilde{\boldsymbol{\beta}}} &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \quad \text{(add-and-subtract strategy)} \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}))'(\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})) \\
&= (\mathbf{e} + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}))'(\mathbf{e} + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})) \quad (\mathbf{e} \coloneqq \mathbf{y} - \mathbf{X}\mathbf{b}) \\
&= \mathbf{e}'\mathbf{e} + \mathbf{e}'\mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) + (\mathbf{b} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{e} + (\mathbf{b} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) \\
&= \mathbf{e}'\mathbf{e} + (\mathbf{b} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) \quad \text{(since by Normal Equations: } \mathbf{X}'\mathbf{e} = \mathbf{0})
\end{aligned}$$

Then unless $\tilde{\boldsymbol{\beta}} = \mathbf{b}$: $SSR_{\tilde{\boldsymbol{\beta}}} \geq SSR_{\mathbf{b}}$

Let us derive the Restricted Least Squares estimator vector $\hat{\boldsymbol{\beta}}$ under H_0. Notice how we impose a set of restrictions by performing hypothesis testing on a selection of coefficients

$$\begin{aligned}
&\mathrm{H}_0\text{: } \mathbf{R}\boldsymbol{\beta} = \mathbf{r} \\
&\mathrm{H}_1\text{: } \mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}
\end{aligned}$$

Thus, the optimization problem to be solved boils down to:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = \underset{\tilde{\boldsymbol{\beta}}}{\arg\min} \ &\left\{ \frac{1}{2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \right\} \\
&\text{s.t } \mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{r}
\end{aligned} \tag{20}$$

The Lagrangian can be formed as:

$$\mathcal{L}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda}'(\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r}) \tag{21}$$

Consequently, since by A.3 a closed-form or Newtonian solution is achievable, we solve the following expression by setting the gradient vector to $\vec{0}$:

$$\hat{\boldsymbol{\beta}} = \underset{\tilde{\boldsymbol{\beta}},\, \boldsymbol{\lambda}}{\arg\min} \ \left\{ \frac{1}{2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda}'(\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r}) \right\} \tag{22}$$

Note the following properties:

$$\texttt{length}(\boldsymbol{\lambda}) = \#\mathbf{r} \quad \texttt{dim}(\mathbf{R}) = \#\mathbf{r} \,\mathrm{x}\, K \quad \texttt{length}(\mathbf{r}) = \#\mathbf{r}$$

Expanding the objective function:

$$SSR(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda}) = \frac{1}{2}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\,\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\,\tilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda}'(\mathbf{R}\,\tilde{\boldsymbol{\beta}} - \mathbf{r})$$

Minimizing $SSR(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda})$:

$$\frac{\partial SSR(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda})}{\partial \tilde{\boldsymbol{\beta}}} = -\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\,\tilde{\boldsymbol{\beta}} + \mathbf{R}'\boldsymbol{\lambda} = 0 \quad \text{(I)}$$

$$\frac{\partial SSR(\tilde{\boldsymbol{\beta}}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \mathbf{R}\,\tilde{\boldsymbol{\beta}} - \mathbf{r} = 0 \qquad \text{(II)}$$

From (I):

$$\begin{aligned} \mathbf{X}'\mathbf{X}\,\tilde{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} - \mathbf{R}'\boldsymbol{\lambda} \\ \hat{\boldsymbol{\beta}} &= \underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}_{\mathbf{b}} - (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}'\boldsymbol{\lambda}) \end{aligned}$$

Note that by A.3 $\mathbf{X}'\mathbf{X}$ is invertible. Plugging this expression into (II):

$$\begin{aligned} \mathbf{R}\mathbf{b} - \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} - \mathbf{r} &= 0 \\ \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} &= \mathbf{R}\mathbf{b} - \mathbf{r} \\ \boldsymbol{\lambda} &= [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \end{aligned}$$

Consider the aforementioned vector dimension properties:

$$\begin{aligned} \texttt{dim}(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}') &= (\#\mathbf{r}\,\mathrm{x}\,\mathrm{K})(\mathrm{K}\,\mathrm{x}\,\mathrm{K})(K\,\mathrm{x}\,\#\mathbf{r}) \\ &= (\#\mathbf{r}\,\mathrm{x}\,\mathrm{K})(\mathrm{K}\,\mathrm{x}\,\#\mathbf{r}) \\ &= (\#\mathbf{r}\,\mathrm{x}\,\#\mathbf{r}) \end{aligned}$$

So $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is a square matrix. Since it is also of full rank, it is Positive Definite and thus invertible.

Plugging $\boldsymbol{\lambda}$ into the above expression for $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \qquad (23)$$

Let us derive SSR for the restricted model:

$$\begin{aligned}
SSR_r &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= (\underbrace{\mathbf{y} - \mathbf{X}\mathbf{b}}_{\mathbf{e}} + \mathbf{X}\mathbf{b} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\underbrace{\mathbf{y} - \mathbf{X}\mathbf{b}}_{\mathbf{e}} + \mathbf{X}\mathbf{b} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad \text{(add-and-subtract strategy)} \\
&= (\mathbf{e} + \mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}}))'(\mathbf{e} + \mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}})) \\
&= \mathbf{e}'\mathbf{e} + \mathbf{e}'\mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}}) + (\mathbf{b} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{e} + (\mathbf{b} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}}) \\
&= \mathbf{e}'\mathbf{e} + (\mathbf{b} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}}) \quad \text{(by Normal Equations } \mathbf{X}'\mathbf{e} = \mathbf{0})
\end{aligned}$$

Then:

$$\begin{aligned}
SSR_r - SSR_u &= \mathbf{e}'\mathbf{e} + (\mathbf{b} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}}) - \mathbf{e}'\mathbf{e} \\
&= (\mathbf{b} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}})
\end{aligned}$$

Plugging in the complete expression of $\hat{\boldsymbol{\beta}}$:

$$[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})]' \underbrace{\mathbf{X}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{I}_k}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})]$$

Consequently:

$$(\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \underbrace{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}}_{\mathbf{I}_{\#\mathbf{r}}}(\mathbf{R}\mathbf{b} - \mathbf{r})]$$

$$\begin{aligned}
SSR_r - SSR_u &= (\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})] \\
&= \boldsymbol{\lambda}'\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \quad \text{(by the definition of } \boldsymbol{\lambda})
\end{aligned}$$

Note the following equality from FOC (I):

$$\begin{aligned}
\mathbf{R}'\boldsymbol{\lambda} &= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{X}'\hat{\boldsymbol{\varepsilon}} \quad \text{(by the definition of a residual vector)}
\end{aligned}$$

Thus:

$$\begin{aligned}
SSR_r - SSR_u &= \boldsymbol{\lambda}'\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \\
&= \hat{\boldsymbol{\varepsilon}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\varepsilon}} \\
&= \hat{\boldsymbol{\varepsilon}}'\mathbf{P}\hat{\boldsymbol{\varepsilon}} \quad \text{(by the definition of the Projection Matrix)}
\end{aligned}$$

Let us come back to the F-statistic:

$$F = \frac{(\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/\#\mathbf{r}}{S^2}$$

$$= \frac{(\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/\#\mathbf{r}}{\mathbf{e}'\mathbf{e}/(n-K)}$$

$$= \frac{(SSR_r - SSR_u)/\#\mathbf{r}}{SSR_u/(n-K)}$$

Since

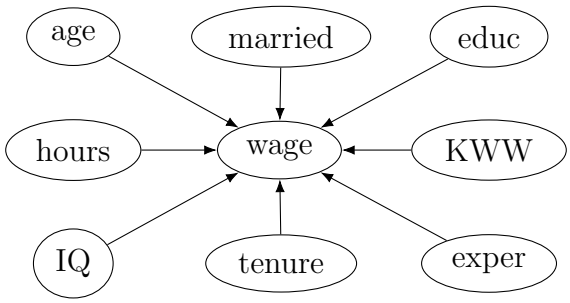
$$SSR_r - SSR_u = (\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})$$
$$SSR_u = \mathbf{e}'\mathbf{e}$$
$$S^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$$

Indeed, we can express the F-ratio as a relation between SSR_r and SSR_u.

Thus, we can perform hypothesis testing comparing the Sum of Squared Residuals from the restricted regression model and the original specification.

2.8 Application of Finite Sample Theory in Causal Inference

It is now time to apply the techniques learned in this section to a real-world problem: wage estimation. We resorted to Woolridge's econometric datasets, specifically `wage2.csv`, a cross-sectional data on wages collected from 935 individuals. A total of 16 features are available. For the sake of coherence, we do not use the entire dataset as to adapt it to the Finite Sample Theory context. Thus, only 200 observations and 8 variables will be used. The resulting DAG is as follows:



Where KWW stands for knowledge of world work score, *exper* means professional experience in years, IQ is intelligence quotient, *hours* represents average working weekly hours and *married* is a one-hot encoded variable that equals 1 whether the i-th observation is married and 0 otherwise.

Note that no confounders are found in the preceding causal diagram. Thus, no back doors have to be closed. This fits A.2 as $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}$. By A.1, the Linear Regression Model to be estimated is:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{24}$$

$$y_i = \text{wage}_i,\ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_8 \end{bmatrix},\ \mathbf{x}_i = \begin{bmatrix} 1 \\ \text{hours}_i \\ \text{age}_i \\ \vdots \\ \text{married}_i \end{bmatrix} \quad n = 200,\ K = 9$$

Applying the functions defined in the Python file `Finite-Sample-Theory.py`:

Table 1: Finite Sample Theory Regression

	Features	Estimate	Standard Error	p-value
b_0	intercept	-552.735	402.6729	0.1714
b_1	hours	1.833	3.7992	0.6298
b_2	IQ	4.187	2.4626	0.4430
b_3	KWW	20.567	4.7370	2.2845e-05**
b_4	educ	7.825	16.4027	0.0109**
b_5	exper	4.798	8.2763	0.9410
b_6	tenure	-10.560	5.7348	0.0671
b_7	age	0.357	11.2547	0.9746
b_8	married	58.767	82.6653	0.4780

Before interpreting results, bear in mind that the vast majority of the dataset has not been considered. A highly restrictive approach tends to yield unrealistic estimates.

Note that all coefficients b_k represent the marginal effect of feature x_k on y_i. Since we estimated a level-level model (no non-linear transformations were applied), the k-th coefficient can be interpreted as the average effect on y_i by a one-unit change in x_k. For instance, for b_4, an additional year of education is associated with an increase in wages of \$7.825. If all assumptions were satisfied, rather than 'associated', we could state that an additional unit of x_k **causes** y_i to increase by b_k. However, this most probably will not be the case.

Firstly, it seems that only KWW and *educ* are individually statistically significant at a 5% significance level. It is highly unlikely that no variation in wages is explained by variables such as *age* (junior employees tend to earn less), *exper* (higher experience is associated with more generous salaries) or *hours* (working more hours tends to increase wages). Furthermore, the sign of b_6, the coefficient of *tenure* is negative. Undoubtedly, our estimation is performing poorly. Even though it was proved that under the Finite Sample Theory model the OLS estimator is unbiased and efficient, our results are indicative of the failure of some assumptions.

A.2 is highly unlikely to hold, *exp*, *age* and *educ* are probably confounded: education years and experience increase with time, and so does age, which in term affects experience and education. IQ and KWW are error-ridden measured variables (does it really exist a test that can perfectly quantify one's skills and intelligence?). A.3 would be no problem since we propose a robust estimation of $(\mathbf{X}'\mathbf{X})^{-1}$. A.4 probably is doomed to fail, since homoskedatic error terms are scarce, to say the least. Feauture engineering could potentially solve this last issue, but with such a limited number of observations in the dataset it might not make a difference.

A joint hypothesis test of significance yields a F-statistic of 0.000159 and p-value equal to 0.999. Seems like the model is not statistically significant at all.

A naive approach would be to estimate Equation (24) without going over Labor Economics. The aforementioned confounding issue and lack of enough data leads to terrible results in terms of Causal Inference.

In the next section, we introduce Large Sample Theory, which will relax some assumptions of the Finite Sample Theore Regression Model. We will leverage Probabilistic Convergence Theory to attain such feat and accommodate the model to larger samples.

Nevertheless, let us briefly explain how to overcome failure of A.3 before proceeding to a new Causal Inference Topic.

2.9 What if A.3 fails? Gradient Descent Algorithm

A.3 necessarily assumes that $n \geq K$, which nowadays may be restrictive. In Big Data problems it is ubiquitous to find datasets in which features are more abundant that training examples. Moreover, the symmetric matrix $\mathbf{X}'\mathbf{X}$ must be nonsingular so it can be inverted, forcing all eigenvalues to be non-negative ($\lambda_i > 0 \ \forall i$). Additionally, for considerably large datasets, inverting said matrix may be computationally expensive. Indeed, the time complexity of calculating the $K \times K$ matrix $\mathbf{X}'\mathbf{X}$ is $O(nK^2)$ and inverting it adds time complexity $O(K^3)$. For the other matrix $\mathbf{X}'\mathbf{y}$ time complexity is lower: $O(nK)$. Thus, asymptotically, time complexity is $O(nK^2 + K^3)$

Gradient Descent is a convenient Algorithm that effectively deals with these issues, as it does not require to invert the cross-product matrix of features. It was proposed by Cauchy (1847) and is still widely used for optimizing AI Algorithms.

Pseudo-code describing Gradient Descent Algorithm can be found below. It can be used to obtain the K dimensional vector of weights $\mathbf{b}$. Within the same GitHub repository the reader may find Python code to implement it themselves.

Algorithm 1: Gradient Descent Algorithm

Input: Feature matrix $\mathbf{X}$, Target vector $\mathbf{y}$, Learning Rate η and Maximum number of iterations N

Output: Estimated Model Coefficients $\mathbf{b}$

Randomly initialize: $b_j^{(0)} \sim N(0,1)$

for $i = 1, 2, \ldots, N$ **do**

$$\hat{\mathbf{y}}^{(i)} \leftarrow \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j^{(i)\prime}\mathbf{b}^{(i)}$$

$$\frac{\partial J(\tilde{\beta}_j)}{\partial b_j}^{(i)} \leftarrow -\frac{1}{K}\sum_{m=1}^{n}\left(\hat{y}_m^{(i)} - y_m^{(i)}\right)$$

$$b_j^{(i)} \leftarrow b_j^{(i)} - 2\eta\frac{\partial J(\tilde{\beta}_j)}{\partial b_j}^{(i)}$$

end

Since summing over all data points in each iteration increases the computational cost, a more efficient approach would require vectorizing said individual data points. Think of this as eliminating an additional explicit *for* loop. Indeed, the pseudo-code illustrating this idea and that finally got implemented is the following:

Algorithm 2: Vectorized Gradient Descent Algorithm

Input: Feature matrix $\mathbf{X}$, Target vector $\mathbf{y}$, Learning Rate η and Maximum number of iterations N

Output: Estimated Model Coefficients $\mathbf{b}$

Randomly initialize: $\mathbf{b}^{(0)} \sim \mathbf{N}_k(\mathbf{0}, \mathbf{I}_k)$

for $i = 1, 2, \ldots, N$ **do**

$\hat{\mathbf{y}}^{(i)} \leftarrow \mathbf{X}^{(i)}\mathbf{b}^{(i)}$

$\nabla J(\tilde{\boldsymbol{\beta}})^{(i)} \leftarrow -\frac{2}{K}\mathbf{X}'^{(i)}(\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)})$

$\mathbf{b}^{(i)} \leftarrow \mathbf{b}^{(i)} - \eta\nabla J(\tilde{\boldsymbol{\beta}})^{(i)}$

end

As pointed out, stacking data into vectors and matrices decreases running time. Numpy arrays served this purpose in the Python implementation.

An easy way to understand Gradient Descent Algorithm is by thinking of obtaining estimated weights as climbing down a hill.

Let us suppose we are at the very top of a mountain (or actually any random point of its skirt) and our objective is to get to the bottom. We start descending by taking several steps (learning rate or step size hyperparameter) in the direction of greatest decline (gradient of the objective function). After some steps (iterations), we finally reach our goal. By and large, it is very difficult not to convergence, although saddle points could potentially arise. Andrew Ng (2019) points out that it is unlikely that Gradient Descent Algorithm will not converge, as normally the vector of derivatives correctly leads the path towards optima and most cost functions are effectively covered by optimization Algorithms such as Adam (Adaptive Moment Estimation) in the context of Neural Networks.

For a randomly initialized value we move towards the optimum vector of weights that minimizes the objective function. Ideally, we would want to use prior knowledge to smartly initialize weights so that convergence is attained faster. This can be either performed by leveraging conjugate distributions or carrying out Bayesian Optimization. The former would be an example of simple Bayesian Statistics whereas the latter would require the use of MCMC methods via Gibbs Sampling Algorithm due to complex distributions. Decreasing learning rate as iterations increase would also smooth out results.