Advanced Econometrics: Finite Sample Theory

José Jaén Delgado

Introduction - In this paper solutions to most exercises from Chapter 1 of Fumio Hayashi's *Econometrics* are provided

1 Questions for Review

1.1 The Classical Linear Regression Model

1. Let us estimate a Linear Regression Model with OLS. Specifically, let it take the semi-log form:

$$\log(wage_i) = \beta_1 + \beta_2 S_i + \beta_3 tenure_i + \beta_4 exp_i + \varepsilon_i$$

We now show that changes in measurement units make a difference:

$$\begin{aligned}
\log(wage_i * 100) &= \beta_1 + \beta_2 S_i + \beta_3 tenure_i + \beta_4 exp_i + \varepsilon_i \\
\log(wage_i) + \log(100) &= \beta_1 + \beta_2 S_i + \beta_3 tenure_i + \beta_4 exp_i + \varepsilon_i \\
\log(wage_i) &= (\beta_1 - 2) + \beta_2 S_i + \beta_3 tenure_i + \beta_4 exp_i + \varepsilon_i
\end{aligned}$$

2. Let us prove $\mathrm{E}[\varepsilon_i \varepsilon_j | \mathbf{X}] = \mathrm{E}[\varepsilon_i | \mathbf{x}_i]\mathrm{E}[\varepsilon_j | \mathbf{x}_j]$ for random sampling when $i \neq j$

$$\begin{aligned}
\mathrm{E}[\varepsilon_i \varepsilon_j | \mathbf{X}] &= \mathrm{E}[\mathrm{E}[\varepsilon_i \varepsilon_j | \mathbf{X}, \varepsilon_j] | \mathbf{X}] \quad \text{(by LIE)} \\
&= \mathrm{E}[\varepsilon_j \mathrm{E}[\varepsilon_i | \mathbf{X}, \varepsilon_j] | \mathbf{X}] \quad \big(\text{by linearity of conditional expectations}\big) \\
&= \mathrm{E}[\varepsilon_j | \mathbf{x}_j]\mathrm{E}[\mathrm{E}[\varepsilon_i | \mathbf{X}, \varepsilon_j] | \mathbf{X}] \quad \big(\text{by random sampling properties}\big) \\
&= \mathrm{E}[\varepsilon_j | \mathbf{x}_j]\mathrm{E}[\varepsilon_i | \mathbf{x}_i]
\end{aligned}$$

Where we leveraged that $\mathrm{E}[\mathrm{E}[Y|X,Z]|X] = \mathrm{E}[Y|X]$ and $\mathrm{E}[\varepsilon_i | \mathbf{X}] = \mathrm{E}[\varepsilon_i | \mathbf{x}_i]$

3. Let us show that A.1 (Linearity) and A.2 (Strict Exogeneity) imply:

$$\mathrm{E}[y_i|\mathbf{X}] = \mathbf{x}_i'\boldsymbol{\beta}$$

We start by the general expression of any Regression Model:

$$\begin{aligned}
y_i &= \mathrm{E}[y_i|\mathbf{X}] + \varepsilon_i \\
y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \big(\text{by A.1: } \mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}\big) \\
\mathrm{E}[y_i|\mathbf{X}] &= \mathrm{E}[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i|\mathbf{X}] \quad \big(\text{taking conditional expectations}\big) \\
\mathrm{E}[y_i|\mathbf{X}] &= \mathrm{E}[\mathbf{x}_i'\boldsymbol{\beta}|\mathbf{X}] + \mathrm{E}[\varepsilon_i|\mathbf{X}] \\
\mathrm{E}[y_i|\mathbf{X}] &= \mathbf{x}_i'\boldsymbol{\beta} \quad \big(\text{by A.2: } \mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0\big)
\end{aligned}$$

Conversely, assuming $\mathrm{E}[y_i|\mathbf{X}] = \mathbf{x}_i'\boldsymbol{\beta}$:

$$\begin{aligned}
y_i &= \mathrm{E}[y_i|\mathbf{X}] + \varepsilon_i \\
y_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad \big(\text{A.1 is now proved}\big) \\
y_i - \mathbf{x}_i'\boldsymbol{\beta} &= \varepsilon_i \\
\mathrm{E}[y_i - \mathbf{x}_i'\boldsymbol{\beta}|\mathbf{X}] &= \mathrm{E}[\varepsilon_i|\mathbf{X}] \\
\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'\boldsymbol{\beta} &= \mathrm{E}[\varepsilon_i|\mathbf{X}] \\
\mathrm{E}[\varepsilon_i|\mathbf{X}] &= 0 \quad \big(\text{A.2 is now proved}\big)
\end{aligned}$$

4. Considering a random sample on Consumption and Disposable Income such that $(CON_i, YD_i) \overset{i.i.d}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we proceed to prove A.1 (Linearity), A.2 (Strict Exogeneity), A.3 (No Multicollinearity) and A.4 (Spherical Error Variance):

A.3 is pretty straightforward: $\Pr(\mathrm{rank}(\mathbf{X}'\mathbf{X}) = K) = 1$ since we are dealing with a random sample. Note that $\mathbf{X}$ is a $n \,\mathrm{x}\, K$ matrix, so we claim it is of full-column rank.

Since CON_i and YD_i are jointly normally distributed, their Conditional Expectation Function (CEF) is linear and the conditional variance of the target variable is equal to the unconditional one.

$$\mathrm{E}[CON_i|\mathbf{YD}] = \beta_1 + \beta_2\mathbf{YD}$$

$$\mathrm{V}[CON_i|\mathbf{YD}] = \mathrm{V}[CON_i]$$

A.1 is satisfied as CEF is linear due to properties of the normal distribution. Then:

$$\begin{aligned}
\mathrm{E}[CON_i|\mathbf{YD}] &= \mathrm{E}[\beta_1 + \beta_2\mathbf{YD} + \varepsilon_i|\mathbf{YD}] \\
\mathrm{E}[CON_i|YD_i] &= \mathrm{E}[\beta_1 + \beta_2 YD_i + \varepsilon_i|YD_i] \quad \text{(by random sampling)} \\
&= \beta_1 + \beta_2 YD_i + \mathrm{E}[\varepsilon_i|YD_i]
\end{aligned}$$

Note that $\mathrm{E}[\varepsilon_i|YD_i] = 0$ since $\mathrm{E}[CON_i|\mathbf{YD}] = \beta_1 + \beta_2\mathbf{YD}$. A.2 is thus proved.

For A.4 let us derive the expression of $\mathrm{E}[\varepsilon_i 2|\mathbf{YD}]$ where $\mathbf{X}$ represents the matrix of regressors $\mathbf{YD}$, and y_i the dependent variable CON_i:

$$\begin{aligned}
\mathrm{E}[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2|\mathbf{X}] &= \mathrm{E}[y_i^2 + (\mathbf{x}_i'\boldsymbol{\beta})'(\mathbf{x}_i'\boldsymbol{\beta}) - 2y_i\mathbf{x}_i'\boldsymbol{\beta}|\mathbf{X}] \\
&= \mathrm{E}[y_i^2|\mathbf{X}] + (\mathbf{x}_i'\boldsymbol{\beta})'(\mathbf{x}_i'\boldsymbol{\beta}) - 2\mathbf{x}_i'\boldsymbol{\beta}\mathrm{E}[y_i|\mathbf{X}] \\
&= \underbrace{\mathrm{V}[y_i|\mathbf{X}] + \mathrm{E}[y_i|\mathbf{X}]^2}_{\mathrm{E}[y_i|\mathbf{X}]^2} + (\mathbf{x}_i'\boldsymbol{\beta})'(\mathbf{x}_i'\boldsymbol{\beta}) - 2\mathbf{x}_i'\boldsymbol{\beta}\mathrm{E}[y_i|\mathbf{X}] \\
&= \mathrm{V}[y_i|\mathbf{X}] + \mathrm{E}[y_i|\mathbf{X}]^2 + (\mathbf{x}_i'\boldsymbol{\beta})'(\mathbf{x}_i'\boldsymbol{\beta}) - 2(\mathbf{x}_i'\boldsymbol{\beta})'\underbrace{(\mathbf{x}_i'\boldsymbol{\beta})}_{\mathrm{E}[y_i|\mathbf{X}]} \\
&= \mathrm{V}[y_i|\mathbf{X}] \quad (\text{since } \mathrm{E}[y_i|\mathbf{X}]^2 = (\mathbf{x}_i'\boldsymbol{\beta})'(\mathbf{x}_i'\boldsymbol{\beta})) \\
&= \mathrm{V}[y_i] \quad \text{(by normal distribution properties)} \\
&= \sigma^2
\end{aligned}$$

A.4 is satisfied as $\mathrm{E}[\varepsilon_i^2|\mathbf{X}] = \sigma^2 \ \forall i$

5. Let us show that the full-column rank condition, i.e: $\Pr(\mathrm{rank}(X) = K) = 1$ in the Simple Regression Model ($K = 2$) implies that $x_{i2} \neq x_{j2}$ for some pairs $i \neq j$.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} \\ 1 & x_{22} \\ \vdots & \vdots \\ 1 & x_{n2} \end{bmatrix}$$

Note that $x_{i1} = 1 \ \forall i$ since a SRM is composed by two regressors: a constant and a random variable.

The rank of any matrix determines how many columns are linearly independent of each other. Thus, if we want the full-column rank condition to hold, there should be K linearly independent columns. Since $K = 2$, the only way to assure $\text{rank}(\mathbf{X}) = 2$ is that each column is different from each other, i.e: $x_{i2} \neq x_{j2}$

6. We prove that A.2: $\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ and A.4: $\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}$ and $\mathrm{E}[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0$ imply:

$$\begin{aligned} \mathrm{V}[\varepsilon_i] &= \sigma^2 \;\; \forall i \\ \text{cov}(\varepsilon_i, \varepsilon_j) &= 0 \quad \forall i \neq j \end{aligned}$$

$$\begin{aligned} \mathrm{E}[\varepsilon_i] &= \mathrm{E}[\mathrm{E}[\varepsilon_i|\mathbf{X}]] \quad \big(\text{by LIE}\big) \\ &= 0 \quad \big(\text{by A.2}\big) \\ \mathrm{E}[x_{jk}\varepsilon] &= \mathrm{E}[\mathrm{E}[x_{jk}\varepsilon_i|x_{jk}]] \quad \big(\text{by LIE}\big) \\ &= \mathrm{E}[x_{jk}\mathrm{E}[\varepsilon_i|x_{jk}]] \quad \big(\text{by linearity of conditional expectations}\big) \\ &= 0 \quad \big(\text{by A.2}\big) \end{aligned}$$

Thus:

$$\begin{aligned} \mathrm{V}[\varepsilon_i] &= \mathrm{E}[\varepsilon_i^2] - \mathrm{E}[\varepsilon_i]^2 \\ &= \mathrm{E}[\varepsilon_i^2] \quad \big(\text{following our results}\big) \\ &= \mathrm{E}[\mathrm{E}[\varepsilon_i^2|\mathbf{X}]] \\ &= \sigma^2 \quad \big(\text{by A.4}\big) \\ \text{cov}(\varepsilon_i, \varepsilon_j) &= \mathrm{E}[\varepsilon_i\varepsilon_j] - \mathrm{E}[\varepsilon_i]\mathrm{E}[\varepsilon_j] \\ &= \mathrm{E}[\varepsilon_i\varepsilon_j] \quad \big(\text{following our results}\big) \\ &= \mathrm{E}[\mathrm{E}[\varepsilon_i\varepsilon_j|\mathbf{X}]] \\ &= 0 \quad \big(\text{by A.4 if } i \neq j\big) \end{aligned}$$

1.2 The Algebra of Least Squares

1. We prove $\mathbf{X}'\mathbf{X}$ is Positive Definite if $\mathbf{X}$ is of full-column rank.

Let $\dim(\mathbf{X}) = n \,\mathrm{x}\, K$ and $\mathbf{Xc} = \mathbf{z}$, we need to show that $\mathbf{c}'\mathbf{X}'\mathbf{Xc} > 0 \;\; \forall \mathbf{c} \neq 0$

$$\begin{aligned}\mathbf{c}'\mathbf{X}'\mathbf{Xc} &= (\mathbf{Xc})'\mathbf{Xc} \\ &= \mathbf{z}'\mathbf{z} \\ &= \sum_{i=1}^{n} z_i^2\end{aligned}$$

By construction: $\sum_{i=1}^{n} z_i^2 \geq 0$ but note that if $\Pr(\mathrm{rank}(\mathbf{X}) = K) = 1$ then $\sum_{i=1}^{n} z_i^2 > 0$

Thus, if $\mathbf{X}$ is of full-column rank, $\mathbf{X}'\mathbf{X}$ is Positive Definite.

2. We verify the following equalities:

$$\mathbf{X}'\mathbf{X}/n = \frac{1}{n}\sum_{i}^{n} \mathbf{x}_i\mathbf{x}_i' \qquad \mathbf{X}'\mathbf{y}/n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i y_i$$

$$\begin{aligned}\mathbf{X}'\mathbf{X}/n &= \frac{1}{n}\begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix}\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \dots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \dots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \dots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}\end{aligned}$$

Where we leveraged that $x_i' = x_i$ for scalars. Also, note that:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix}$$

Thus:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \dots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \dots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \dots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}$$

Consequently, $\mathbf{X}'\mathbf{X}/n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$

$$\begin{aligned} \mathbf{X}'\mathbf{y}/n &= \frac{1}{n}\begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i \end{aligned}$$

3. For the Simple Regression Model ($K = 2$, $x_{i1} = 1\ \forall i$) let us prove that:

$$\mathbf{S_{xx}} = \begin{bmatrix} 1 & \bar{x}_2 \\ \bar{x}_2 & \frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 \end{bmatrix} \qquad \mathbf{S_{xy}} = \begin{bmatrix} \bar{y} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2}y_i \end{bmatrix}$$

$$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2} \qquad b_1 = \bar{y} - \bar{x}_2 b_2$$

Where

$$\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{x}_2 := \frac{1}{n}\sum_{i=1}^{n} x_{i2}$$

$$\mathbf{S}_{xx} = \mathbf{X}'\mathbf{X}/n = \frac{1}{n}\begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{bmatrix}\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}$$

$$= \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 \end{bmatrix} = \begin{bmatrix} 1 & \bar{x}_2 \\ \bar{x}_2 & \frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 \end{bmatrix} \quad (\text{since} \quad x_{i1} = 1\ \forall i)$$

$$\mathbf{S}_{xy} = \mathbf{X}'\mathbf{y}/n = \frac{1}{n}\begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} x_{i1}y_i \\ \sum_{i=1}^{n} x_{i2}y_i \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2}y_i \end{bmatrix}$$

By minimizing the objective function $\mathbf{e}'\mathbf{e}$ we obtain $\mathbf{b} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xy}}$ and applying brute force matrix inversion:

$$\mathbf{S}_{xx}^{-1} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 - \bar{x}_2^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{bmatrix}$$

We will leverage the following equality to derive $\mathbf{b}$:

$$\frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 - \bar{x}_2^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2$$

Thus:

$$\begin{aligned} \mathbf{b} &= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{bmatrix} \begin{bmatrix} \bar{y} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2}y_i \end{bmatrix} \\ &= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2} \begin{bmatrix} \bar{y}\frac{1}{n}\sum_{i=1}^{n} x_{i2}^2 - \bar{x}_2\frac{1}{n}\sum_{i=1}^{n} x_{i2}y_i \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2}y_i - \bar{x}_2\bar{y} \end{bmatrix} \end{aligned}$$

Notice that the estimator vector $\mathbf{b}$ consists of two elements, b_1 and b_2:

$$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{i2}y_i - \bar{x}_2\bar{y}}{\frac{1}{n}\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}$$

Since:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)(y_i - \bar{y}) &= \frac{1}{n}\sum_{i=1}^{n}x_{i2}y_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n}x_{i2} - \bar{x}_2\frac{1}{n}\sum_{i=1}^{n}y_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}_2\bar{y} \\ &= \frac{1}{n}\sum_{i=1}^{n}x_{i2}y_i - \bar{y}\bar{x}_2 - \bar{x}_2\bar{y} + \bar{x}_2\bar{y} = \frac{1}{n}\sum_{i=1}^{n}x_{i2}y_i - \bar{x}_2\bar{y} \end{aligned}$$

$$
\begin{aligned}
b_1 &= \frac{\bar{y}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(x_{i2}^2 - 2^2 \bar{+} \bar{x}_2^2\right) - \bar{x}_2\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{i2}y_i}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2} \\
&= \bar{y} - \bar{x}_2\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{i2}y_i}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2} - \bar{y}\bar{x}_2^2 = \bar{y} - \bar{x}_2 \underbrace{\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{i2}y_i - \bar{x}_2\bar{y}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}}_{b_2} \\
&= \bar{y} - \bar{x}_2 b_2
\end{aligned}
$$

4. Let the projection matrix be denoted as $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the annihilator matrix as $\mathbf{M} = (\mathbf{I}_n - \mathbf{P})$, we prove they are both symmetric and idempotent.

Note that a symmetric matrix $\mathbf{A}$ satisfies $\mathbf{A} = \mathbf{A}'$ and a matrix $\mathbf{C}$ is said to be idempotent if $\mathbf{C} = \mathbf{C}^2$

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

5. Let us prove some matrix algebra properties of fitted values and residuals:

i) $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$

$$
\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \quad \big(\text{by definition of } \hat{\mathbf{y}}\big) \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \big(\text{since } \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\big) \\
&= \mathbf{P}\hat{\mathbf{y}}
\end{aligned}
$$

ii) $\mathbf{e} = \mathbf{M}\mathbf{y}$ and $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$

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

iii) $\text{SSR} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$

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

6. R^2 is not altered by changes in the unit of measurement of ($\mathbf{y}$, $\mathbf{X}$):

$$R^2 = \frac{\displaystyle\sum_{i=1}^{n}(\mathbf{k}\hat{y}_i - \mathbf{k}\bar{y})^2}{\displaystyle\sum_{i=1}^{n}(\mathbf{k}y_i - \mathbf{k}\bar{y})^2}$$

$$= \frac{\mathbf{k}}{\mathbf{k}}\frac{\displaystyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In the case of regressors, let us study whether $\mathbf{b}$ is modified or not:

$$\begin{aligned}\mathbf{b} &= (\mathbf{k}\mathbf{X}'\mathbf{k}\mathbf{X})^{-1}\mathbf{k}\mathbf{X}'\mathbf{y} \\ &= [\mathbf{k}(\mathbf{X}'\mathbf{X})]^{-1}\mathbf{k}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

7. Given that $R^2_{uc} = 1 - \dfrac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}}$ and $R^2 = 1 - \dfrac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

We prove the following relationship:

$$1 - R^2 = \left(1 + \frac{n\bar{y}^2}{\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2}\right)(1 - R^2_{uc})$$

$$1 - R^2 = \frac{\displaystyle\sum_{i=1}^{n} e_i^2}{\displaystyle\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Note that:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}y_i^2+\sum_{i=1}^{n}\bar{y}^2-2\sum_{i=1}^{n}y_i\bar{y}\\
&= \sum_{i=1}^{n}y_i^2+n\bar{y}^2-2\bar{y}\sum_{i=1}^{n}y_i\\
&= \sum_{i=1}^{n}y_i^2+n\bar{y}^2-2n\bar{y}^2\\
&= \sum_{i=1}^{n}y_i^2-n\bar{y}^2
\end{aligned}$$

Then:

$$\begin{aligned}
1-R^2 &= \frac{\displaystyle\sum_{i=1}^{n}e_i^2}{\displaystyle\sum_{i=1}^{n}y_i^2-n\bar{y}^2}=\frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}-n\bar{y}^2}\\
&= \frac{(1-R_{uc}^2)\mathbf{y}'\mathbf{y}}{\mathbf{y}'\mathbf{y}-n\bar{y}^2}\quad \text{(since } \mathbf{e}'\mathbf{e}=(1-R_{uc}^2)\mathbf{y}'\mathbf{y})\\
&= \frac{\mathbf{y}'\mathbf{y}}{\mathbf{y}'\mathbf{y}-n\bar{y}^2}(1-R_{uc}^2)\\
&= \frac{\displaystyle\sum_{i=1}^{n}y_i^2}{\displaystyle\sum_{i=1}^{n}y_i^2-n\bar{y}^2}(1-R_{uc}^2)\\
&= \frac{\sum_{i=1}^{n}y_i^2-n\bar{y}^2+n\bar{y}^2}{\sum_{i=1}^{n}y_i^2-n\bar{y}^2}(1-R_{uc}^2)
\end{aligned}$$

Notice that we applied the add-and-subtract strategy and since:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

We finally obtain:

$$1 - R^2 = \left(1 + \frac{n\bar{y}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \right) (1 - R_{uc}^2)$$

8. Let us show that $R_{uc}^2 = \dfrac{\mathbf{y}'\mathbf{P}\mathbf{y}}{\mathbf{y}'\mathbf{y}}$

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{P}\mathbf{y} \end{aligned}$$

Since $\mathbf{P}$ is symmetric and idempotent:

$$\mathbf{y}'\mathbf{P}\mathbf{y} = \mathbf{y}'\mathbf{P}'\mathbf{P}\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}}$$

Then:

$$R_{uc}^2 = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{\mathbf{y}'\mathbf{y}}$$

9. Proof that regression coefficients and related statistics can be calculated from sample averages $\mathbf{S_{xx}}, \mathbf{S_{xy}}, \bar{y}, \mathbf{y}'\mathbf{y}/n$. Thus, they only need to be computed once.

$$\begin{aligned} \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (n^{-1}\mathbf{X}'\mathbf{X})^{-1}n^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{S_{xx}^{-1}}\mathbf{S_{xy}} \end{aligned}$$

$$
\begin{aligned}
SSR &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - (\mathbf{X}\mathbf{b})'\mathbf{y} + (\mathbf{X}\mathbf{b})'(\mathbf{X}\mathbf{b}) \\
&= \mathbf{y}'\mathbf{y}/n - 2\mathbf{b}\mathbf{X}'\mathbf{y}/n + \mathbf{b}'\mathbf{X}'\mathbf{X}/n\mathbf{b} \\
&= \mathbf{y}'\mathbf{y}/n - 2\mathbf{b}\mathbf{S}_{\mathbf{xy}} + \mathbf{b}'\mathbf{S}_{\mathbf{xx}}\mathbf{b}
\end{aligned}
$$

$$
\begin{aligned}
S^2 &= \frac{SSR}{n-k} \\
&= \frac{\mathbf{y}'\mathbf{y}/n - 2\mathbf{b}\mathbf{S}_{\mathbf{xy}} + \mathbf{b}'\mathbf{S}_{\mathbf{xx}}\mathbf{b}}{n-k}
\end{aligned}
$$

$$
\begin{aligned}
R^2 &= 1 - \frac{SSR}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= 1 - \frac{\mathbf{y}'\mathbf{y}/n - 2\mathbf{b}\mathbf{S}_{\mathbf{xy}} + \mathbf{b}'\mathbf{S}_{\mathbf{xx}}\mathbf{b}}{\mathbf{y}'\mathbf{y}/n - \bar{y}^2}
\end{aligned}
$$

1.3 Finite Sample Properties of OLS

1. We prove the Gauss-Markov Theorem. Additionally, we prove the adequacy of S^2 as an estimator for σ^2. Our Linear Regression Model satisfies the following assumptions:

- A.1 (Linearity):

$$\mathrm{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- A.2 (Strict Exogeneity):

$$\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$$

- A.3 (Rank Condition):

$$\Pr(\mathrm{rank}(\mathbf{X}) = K) = 1$$

- A.4 (Spherical Error Variance):

$$\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I} \qquad \mathrm{E}[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0 \quad \text{if } i \neq j$$

i) The OLS estimator $\mathbf{b}$ is unbiased:

$$\begin{aligned}
\mathrm{E}[\mathbf{b}|\mathbf{X}] &= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \quad \big(\text{by minimizing SSR}\big) \\
&= \mathrm{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] \quad \big(\text{by A.1}\big) \\
&= \mathrm{E}[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \boldsymbol{\beta} + \mathrm{E}[\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] \quad \big(\mathbf{A} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big) \\
&= \boldsymbol{\beta} + \mathbf{A}\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \quad \big(\text{by linearity of conditional expectations}\big) \\
&= \boldsymbol{\beta} \quad \big(\text{by A.2}\big)
\end{aligned}$$

Note that $\mathbf{X}'\mathbf{X}$ is invertible by A.3, so all previous results hold. Now let us focus on the variance of $\mathbf{b}$:

$$\begin{aligned}
\mathrm{V}[\mathbf{b}|\mathbf{X}] &= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \\
&= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] \quad \big(\text{by A.1}\big) \\
&= \mathrm{V}[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \mathrm{V}[\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}] \quad \big(\text{as } \boldsymbol{\beta} \text{ is a fixed parameter}\big) \\
&= \mathbf{A}\mathrm{V}[\boldsymbol{\varepsilon}|\mathbf{X}]\mathbf{A}' \quad \big(\text{by properties of the variance}\big) \\
&= \mathbf{A}\sigma^2\mathbf{I}\mathbf{A}' \quad \big(\text{by A.4}\big) \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Where we claimed $\boldsymbol{\beta}$ is a fixed parameter as we are taking a Frequentist Approach (may Bayesians forgive me).

ii) The OLS estimator $\mathbf{b}$ is efficient.

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 $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
Then $\hat{\boldsymbol{\beta}} = (\mathbf{D} + \mathbf{A})\mathbf{y}$

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

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

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

Deriving the sampling error:

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

Then:

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

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

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

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 $\mathbf{DD} \geq 0$. 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.

iii) S^2 is an 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

$$SSR/n - k = \mathbf{e}'\mathbf{e}/n - k$$

We proved $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ earlier, so:

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

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.

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

Thus:

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

We have just proved that the estimator S^2 is unbiased. Note that dividing SSR by the degrees of freedom $n-k$ was imperative to achieve unbiasedness.

2. We proved the Gauss-Markov Theorem, which establishes that **b** is unbiased and efficient in the class of linear unbiased estimators. This does not entail that any other linear estimator is excluded from unbiasedness. Let us show an example for the following Linear Regression model satisfying the previously stated assumptions:

$$CON_i = \beta_1 + \beta_2 YD_i + \varepsilon_i$$

Let our synthetic estimator be denoted as:

$$\hat{\beta}_2 = \tfrac{CON_2 - CON_1}{YD_2 - YD_1}$$

We proceed to show it is unbiased:

$$\mathrm{E}[\hat{\beta}_2|\mathbf{YD}] = \mathrm{E}\left[\frac{\beta_1 + \beta_2 YD_2 + \varepsilon_2 - \beta_1 - \beta_2 YD_1 - \varepsilon_1}{YD_2 - YD_1}|\mathbf{YD}\right]$$

$$= \frac{1}{YD_2 - YD_1}\mathrm{E}\left[\beta_2 YD_2 - \beta_2 YD_1 + \varepsilon_2 - \varepsilon_1|\,\mathbf{YD}\right]$$

$$= \frac{1}{YD_2 - YD_1}\mathrm{E}[\beta_2(YD_2 - YD_1)|\mathbf{YD}] + \mathrm{E}[\varepsilon_2 - \varepsilon_1|\mathbf{YD}]$$

$$= \beta_2\left(\frac{YD_2 - YD_1}{YD_2 - YD_1}\right) + \frac{1}{YD_2 - YD_1}\left(\mathrm{E}[\varepsilon_2|\mathbf{YD}] - \mathrm{E}[\varepsilon_1|\mathbf{YD}]\right)$$

$$= \beta_2 \quad \big(\text{by A.2: Strict Exogeneity Assumption}\big)$$

3. Likewise, there exist linear but not necessarily unbiased estimators that have smaller variance than the OLS estimator **b**. We present an example here:

Let $\hat{\beta} = c$, where c represents any real number (and thus a constant). Unless $\beta = c$, our estimator $\hat{\beta}$ is badly biased. However, note that $\mathrm{V}[\hat{\beta}] = 0$ by construction, as c lacks any sort of stochasticity.

Clearly: $\mathrm{V}[\mathbf{b}] > \mathrm{V}[\hat{\beta}]$

The purpose of questions **2)** and **3)** is to show that unbiasedness and low variance are not sufficient properties by themselves to determine whether an estimator is good or not, as well as to clarify what the Gauss-Markov Theorem really means.

4. We prove the following equality for the unconditional variance:

$$\mathrm{V}[\hat{\boldsymbol{\beta}}] = \mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] + \mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right]$$

Using the add-and-subtract strategy and defining:

$$\mathbf{d} := \hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] \qquad \mathbf{a} := \hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}] \qquad \mathbf{c} := \mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathrm{E}[\hat{\boldsymbol{\beta}}]$$

Then

$$\mathbf{d} = \mathbf{a} - \mathbf{c} \qquad \mathrm{E}[\mathbf{d}\mathbf{d}'] = \mathrm{E}[(\mathbf{a}-\mathbf{c})(\mathbf{a}-\mathbf{c})'] = \mathrm{E}[\mathbf{a}\mathbf{a}'] - \mathrm{E}[\mathbf{c}\mathbf{a}'] - \mathrm{E}[\mathbf{a}\mathbf{c}'] + \mathrm{E}[\mathbf{c}\mathbf{c}']$$

By definition:

$$\mathrm{V}[\mathbf{x}] = \sum_{i=1}^{n} (x_i - \mathrm{E}[\mathbf{x}])^2 = (\mathbf{x} - \mathrm{E}[\mathbf{x}])(\mathbf{x} - \mathrm{E}[\mathbf{x}])'$$

Following the previous expression:

$$\mathrm{E}[\mathbf{a}\mathbf{a}'] = \mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)\left(\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)'\right] = \mathrm{V}[\hat{\boldsymbol{\beta}}]$$

$$\begin{aligned}
\mathrm{E}[\mathbf{c}\mathbf{c}'] &= \mathrm{E}\left[\left(\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)\left(\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)'\right] \\
&= \mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] \quad \left(\text{since } \mathrm{E}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] = \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[\mathbf{c}\mathbf{a}'] &= \mathrm{E}[\mathrm{E}[\mathbf{c}\mathbf{a}'|\mathbf{X}]] \quad (\text{by LIE}) \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)\left(\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)' | \,\mathbf{X}\right]\right] \\
&= \mathrm{E}\left[\left(\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)\left(\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] - \mathrm{E}[\hat{\boldsymbol{\beta}}]\right)'\right] \\
&= \mathrm{E}\left[\mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right]\right] = \mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right]
\end{aligned}$$

This last equality holds since

$$\begin{aligned} \mathrm{E}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]|\mathbf{X}\right] &= \mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] \\ \mathrm{E}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] &= \mathrm{E}[\hat{\boldsymbol{\beta}}] \end{aligned}$$

Note that $\mathrm{E}[\mathbf{ca}'] = \mathrm{E}[\mathbf{ac}']$. Thus:

$$\begin{aligned} \mathrm{E}[\mathbf{dd}'] &= \mathrm{E}\left[\mathrm{E}[\mathbf{dd}'|\mathbf{X}]\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\left(\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right)\left(\hat{\boldsymbol{\beta}} - \mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right)'|\mathbf{X}\right]\right] \\ &= \mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] \end{aligned}$$

Finally:

$$\begin{aligned} \mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] &= \mathrm{V}[\hat{\boldsymbol{\beta}}] - \mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] \\ \mathrm{V}[\hat{\boldsymbol{\beta}}] &= \mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] + \mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] \end{aligned}$$

Leveraging this derivation we show that $\mathrm{V}[\hat{\boldsymbol{\beta}}] \geq \mathrm{V}[\mathbf{b}]$

Note that

$$\mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] = \mathrm{V}\left[\mathrm{E}[\mathbf{b}|\mathbf{X}]\right] = \mathrm{V}[\boldsymbol{\beta}] = 0$$

So

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

Clearly:

$$\mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] \geq \mathrm{E}\left[\mathrm{V}[\mathbf{b}|\mathbf{X}]\right] \implies \mathrm{V}[\hat{\boldsymbol{\beta}}] \geq \mathrm{V}[\mathbf{b}]$$

5. It was previously noted that dividing by the degrees of freedom actually contributed to obtaining an unbiased estimator for σ^2. Think of this procedure as an operation similar to Bessel's correction. We now prove that it would not be necessary if we had data on $\boldsymbol{\varepsilon}$, i.e, the error term were observable.

Let our theoretical estimator be denoted as:

$$\Psi := \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n}$$

Then

$$\begin{aligned}
\mathrm{E}[\Psi|\mathbf{X}] = \mathrm{E}\left[\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n}|\mathbf{X}\right] &= \frac{1}{n}\mathrm{E}[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\varepsilon_i^2|\mathbf{X}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\varepsilon_i^2|\mathbf{X}] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 \quad \big(\text{by A.4}\big) \\
&= \frac{n}{n}\sigma^2 = \sigma^2
\end{aligned}$$

7. Let us prove that the i-th element of the matrix $\mathbf{P}$ is bounded: $p_i \in [0,1]$

$$p_i := \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

Since trace($\mathbf{P}$) $= K$ then $\sum_{i=1}^{n} p_i = K$ so it must be that $p_i \geq 0\ \forall i$ as the diagonal of $\mathbf{P}$ is nonnegative ($\mathbf{P}$ is Positive Semidefinite).

Let $\mathbf{e}_i$ be a vector containing the i-th residual so $\mathbf{e}_{-i} = 0$. Then $p_i = \mathbf{e}_i'\mathbf{P}\mathbf{e}_i$

$$\begin{aligned}
p_i &= \mathbf{e}_i'\mathbf{P}\mathbf{e}_i \\
&= \mathbf{e}_i'(\mathbf{I}_n - \mathbf{M})\mathbf{e}_i \quad (\text{since } \mathbf{M} = \mathbf{I}_n - \mathbf{P}) \\
&= \mathbf{e}_i'\mathbf{e}_i - \mathbf{e}_i'\mathbf{M}\mathbf{e}_i \\
&= 1 - \mathbf{e}_i'\mathbf{M}\mathbf{e}_i \\
p_i &\leq 1 \quad (\text{since } \mathbf{M} \text{ is Positive Semidefinite})
\end{aligned}$$

6. Under the previously stated assumptions we proceed to prove that the OLS estimator $\mathbf{b}$ and the residual vector $\mathbf{e}$ are conditionally uncorrelated.

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

$$\begin{aligned}
\mathbf{b}-\mathrm{E}[\mathbf{b}|\mathbf{X}] &= \mathbf{b}-\boldsymbol{\beta} \quad \big(\text{since } \mathbf{b} \text{ is unbiased}\big)\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}-\boldsymbol{\beta}\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\beta \quad \big(\text{by A.1}\big)\\
&= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta}\\
&= \mathbf{A}\boldsymbol{\varepsilon} \quad \big(\mathbf{A} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\big)\\
\mathbf{e}-\mathrm{E}[\mathbf{e}|\mathbf{X}] &= \mathbf{b}-\mathrm{E}[\mathbf{y}-\mathbf{X}\mathbf{b}|\mathbf{X}]\\
&= \mathbf{e}-\mathrm{E}[\mathbf{y}|\mathbf{X}]+\mathrm{E}[\mathbf{X}\mathbf{b}|\mathbf{X}]\\
&= \mathbf{e}-\mathrm{E}[\mathbf{X},\boldsymbol{\beta}+\boldsymbol{\varepsilon}|\mathbf{X}]+\mathbf{X}\mathrm{E}[\mathbf{b}|\mathbf{X}]\\
&= \mathbf{e}-\mathbf{X}\boldsymbol{\beta}+\mathrm{E}[\varepsilon|\mathbf{X}]+\mathbf{X}\boldsymbol{\beta}\\
&= \mathbf{e} \quad \big(\text{by A.2}\big)
\end{aligned}$$

Then

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

This last equality holds since:

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

1.4 Hypothesis Testing Under Normality

1. From now on we introduce an additional assumption to our Finite Sample Linear Regression Model:

- A.5 (Normality):

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\vec{0}, \sigma^2 \mathbf{I}_n)$$

Meaning that the distribution of the error term conditional on $\mathbf{X}$ is jointly normal. Furthermore, the marginal distribution of $\boldsymbol{\varepsilon}$ is also normal since both, the mean and variance do not depend on the data. We now derive the distribution of $\mathbf{b}$ conditional on $\mathbf{X}$ and analyze whether its marginal distribution can be said to be normal or not.

Firstly, since we only make assumptions about the distribution of $\boldsymbol{\varepsilon}|\mathbf{X}$, we need to find a way to express $\mathbf{b}|\mathbf{X}$ as a function of the former.

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \quad (\text{by A.1}) \\ &= \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} \quad (\mathbf{A} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\end{aligned}$$

Since the only stochastic component of $\mathbf{bX}$ is $\boldsymbol{\varepsilon}$, then $\mathbf{b}|\mathbf{X}$ is also jointly normal. Leveraging that we already derived the mean and variance of the OLS estimator $\mathbf{b}$:

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

Clearly, the variance of the OLS estimator $\mathbf{b}$ depends on $\mathbf{X}$, and thus on the feature data matrix. Consequently, we cannot claim that the marginal distribution of $\mathbf{b}$ is normal. Note, however, that statistics z_k, t_k, F are independently distributed of $\mathbf{X}$. This was crucial in hypothesis testing as only one probability density table had to be calculated. Nowadays, statistical software can easily derive probability density and distribution functions for non-standardized random variables. Think of `pnorm()` in the case of **R** or `scipy.stats.norm.cdf()` in **Python**.

$$z_k|\mathbf{X} = \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_{kk}}} \sim N(0,1) \quad t_k|\mathbf{X} = \frac{b_k - \bar{\beta}_k}{\sqrt{S^2\mathbf{X})_{kk}}} \sim t_{n-k} \quad F|\mathbf{X} = \frac{w/\#r}{q/n-k} \sim F(\#r, n-k)$$

2. It was already verified that $\{\mathbf{b}, SSR, S^2, R^2\}$ can be calculated from sample averages $\{\mathbf{S_{xx}}, \mathbf{S_{xy}}, \mathbf{y}'\mathbf{y}/n, \bar{y}\}$ so let us show that we can also express $\text{SE}(b_k)$ with the former set.

$$\text{SE}(b_k) = \sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n}(\mathbf{X}'\mathbf{X})^{-1}}$$

Note that we already showed that

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y}/n - 2\tilde{\boldsymbol{\beta}}'\frac{\mathbf{X}'\mathbf{y}}{n}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\frac{\mathbf{X}'\mathbf{X}}{n}\tilde{\boldsymbol{\beta}}$$

Thus, $\text{SE}(b_k)$ can be calculated from the aforementioned sample averages.

3. Under the Classical Assumptions we have been using throughout the entirety of this paper, let us prove that the matrix $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is Positive Definite.

We already showed that $(\mathbf{Xc})'\mathbf{Xc} > \mathbf{0}\ \ \forall \mathbf{c} \neq \mathbf{0}$ since by A.3: $\Pr(\text{rank}(\mathbf{X}) = K) = 1$

Note that $\mathbf{R}$ is of full row rank since we impose non-linearly dependent restrictions, so $\Pr(\text{rank}(\mathbf{R}) = \#r) = 1$. Let $\mathbf{Rc} = \mathbf{z}'$:

$$\mathbf{Rc}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{Rc})' = \mathbf{z}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z} > 0$$

4. We show that the size of the following one-tailed test is α:

$$\begin{aligned} &\text{H}_0\text{: } \beta_k = \bar{\beta}_k \\ &\text{H}_1\text{: } \beta_k > \bar{\beta}_k \end{aligned}$$

Let $\Pr(t_k > t_\alpha) = \alpha$ and the Rejection Region be defined as RR_α: $\{t_k > t_\alpha\}$

$$\begin{aligned} \text{Type I Error} &= \Pr(\text{Reject}\text{H}_0|\text{H}_1 \text{ is true}) \\ &= \Pr\left(\frac{b_k - \beta_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} > t_\alpha|\beta_k = \bar{\beta}_k\right) \quad (\text{Under H}_0) \\ &= \Pr\left(\frac{b_k - \bar{\beta}_k}{\sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} > t_\alpha|\beta_k = \bar{\beta}_k\right) = \alpha \end{aligned}$$

5. The F-statistic is represented as follows:

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

Note that if rather than testing $\#r$ restrictions we limit ourselves to just $\#r = 1$ then:

$$F = (\mathbf{b}_k - \bar{\beta}_k)'[S^2(\mathbf{X}'\mathbf{X})^{-1\prime}_{kk}]^{-1}(\mathbf{b}_k - \bar{\beta}_r)$$

Since $\mathbf{R} = [0 \quad 0 \quad \dots \underbrace{1}_{\text{Position } k} \dots \quad 0 \quad 0]$ and $\mathbf{r} = \bar{\beta}_k$

So $F(1, n-k) = t^2_{n-k}$

6. In order to test multiple hypothesis the F-statistic is preferred over several t-statistics as the former does adjust the critical value so that the size of the test remains unchanged. Joint hypothesis testing with the latter results in $\tilde{\alpha} \neq \alpha$

7. We derive the variance of the random variable S^2:

$$\begin{aligned}
\mathrm{V}[S^2|\mathbf{X}] &= \mathrm{V}\left[\frac{\mathbf{e}'\mathbf{e}}{n-k}|\mathbf{X}\right] \\
&= \left(\frac{1}{n-k}\right)^2 \mathrm{V}[\mathbf{e}'\mathbf{e}|\mathbf{X}] \\
&= \left(\frac{1}{n-k}\right)^2 \mathrm{V}[\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}|\mathbf{X}] \\
&= \left(\frac{1}{n-k}\right)^2 \mathrm{V}\left[\frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{M}\frac{\boldsymbol{\varepsilon}}{\sigma}\sigma^2|\mathbf{X}\right] \\
&= \left(\frac{1}{n-k}\right)^2 2(n-k)\sigma^4 \\
&= \frac{2\sigma^4}{n-k}
\end{aligned}$$

By A.5 the error term is normally distributed and dividing by σ standardizes it. Since $\mathbf{Y}^2 = \mathbf{x}'\mathbf{A}\mathbf{x}$ with $\mathbf{x} \sim N(\vec{0}, \vec{1})$ then $\mathbf{Y}^2 \sim \chi^2_m$ satisfying $\mathrm{V}[\mathbf{Y}^2] = 2m$ the same applies to $\frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{M}\frac{\boldsymbol{\varepsilon}}{\sigma}$ with $\mathrm{rank}(\mathbf{M}) = n-k$

1.5 Generalized Least Squares (GLS)

1. In most econometric applications it is not common for A.4 to hold. Thus, let us introduce an additional assumption to overcome this issue:

- A.6 (Non-Spherical Error Variance):

$$\mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{V}(\mathbf{X})$$

Where $\mathbf{V}(\mathbf{X})$ is a symmetric, nonsingular and known matrix. For now on we will denote it simply as $\mathbf{V}$ for the sake of simplicity. Also note that $\dim(\mathbf{V}) = n \,\mathrm{x}\, n$. Consequently, $\mathbf{V}$ is Positive Definite since $\mathrm{rank}(\mathbf{V}) = n$ as it is nonsingular and the lenght of its columns is identical to that of its rows.

Furthermore, let $\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}$ where $\mathbf{C}$ is a nonsingular and symetric matrix such that $\dim(\mathbf{C}) = n \,\mathrm{x}\, n$. Additionally, let us suppose that such decomposition of $\mathbf{V}$ is known. Consider a new Regression Model:

$$\begin{gathered}\mathbf{Cy} = \mathbf{CX}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\varepsilon} \\ \tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}\end{gathered}$$

Clearly, A.1 is satisfied since it is still a Linear Regression Model. Let us see if this is the case for A.2 (Strict Exogeneity):

$$\begin{aligned}\mathrm{E}[\tilde{\boldsymbol{\varepsilon}}|\mathbf{X}] &= \mathrm{E}[\mathbf{C}\boldsymbol{\varepsilon}|\mathbf{X}] \\ &= \mathbf{C}\mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] \quad \big(\text{since } \mathbf{C} \text{ is a function of } \mathbf{X}\big) \\ &= 0 \quad \big(\text{by A.2}\big)\end{aligned}$$

Since $\mathbf{C}$ is a full-rank matrix and by A.3 $\Pr(\mathrm{rank}(\mathbf{X}) = K) = 1$, entailing $n \geq K$, then $\mathrm{rank}(\mathbf{CX}) = K$ as $\mathbf{C}$ is nonsingular. A.3 does still hold. Let's check A.4:

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

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.

2. Now let us obtain the new Best Linear Unbiased Estimator: $\hat{\boldsymbol{\beta}}^{GLS}$

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

Where $\tilde{\boldsymbol{\beta}}$ is a running parameter. Note that since A.3 is satisfied, we can obtain a closed-form solution. Before minimizing the objective function $S\tilde{S}R$, let us write down imperative linear algebra derivate 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

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

Thus

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

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

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

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

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

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

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

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

2 Analytical Exercises

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

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

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

2. Let us denote the annihilator associated with the vector of ones ($\mathbf{1}$) as:

$$\mathbf{M}_1 \coloneqq \mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$$

i) $\mathbf{M}_1$ is symmetric and idempotent

$$\begin{aligned}
\mathbf{M}_1 &= (\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')' \\
&= \mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' \quad \big(\text{since } \mathbf{I}_n' = \mathbf{I}_n \text{ and } (\mathbf{AB})' = \mathbf{B}'\mathbf{A}'\big) \\
\mathbf{M}_1'\mathbf{M}_1 &= (\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')'(\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}') \\
&= (\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}') - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} + \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\underbrace{\mathbf{1}'\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}}_{\mathbf{I}_n}\mathbf{1}' \\
&= \mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'
\end{aligned}$$

ii) $\mathbf{M}_1\mathbf{1} = 0$

$$\begin{aligned}
\mathbf{M}_1\mathbf{1} &= (\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')\mathbf{1} \\
&= \mathbf{1} - \mathbf{1}\underbrace{(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{1}}_{\mathbf{I}_n} \\
&= 0
\end{aligned}$$

iii) $\mathbf{M}_1\mathbf{y} = \mathbf{y} - \bar{y}\,\mathbf{1}$

$$\begin{aligned}\mathbf{M}_1\mathbf{y} &= (\mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}')\mathbf{y} \\ &= \mathbf{y} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y}\end{aligned}$$

Note that

$$\begin{aligned}\mathbf{1}'\mathbf{1} &= \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} \\ &= \sum_{i=1}^{n} 1 = n\end{aligned}$$

$$\begin{aligned}\mathbf{1}'\mathbf{y} &= \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix} \\ &= \sum_{i=1}^{n} y_i\end{aligned}$$

$$\mathbf{M}_1\mathbf{y} = \mathbf{y} - \bar{y}\,\mathbf{1} \quad (\text{since } \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i)$$

iv) $\mathbf{M}_1\mathbf{X} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$

Intuition is the same as in iii). However, note that $\bar{\mathbf{x}}$ is a vector of K means since $\mathbf{X}$ is a $n \,\mathrm{x}\, K$ matrix. Just like in Multivariate Analysis, we take the mean of each column of the feature matrix. In the preceeding exercise we did not need to take this into account as $\mathbf{y}$ is the target vector with dimension $n \,\mathrm{x}\, 1$

3 & 4. Let the feature matrix $\mathbf{X}$ be partitioned as

$$\mathbf{X} = \left[\mathbf{X}_1 \vdots \mathbf{X}_2\right]$$

Where $\dim(\mathbf{X}_1) = n \,\mathrm{x}\, K_1$ and $\dim(\mathbf{X}_2) = n \,\mathrm{x}\, K_2$.

Partitioning $\boldsymbol{\beta}$ accordingly:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

Where $\boldsymbol{\beta}_1$ is a $K_1 \,\mathrm{x}\, 1$ coefficient vector and length$(\boldsymbol{\beta}_2) = K_2$

Thus, the Linear Regression Model to be estimated can be written as:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Our goal is to prove the Frisch-Waugh Theorem, which states that the OLS estimator $\mathbf{b}_2$ can be obtained by regressing the residual vector from the regression of $\mathbf{y}$ on $\mathbf{X}_1$ on the matrix of residuals from regressing $\mathbf{X}_2$ on $\mathbf{X}_1$.

In plain English, this Theorem claims that the Multiple Regression coefficient of any single feature can also be obtained by first netting out the effect of other features in the Regression Model from both, the target variable and the independent variable.

For the sake of simplicity, let us denote the aforementioned concepts with the following expressions:

$$\begin{aligned} \mathbf{P}_1 &= \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' \\ \mathbf{M}_1 &= \mathbf{I} - \mathbf{P}_1 \\ \tilde{\mathbf{X}}_2 &= \mathbf{M}_1\mathbf{X}_2 \\ \tilde{\mathbf{y}} &= \mathbf{M}_1\mathbf{y} \end{aligned}$$

Note that all previous finite-sample assumptions that were made throughout the entirety of this paper are also satisfied for the present problem. In addition, since we proved the properties of the Projection Matrix and the Annihilator Matrix we will apply them without providing explicit demonstration so as to eschew redundancy.

We first proceed to deriving the Normal Equations for our Linear Regression Model and then pointing out relevant conclusions that can be extracted from our operations.

Since all of the classical assumptions hold, we can retrieve a closed-form solution by equalizing the gradient vector to zero (said gradient stems from minimizing the objective function SSR).

$$\mathbf{b} = \underset{\tilde{\boldsymbol{\beta}}_1,\, \tilde{\boldsymbol{\beta}}_2}{\arg\min} \left\{ (\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - \mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'(\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - \mathbf{X}_2\tilde{\boldsymbol{\beta}}_2) \right\}$$

Note that

$$\begin{aligned} SSR(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2) &= (\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - \mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'(\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - \mathbf{X}_2\tilde{\boldsymbol{\beta}}_2) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - \mathbf{y}'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 - (\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'\mathbf{y} + (\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 + \\ &\quad (\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 - (\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'\mathbf{y} + (\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'(\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1) + (\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - 2\mathbf{y}'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 + 2(\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 + (\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'(\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1) + \\ &\quad (\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'(\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2) \end{aligned}$$

Thus

$$\frac{\partial SSR(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2)}{\partial \tilde{\boldsymbol{\beta}}_1} = -2\mathbf{y}'\mathbf{X}_1 + 2\mathbf{X}_1'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 + \mathbf{X}_1'\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 + (\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'\mathbf{X}_1 = 0$$

$$\frac{\partial SSR(\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2)}{\partial \tilde{\boldsymbol{\beta}}_2} = -2\mathbf{y}'\mathbf{X}_2 + 2(\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)'\mathbf{X}_2 + \mathbf{X}_2'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2 + (\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2)'\mathbf{X}_2 = 0$$

Solving the equations

$$2\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 + 2\mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1 = 2\mathbf{X}_1'\mathbf{y} \implies \mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_1'\mathbf{y} \quad \text{(I)}$$

$$2\mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1 + 2\mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = 2\mathbf{X}_2'\mathbf{y} \implies \mathbf{X}_2'\mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 = \mathbf{X}_2'\mathbf{y} \quad \text{(II)}$$

We derived the Normal Equations related to the partitioned optimization problem.

From (I):

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2$$

Plugging it into (II):

$$\begin{aligned}
\mathbf{X}_2'\mathbf{y} &= \mathbf{X}_2'\mathbf{X}_1[\underbrace{(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2}_{\mathbf{b}_1}] + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 \\
\mathbf{X}_2'\mathbf{y} &= \mathbf{X}_2'\underbrace{\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'}_{\mathbf{P}_1}\mathbf{y} - \mathbf{X}_2'\underbrace{\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'}_{\mathbf{P}_1}\mathbf{X}_2\mathbf{b}_2 + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 \\
\mathbf{X}_2'\mathbf{y} &= \mathbf{X}_2'\mathbf{P}_1\mathbf{y} + \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 - \mathbf{X}_2'\mathbf{P}_1\mathbf{X}_2\mathbf{b}_2 \quad \text{(rearranging terms)} \\
\mathbf{X}_2'\mathbf{y} &= \mathbf{X}_2'\mathbf{P}_1\mathbf{y} + \mathbf{X}_2'(\mathbf{I} - \mathbf{P}_1)\mathbf{X}_2\mathbf{b}_2 \quad \text{(factorizing)} \\
\mathbf{X}_2'\mathbf{y} &= \mathbf{X}_2'\mathbf{P}_1\mathbf{y} + \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2 \quad (\text{since } \mathbf{M}_1 = \mathbf{I} - \mathbf{P}_1) \\
\mathbf{X}_2'\mathbf{y} - \mathbf{X}_2'\mathbf{P}_1\mathbf{y} &= \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2 \\
\mathbf{X}_2'(\mathbf{I} - \mathbf{P}_1)\mathbf{y} &= \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2 \quad \text{(factorizing)} \\
\mathbf{X}_2'\mathbf{M}_1\mathbf{y} &= \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\mathbf{b}_2 \quad (\text{by definition of } \mathbf{M}_1) \\
\mathbf{b}_2 &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} \quad (\text{since } \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2 \text{ is of full rank}) \\
\mathbf{b}_2 &= [(\mathbf{M}_1\mathbf{X}_2)'\mathbf{M}_1\mathbf{X}_2]^{-1}(\mathbf{M}_1\mathbf{X}_2)'\mathbf{M}_1\mathbf{y} \quad (\text{since } \mathbf{M}_1' = \mathbf{M}_1 \text{ and } \mathbf{M}_1 = \mathbf{M}_1^2) \\
\mathbf{b}_2 &= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}}
\end{aligned}$$

Thus, we can obtain $\mathbf{b}_2$ by regressing the residuals $\tilde{\mathbf{y}}$ on the matrix of residuals $\tilde{\mathbf{X}}_2$.

Moreover, we can show that that the residuals from regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_2$ numerically equals $\mathbf{e}$, which is the vector originated from regressing $\mathbf{y}$ on $\mathbf{X}$. Let the former Linear Regression Model estimation be expressed as follows:

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e} \\
\mathbf{M}_1\mathbf{y} &= \mathbf{M}_1\mathbf{X}_1\mathbf{b}_1 + \mathbf{M}_1\mathbf{X}_2\mathbf{b}_2 + \mathbf{M}_1\mathbf{e} \quad (\text{premultiplying by } \mathbf{M}_1) \\
\tilde{\mathbf{y}} &= \tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{M}_1\mathbf{e} \quad (\text{since } \mathbf{M}_1\mathbf{X}_1 = 0) \\
\tilde{\mathbf{y}} &= \tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{e}
\end{aligned}$$

This last equality holds since:

$$
\begin{aligned}
\mathbf{M}_1\mathbf{e} &= (\mathbf{I} - \mathbf{P}_1)\mathbf{e} \\
&= \mathbf{e} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{e} \\
&= \mathbf{e} \quad \left(\text{since } \mathbf{X}_1'\mathbf{e} = 0 \text{ by Normal Equations}\right)
\end{aligned}
$$

Note that we can also derive $\mathbf{b}_2$ by regressing $\mathbf{y}$ on the matrix of residuals $\tilde{\mathbf{X}}_2$:

$$
\begin{aligned}
\mathbf{b}_2 &= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \\
&= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}(\mathbf{M}_1\mathbf{X}_2)'\mathbf{M}_1\mathbf{y} \\
&= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{y} \quad \left(\text{since } \mathbf{M}_1' = \mathbf{M}_1\right) \\
&= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} \quad \left(\text{since } \mathbf{M}_1 = \mathbf{M}_1^2\right) \\
&= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}(\mathbf{M}_1\mathbf{X}_2)'\mathbf{y} \\
&= (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\mathbf{y}
\end{aligned}
$$

Nevertheless, the residuals from said regression are not numerically equal to the vector $\mathbf{e}$:

$$
\begin{aligned}
\mathbf{y} - \tilde{\mathbf{X}}_2\mathbf{b}_2 &= (\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \tilde{\mathbf{X}}_2\mathbf{b}_2) \quad \left(\text{add-and-subtract}\right) \\
&= (\mathbf{y} - \mathbf{M}_1\mathbf{y}) + (\hat{\mathbf{y}} - \tilde{\mathbf{X}}_2\mathbf{b}_2) \\
&= (\mathbf{y} - \mathbf{M}_1\mathbf{y}) + \mathbf{e} \\
&= \mathbf{y} - \tilde{\mathbf{y}} + \mathbf{P}_1\mathbf{y} + \mathbf{e}
\end{aligned}
$$

Since $\mathbf{P}_1\mathbf{y} \neq 0$ then $\mathbf{y} - \tilde{\mathbf{X}}_2\mathbf{b}_2 \neq \mathbf{e}$. This is also the case for $SSR(\mathbf{b}_2)$:

$$
\begin{aligned}
SSR(\mathbf{b}_2) &= (\mathbf{y} - \tilde{\mathbf{X}}_2\mathbf{b}_2)'(\mathbf{y} - \tilde{\mathbf{X}}_2\mathbf{b}_2) \\
&= (\mathbf{P}_1\mathbf{y} + \mathbf{e})'(\mathbf{P}_1\mathbf{y} + \mathbf{e}) \\
&= (\mathbf{P}_1\mathbf{y})'\mathbf{P}_1\mathbf{y} + (\mathbf{P}_1\mathbf{y})'\mathbf{e} + \mathbf{e}'\mathbf{P}_1\mathbf{y} + \mathbf{e}'\mathbf{e} \\
&= (\mathbf{P}_1\mathbf{y})'\mathbf{P}_1\mathbf{y} + \mathbf{e}'\mathbf{e} \quad \left(\text{by Normal Equations}\right) \\
&= \mathbf{y}'\mathbf{y} + \mathbf{e}'\mathbf{e} \quad \left(\text{since } \mathbf{P}_1 = \mathbf{P}_1^2\right)
\end{aligned}
$$

Leveraging our results we can show another relevant derivation concerning the squared residual vectors:

$$\begin{aligned}
\tilde{\mathbf{y}}'\tilde{\mathbf{y}} &= (\tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{e})'(\tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{e}) \\
&= \mathbf{b}_2'\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{b}_2'\tilde{\mathbf{X}}_2'\mathbf{e} + \mathbf{e}'\tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{e}'\mathbf{e} \\
&= \mathbf{b}_2'\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2\mathbf{b}_2 + \mathbf{e}'\mathbf{e} \quad \text{(by Normal Equations)} \\
\tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \mathbf{e}'\mathbf{e} &= \tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\underbrace{\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}}_{\mathbf{I}}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \quad \text{(note that } \mathbf{b}_2 = (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}}) \\
&= (\mathbf{M}_1\mathbf{y})'\mathbf{M}_1\mathbf{X}_2[(\mathbf{M}_1\mathbf{X}_2)'\mathbf{M}_1\mathbf{X}_2]^{-1}(\mathbf{M}_1\mathbf{X}_2)'\mathbf{M}_1\mathbf{y} \\
&= \tilde{\mathbf{y}}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\tilde{\mathbf{y}} \quad \text{(since } \mathbf{M}_1' = \mathbf{M}_1 \text{ \& } \mathbf{M}_1^2 = \mathbf{M}_1)
\end{aligned}$$

Consequently, were we to regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_2$ the SSR associated with the estimation of said Linear Regression Model would be none other than $\mathbf{e}'\mathbf{e}$:

$$\begin{aligned}
SSR &= (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_2\mathbf{b}_2)'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_2\mathbf{b}_2) \quad \text{(by definition of } SSR) \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2\mathbf{b}_2 - \mathbf{b}_2'\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} + \mathbf{b}_2'\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2\mathbf{b}_2 \\
&= \tilde{\mathbf{y}}'\mathbf{y} - \underbrace{\tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2\mathbf{b}_2}_{\mathbf{b}_2'\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}}} - \underbrace{\tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}}}_{\text{since } \mathbf{b}_2 = (\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}}} + \tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2\underbrace{(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2}_{\mathbf{I}}(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - 2\tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} + \tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\tilde{\mathbf{X}}_2(\tilde{\mathbf{X}}_2'\tilde{\mathbf{X}}_2)^{-1}\tilde{\mathbf{X}}_2'\tilde{\mathbf{y}} \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\tilde{\mathbf{y}} \quad \text{(like we showed above)} \\
&= \mathbf{e}'\mathbf{e} \quad \text{(since } \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \mathbf{e}'\mathbf{e} = \tilde{\mathbf{y}}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\tilde{\mathbf{y}})
\end{aligned}$$

If we regress $\tilde{\mathbf{y}}$ on $\mathbf{X}_1$ we get the following SSR:

$$\begin{aligned}
(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_1\mathbf{b}_1)'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_1\mathbf{b}_1) &= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{X}_1\mathbf{b}_1 - \mathbf{b}_1'\mathbf{X}_1'\tilde{\mathbf{y}} + \mathbf{b}_1\mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1 \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} + \tilde{\mathbf{y}}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\tilde{\mathbf{y}} \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}}
\end{aligned}$$

Note that $\tilde{\mathbf{y}} = \mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\mathbf{X}_1 = 0$ so the regression above confirms that $\mathbf{X}_1$ has no explanatory power. Indeed, we are trying to explain a target variable whose variation related with $\mathbf{X}_1$ was previously cleaned out.

If we now regress $\tilde{\mathbf{y}}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$ we can apply the already derived Frisch-Waugh Theorem to obtain $SSR = \mathbf{e}'\mathbf{e}$. The case of regressing $\tilde{\mathbf{y}}$ on $\mathbf{X}_2$ is different:

$$\begin{aligned}
(\tilde{\mathbf{y}} - \mathbf{X}_2\mathbf{b}_2)'(\tilde{\mathbf{y}} - \mathbf{X}_2\mathbf{b}_2) &= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{X}_2\mathbf{b}_2 - \mathbf{b}_2'\mathbf{X}_2'\tilde{\mathbf{y}} + \mathbf{b}_2\mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2 \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\tilde{\mathbf{y}} \\
&= \tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\tilde{\mathbf{y}}
\end{aligned}$$

5. 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:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\tilde{\beta}} \left\{ \frac{1}{2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \right\}$$
$$\text{s.t } \mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{r}$$

The Lagrangian can be formed as:

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

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

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

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

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

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

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 (proof in exercise 3 from section 1.4).

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

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

Let us derive SSR for the restricted model:

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

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:

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

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

Note the following equality from FOC (I):

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

Thus:

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

Let us come back to the F-ratio presented in exercise 1 from section 1.4:

$$
\begin{aligned}
F &= \frac{(\mathbf{Rb}-\mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{r})/\#\mathbf{r}}{S^2} \\[1em]
&= \frac{(\mathbf{Rb}-\mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{r})/\#\mathbf{r}}{\mathbf{e}'\mathbf{e}/(n-K)} \\[1em]
&= \frac{(SSR_r - SSR_u)/\#\mathbf{r}}{SSR_u/(n-K)}
\end{aligned}
$$

Since

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

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.

Notice how we applied previously introduced lineal algebra derivative properties:

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

6. Let the restricted model be a regression where the only feature is a constant and the unrestricted specification be composed of a constant and other $K - 1$ random variables. Then, in the former scenario the input data matrix and the vector of coefficients can be expressed as:

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \ \ \text{(by minimizing } SSR\text{)}$$

$$\mathbf{1}'\mathbf{1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= \sum_{i=1}^{n} 1 = n$$

$$\mathbf{1}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \sum_{i=1}^{n} y_i$$

$$(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \bar{y} \ \ \left(\text{since } \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$\hat{\boldsymbol{\beta}} = \bar{y}$$

Thus, SSR_r can be expressed as:

$$\begin{aligned}
SSR_r &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \quad \left(\text{since } \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}\right) \\
&= \mathbf{y}'\mathbf{y} - 2\bar{y}'\sum_{i=1}^{n} y_i + n\bar{y}'\bar{y} \quad \left(\text{since } \hat{\boldsymbol{\beta}} = \bar{y}\right)
\end{aligned}$$

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\
&= \sum_{i=1}^{n} y_i^2
\end{aligned}$$

$$\begin{aligned}
-2\bar{y}'\sum_{i=1}^{n} y_i &= -2\sum_{i=2}^{n} y_i\bar{y} \\
n\bar{y}'\bar{y} &= \sum_{i=1}^{n} \bar{y}^2
\end{aligned}$$

Thus:

$$\begin{aligned}
SSR_r &= \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \bar{y}^2 - 2\sum_{i=2}^{n} y_i\bar{y} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \left(\text{by notable product definition}\right)
\end{aligned}$$

Note that:

$$\begin{aligned} SSR_r - SSR_u &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned}$$

Let us also leverage the derivation from the previous exercise:

$$\begin{aligned} SSR_r - SSR_u &= (\mathbf{b} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}}) \\ &= (\mathbf{X}\mathbf{b} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}\mathbf{b} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \\ &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \end{aligned}$$

This holds since:

$$\begin{aligned} \mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{1}\bar{y} \\ &= \bar{\mathbf{y}} \\ \mathbf{X}\mathbf{b} &:= \hat{\mathbf{y}} \end{aligned}$$

So

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \end{aligned}$$

Consequently, another expression for the F-ratio under this scenario is:

$$F = \frac{(SSR_r - SSR_u)/\#\mathbf{r}}{SSR_u/(n-K)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/\#\mathbf{r}}{\sum_{i=1}^{n} e_i^2/(n-K)}$$

$$= \frac{\frac{1}{\#\mathbf{r}} \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\frac{1}{n-K} \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

Where we divided by $\sum_{i=1}^{n}(y_i - \bar{y})^2$ on both sides of the fraction and reproduced the definition of R^2:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Also note that $\#\mathbf{r}$ contains the number of restrictions, which in this case is $K-1$ as we exclude all stochastic regressors from the original specification.

Thus, we can also carry out hypothesis testing by comparing two linear regression models using their R^2 statistic.

7. Let us prove Hausman principle in Finite Sample Theory, namely, the GLS estimator is uncorrelated with the difference between the OLS estimator and itself.

$$\mathrm{E}\left[\left(\hat{\boldsymbol{\beta}}^{GLS} - \mathrm{E}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right]\right)\left(\mathbf{b} - \hat{\boldsymbol{\beta}}^{GLS} - \mathrm{E}\left[\mathbf{b} - \hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right]\right)'|\mathbf{X}\right]$$

Focusing on the Left Hand Side of the equation:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{GLS} - \mathrm{E}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right] &= \hat{\boldsymbol{\beta}}^{GLS} - \boldsymbol{\beta} \quad \left(\text{since } \hat{\boldsymbol{\beta}}^{GLS} \text{ is unbiased}\right)\\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}' - \boldsymbol{\beta}\\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta} \quad \left(\text{by A.1: Linearity}\right)\\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} - \boldsymbol{\beta}\\
&= \mathbf{D}\boldsymbol{\varepsilon} \quad (\mathbf{D} := (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})
\end{aligned}$$

Off to the RHS:

$$\begin{aligned}
\mathbf{b} - \hat{\boldsymbol{\beta}}^{GLS} - \mathrm{E}\left[\mathbf{b} - \hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right] &= \mathbf{b} - \hat{\boldsymbol{\beta}}^{GLS} \quad \left(\text{unbiasedness}\right)\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\\
&= \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} - \boldsymbol{\beta} - \mathbf{D}\boldsymbol{\varepsilon} \quad (\mathbf{A} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\\
&= (\mathbf{A} - \mathbf{D})\boldsymbol{\varepsilon}
\end{aligned}$$

Combining the results:

$$\begin{aligned}
\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}^{GLS}, \mathbf{b} - \hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right) &= \mathrm{E}\left[\mathbf{D}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{A} - \mathbf{D})'|\mathbf{X}\right]\\
&= \mathbf{D}\mathrm{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}\right](\mathbf{A} - \mathbf{D}) \quad \left(\text{by linearity of conditional expectations}\right)\\
&= \mathbf{D}\sigma^2\mathbf{V}(\mathbf{A} - \mathbf{D})' \quad \left(\text{by A.6: Non-SEV}\right)\\
&= \sigma^2\mathbf{D}\mathbf{V}(\mathbf{A} - \mathbf{D})' \quad \left(\text{since } \sigma^2 \text{ is a scalar}\right)\\
&= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\right]\\
&= \sigma^2\left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\right] = 0
\end{aligned}$$

Let $\tilde{\boldsymbol{\beta}}$ be any unbiased estimator and $\mathbf{q} \coloneqq \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{GLS}$. Notice how we set up an identical scenario to the one above if $\tilde{\boldsymbol{\beta}} = \mathbf{b}$. Furthermore, let $\mathbf{V_q} \coloneqq \mathrm{V}[\mathbf{q}|\mathbf{X}]$, where all eigenvalues of $\mathbf{V_q}$ are nonzero ($\lambda_i \neq 0 \ \forall i$) and so $\mathbf{V_q}$ is nonsingular.

Let us define $\hat{\boldsymbol{\beta}} \coloneqq \hat{\boldsymbol{\beta}}^{GLS} + \mathbf{Hq}$ for some matrix $\mathbf{H}$. We will prove $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^{GLS}, \mathbf{q}|\mathbf{X}) = 0$ by contradiction.

$$\mathrm{V}[\hat{\boldsymbol{\beta}}] = \mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] + \mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] \quad \big(\text{by LIE}\big)$$

Focusing on RHS:

$$\begin{aligned}
\mathrm{V}\left[\mathrm{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] &= \mathrm{V}\left[\mathrm{E}\left[\hat{\boldsymbol{\beta}}^{GLS} + \mathbf{Hq}|\mathbf{X}\right]\right] \\
&= \mathrm{V}\left[\boldsymbol{\beta} + \mathbf{H}\mathrm{E}\left[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right]\right] \quad \big(\text{as } \hat{\boldsymbol{\beta}}^{GLS} \text{ is unbiased}\big) \\
&= \mathrm{V}\left[\boldsymbol{\beta} + \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta})\right] \quad \big(\text{as } \tilde{\boldsymbol{\beta}} \text{ is also unbiased}\big) \\
&= \mathrm{V}[\boldsymbol{\beta}] \\
&= 0 \quad \big(\text{since Hayashi is a frequentist}\big)
\end{aligned}$$

Moving on to RHS:

$$\begin{aligned}
\mathrm{E}\left[\mathrm{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}]\right] &= \mathrm{E}\left[\mathrm{V}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right] + \mathrm{V}[\mathbf{Hq}|\mathbf{X}] + \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}^{GLS}, \mathbf{Hq}|\mathbf{X}\right) + \mathrm{Cov}\left(\mathbf{Hq}, \hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right)\right] \\
&= \mathrm{V}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right] + \mathbf{HV_qH}' + \mathbf{CH}' + \mathbf{HC}'
\end{aligned}$$

Suppose $\mathbf{H} \coloneqq -\mathbf{CV_q^{-1}}$

$$\begin{aligned}
\mathrm{V}[\hat{\boldsymbol{\beta}}] &= \mathrm{V}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right] + \mathbf{CV_q^{-1}V_qV_q^{-1}C}' - \mathbf{CV_q^{-1}C}' - \mathbf{CV_q^{-1}C}' \\
&= \mathrm{V}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right] - \mathbf{CV_q^{-1}C}'
\end{aligned}$$

Note that if $\mathbf{C} \neq 0$ then $\exists \mathbf{z}$ such that $\mathbf{v} \coloneqq \mathbf{C}'\mathbf{z}$ $(\neq 0)$

$$\mathbf{z}'\left[\mathrm{V}[\hat{\boldsymbol{\beta}}] - \mathrm{V}\left[\hat{\boldsymbol{\beta}}^{GLS}|\mathbf{X}\right]\right] = -\mathbf{v}'\mathbf{V_q^{-1}v} < 0$$

This contradicts Gauss-Markov Theorem, so $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^{GLS}, \mathbf{q}|\mathbf{X}) = 0$