Chapter 4

Ordinary Least Squares

4.1 Finite-Sample Properties of OLS

4.1.1 The Classical Linear Regression Model

Notation:

- y_i dependent variable.
- x_{ik} kth independent variable (or regressor) with $k = 1, \ldots, K$. Can be stochastic or deterministic.
- ε_i stochastic error term
- i indexes the ith individual with $i = 1, \ldots, n$, where n is the sample size

Assumption 1.1: Linearity

$$y_i = \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{4.1}$$

Usually, a constant (or intercept) is included, in this case $x_{i1} = 1$ for all i. In the following we will always assume that a constant is included in the linear model, unless otherwise stated. A special case of the above defined linear model is the so-called *simple linear model*, defined as

$$y_i \;=\; \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{4.2}$$

Often it is convenient to write Eq. (4.1) using matrix notation

$$y_i \;=\; \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. Stacking all individual rows i leads to

$$\underset{(n\times 1)}{\mathbf{y}} \;=\; \underset{(n\times K)}{\mathbf{X}}\underset{(K\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}, \tag{4.3}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{nK} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

We begin our analysis of the model in Eq. (4.3) under the framework of the so-called *classic assumptions*.

Assumption 1.2: Strict Exogeneity

$$\mathbb{E}(\varepsilon_i |\, \mathbf{X}) = 0$$

or equivalently stated for the vector $\boldsymbol{\varepsilon}$

$$\mathbb{E}(\boldsymbol{\varepsilon} \,|\, \mathbf{X}) = \mathbf{0}.$$

Notice that in the presence of a constant regressor, setting the expectation to zero is a normalization. Note that in econometrics, where we typically have to work with quasi-experimental data, strict exogeneity is a very strong assumption. It also cannot be fulfilled when the regressors include lagged dependent variables.

Some Implications of Strict Exogeneity:

- The unconditional mean of the error term is zero:

$$\mathbb{E}(\varepsilon_i) \;=\; 0 \quad (i = 1, \ldots, n) \tag{4.4}$$

Proof:
From the *Law of Total Expectations* (i.e., $\mathbb{E}(\mathbb{E}(y|\mathbf{x})) = \mathbb{E}(y)$) it follows that

$$\mathbb{E}(\varepsilon_i) = \mathbb{E}(\mathbb{E}(\varepsilon_i \,|\, \mathbf{X})).$$

The strict exogeneity assumption then yields

$$\mathbb{E}(\mathbb{E}(\varepsilon_i \,|\, \mathbf{X})) = \mathbb{E}(0) = 0. \quad \Box$$

- Generally, two random variables x and y are said to be **orthogonal** if their cross moment is zero: $\mathbb{E}(xy) = 0$. Under strict exogeneity, the regressors are orthogonal to the error term for *all* observations, i.e.,

$$\mathbb{E}(x_{jk}\,\varepsilon_i) \;=\; 0 \quad (i, j = 1, \ldots, n; k = 1, \ldots, K) \tag{4.5}$$

Proof:

$$\begin{aligned}\mathbb{E}(x_{jk}\,\varepsilon_i) \;&=\; \mathbb{E}(\mathbb{E}(x_{jk}\,\varepsilon_i\,|x_{jk})) \quad \text{(Law of Total Expect.)}\\ &=\; \mathbb{E}(x_{jk}\,\mathbb{E}(\varepsilon_i\,|x_{jk})) \;\; \text{(Linearity of } \mathbb{E}\text{-operator)}\end{aligned}$$

Now, to show that $\mathbb{E}(x_{jk}\,\varepsilon_i) = 0$, we need to show that $\mathbb{E}(\varepsilon_i\,|x_{jk}) = 0$, which is done in the following:
Since x_{jk} is an element of $\mathbf{X}$, the *Law of Iterated Expectations* (i.e., $\mathbb{E}(\mathbb{E}(y|\mathbf{x},\mathbf{z})|\mathbf{x}) = \mathbb{E}(y|\mathbf{x})$) implies that

$$\mathbb{E}(\mathbb{E}(\varepsilon_i\,|\,\mathbf{X})|x_{jk}) = \mathbb{E}(\varepsilon_i\,|x_{jk}).$$

The strict exogeneity assumption yields

$$\mathbb{E}(\mathbb{E}(\varepsilon_i\,|\,\mathbf{X})|x_{jk}) = \mathbb{E}(0|x_{jk}) = 0.$$

I.e., we have that

$$\mathbb{E}(\varepsilon_i\,|x_{jk}) = 0,$$

which allows us to conclude that

$$\mathbb{E}(x_{jk}\,\varepsilon_i) = \mathbb{E}(x_{jk}\,\mathbb{E}(\varepsilon_i\,|x_{jk})) = \mathbb{E}(x_{jk}0) = 0. \quad \Box$$

- Because the mean of the error term is zero ($\mathbb{E}(\varepsilon_i) = 0$ for all i), it follows that the orthogonality property ($\mathbb{E}(x_{jk}\,\varepsilon_i) = 0$, for all i, j, k) is equivalent to a zero-correlation property. I.e., that

$$\text{Cov}(\varepsilon_i, x_{jk}) \;=\; 0;\; i,j = 1,\ldots,n; k = 1,\ldots,K \quad (4.6)$$

Therefore, the strict exogeneity assumption implies the requirement that regressors are uncorrelated with the current ($i = j$), the past ($i < j$) and the future ($i > j$) error terms. Of course, this is usually found to be a too strong assumption - particularly in time-series contexts.

Proof:

$$\begin{aligned} \text{Cov}(\varepsilon_i, x_{jk}) &= \mathbb{E}(x_{jk}\,\varepsilon_i) - \mathbb{E}(x_{jk})\,\mathbb{E}(\varepsilon_i) \quad \text{(Def. of Cov)} \\ &= \mathbb{E}(x_{jk}\,\varepsilon_i) \quad \text{(Since } \mathbb{E}(\varepsilon_i) = 0\text{; see (4.4))} \\ &= 0 \quad \text{(Orthogonality; see (4.5))} \quad \square \end{aligned}$$

Assumption 1.3: Rank Condition

$$\text{rank}(\mathbf{X}) = K \quad \text{a.s.}$$

This assumption demands that the event of one regressor being linearly dependent on the others occurs with a probability equal to zero. (This is the literal translation of the ”almost surely (a.s.)” concept.) This assumption also implies the assumption that $n \geq K$.

This assumption is a bit dicey and its violation belongs to one of the classic problems in applied econometrics (keywords: multicollinearity, dummy variable trap, variance inflation). The violation of this assumption harms any economic interpretation as we cannot disentangle the regressors' individual effects on $\mathbf{y}$. Therefore, this assumption is often referred to as an *identification* assumption.

Assumption 1.4: Spherical Error

$$\begin{aligned} \mathbb{E}(\varepsilon_i^2\,|\mathbf{X}) &= \sigma^2 > 0 \\ \mathbb{E}(\varepsilon_i\,\varepsilon_j\,|\mathbf{X}) &= 0, \qquad i \neq j. \end{aligned}$$

Or more compactly written as,

$$\mathbb{E}(\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'\,|\mathbf{X}) = \sigma^2 I_n, \qquad \sigma^2 > 0.$$

Thus, we assume that, for a given realization of $\mathbf{X}$, the error process is uncorrelated ($\mathbb{E}(\varepsilon_i \, \varepsilon_j \, |\mathbf{X}) = 0$, for all $i \neq j$) and homoscedastic (same σ^2, for all i).

4.1.2 The Algebra of Least Squares

The OLS estimator $\mathbf{b}$ is defined as the minimizer of a specific loss function termed *the sum of squared residuals*

$$SSR(\mathbf{b}^*) \; = \; \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\mathbf{b}^*)^2 \; = \; (\mathbf{y} - \mathbf{X}\mathbf{b}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b}^*).$$

I.e., we have

$$\mathbf{b} \; := \; \arg \min_{\mathbf{b}^* \in \mathbb{R}^K} S(\mathbf{b}^*),$$

We can easily minimize $SSR(\mathbf{b}^*)$ in closed form:

$$\begin{aligned} SSR(\mathbf{b}^*) &= (\mathbf{y} - \mathbf{X}\mathbf{b}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b}^*) \\ &= \mathbf{y}'\mathbf{y} - (\mathbf{X}\mathbf{b}^*)'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b}^* + \mathbf{b}^{*'}\mathbf{X}'\mathbf{X}\mathbf{b}^* \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b}^* + \mathbf{b}^{*'}\mathbf{X}'\mathbf{X}\mathbf{b}^* \\ \Rightarrow \quad \frac{d}{d\mathbf{b}^*} SSR(\mathbf{b}^*) &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}^* \end{aligned}$$

Setting the first derivative to zero yields the so-called *normal equations*

$$\mathbf{X}'\mathbf{X}\mathbf{b} \; = \; \mathbf{X}'\mathbf{y},$$

which lead to the OLS estimator

$$\mathbf{b} \; = \; (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{4.7}$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ exists (a.s.) because of our full rank assumption (Assumption 3).

Often it is useful to express $\mathbf{b}$ (and similar other estimators) in sample moment notation:

$$\mathbf{b} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{xy}},$$

where $\mathbf{S}_{\mathbf{xx}} = n^{-1}\,\mathbf{X}'\,\mathbf{X} = n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{s}_{\mathbf{xy}} = n^{-1}\,\mathbf{X}'\,\mathbf{y} = n^{-1}\sum_i \mathbf{x}_i y_i$. This notation is more convenient for developing our large sample results.

Some quantities of interest:

- The *(OLS) fitted value*: $\hat{y}_i = \mathbf{x}_i\mathbf{b}$
  In matrix notation: $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- The *(OLS) residual*: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
  In matrix notation: $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{y}$,

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a so-called orthogonal projection matrix that projects any vector into the column space spanned by $\mathbf{X}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the associated orthogonal projection matrix that projects any vector into the vector space that is orthogonal to that spanned by $\mathbf{X}$. Projection matrices have some nice properties, listed in the following lemma.

Lemma 4.1.1 (Orthogonal projection matrices)
For $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ *and* $\mathbf{M} = \mathbf{I}_n - \mathbf{P}$ *with* $\mathbf{X}$ *being of full rank it holds:*

(i) $\mathbf{P}$ *and* $\mathbf{M}$ *are symmetric and idempotent, i.e.:*

$$\mathbf{PP} = \mathbf{P} \quad \textit{and} \quad \mathbf{MM} = \mathbf{M}.$$

(ii) Further properties:

$$\mathbf{X}'\mathbf{P} = \mathbf{X}', \quad \mathbf{X}'\mathbf{M} = \mathbf{0}, \quad \textit{and} \quad \mathbf{PM} = \mathbf{0}.$$

Proofs follow directly from the definitions of **P** and **M**.

Using these results we obtain the following proposition on the OLS residuals and OLS fitted values.

Proposition 4.1.2 (OLS residuals) *For the OLS residuals and the OLS fitted values it holds that*

$$\begin{aligned} \mathbf{X}'\hat{\boldsymbol{\varepsilon}} &= \mathbf{0}, \quad \textit{and} \\ \mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \end{aligned}$$

Proof:
The first result can be shown as following:

$$\begin{aligned} \mathbf{X}'\hat{\boldsymbol{\varepsilon}} &= \mathbf{X}'\mathbf{M}\mathbf{y} \quad \text{(By Def. of } \mathbf{M}\text{)} \\ &= \mathbf{0}\mathbf{y} \quad \text{(By Lemma 4.1.1 part (ii))} \\ &= \underset{(K\times 1)}{\mathbf{0}} \end{aligned}$$

The second result follows from:

$$\begin{aligned} \mathbf{y}'\mathbf{y} &= (\mathbf{Py}+\mathbf{My})'(\mathbf{Py}+\mathbf{My}) \quad \text{(By Def. of } \mathbf{P} \text{ and } \mathbf{M}\text{)} \\ &= (\mathbf{y}'\mathbf{P}'+\mathbf{y}'\mathbf{M}')(\mathbf{Py}+\mathbf{My}) \\ &= \mathbf{y}'\mathbf{P}'\mathbf{Py}+\mathbf{y}'\mathbf{M}'\mathbf{My}+\mathbf{0} \quad \text{(By Lemma 4.1.1 part (ii))} \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}}+\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \quad \square \end{aligned}$$

The vector of residuals $\hat{\boldsymbol{\varepsilon}}$ has only $n-K$ so-called *degrees of freedom*. The vector looses K degrees of freedom, since it has to satisfy the K linear restrictions ($\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$). Particularly, in the case with intercept we have that $\sum_{i=1}^{n} \hat{\boldsymbol{\varepsilon}}_i = \mathbf{0}$.

This loss of K degrees of freedom also appears in the definition of the *unbiased* variance estimator

$$s^2 = \frac{1}{n-K} \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{4.8}$$

Coefficient of determination

The total sample variance of the dependent variable $\sum_{i=1}^{n} (y_i - \bar{y})^2$, where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, can be decomposed as following:

Proposition 4.1.3 (Variance decomposition) *For the OLS regression of the linear model* (4.1) *with intercept it holds that*

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\textit{total variance}} = \underbrace{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{\hat{y}}\right)^2}_{\textit{explained variance}} + \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}_{\textit{unexplained variance}} .$$

Proof:

- As a consequence of Prop. 4.1.2 we have for regressions with intercept: $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$. Hence, from $y_i = \hat{y}_i + \hat{\varepsilon}_i$ it follows that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} y_i &= \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i + \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i \\ \bar{y} &= \bar{\hat{y}}_i + 0 \end{aligned}$$

- From Prop. 4.1.2 we know that:

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\
\mathbf{y}'\mathbf{y} - n\bar{y}^2 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y}^2 + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\
\mathbf{y}'\mathbf{y} - n\bar{y}^2 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{\hat{y}}^2 + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \quad \text{(By our result above.)} \\
\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 &= \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{\hat{y}}^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \\
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad \square
\end{aligned}$$

The larger the proportion of the explained variance, the better is the fit of the model. This motivates the definition of the so-called R^2 coefficient of determination:

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

Obviously, we have that $0 \leq R^2 \leq 1$. The closer R^2 lies to 1, the better is the fit of the model to the observed data. However, a high/low R^2 does not mean a validation/falsification of the estimated model. Any relation (i.e., model assumption) needs a plausible explanation from relevant economic theory.
The most often criticized disadvantage of the R^2 is that additional regressors (relevant or not) will always increase the R^2.

Proposition 4.1.4 (R^2 increase) *Let R_1^2 and R_2^2 result from*

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}_1\,\mathbf{b}_{11} + \hat{\boldsymbol{\varepsilon}}_1 \quad \textit{and} \\
\mathbf{y} &= \mathbf{X}_1\,\mathbf{b}_{21} + \mathbf{X}_2\,\mathbf{b}_{22} + \hat{\boldsymbol{\varepsilon}}_2.
\end{aligned}$$

It then holds that $R_2^2 \geq R_1^2$.

Proof:
Consider the sum of squared residuals,

$$S(\mathfrak{b}_{21}, \mathfrak{b}_{22}) = (\mathbf{y} - \mathbf{X}_1\,\mathfrak{b}_{21} + \mathbf{X}_2\,\mathfrak{b}_{22})'(\mathbf{y} - \mathbf{X}_1\,\mathfrak{b}_{21} + \mathbf{X}_2\,\mathfrak{b}_{22})$$

By definition, this sum is minimized by the OLS estimators $\mathbf{b}_{21}$ and $\mathbf{b}_{22}$, i.e., $S(\mathbf{b}_{21}, \mathbf{b}_{22}) \leq S(\mathfrak{b}_{21}, \mathfrak{b}_{22})$. Consequently,

$$\hat{\boldsymbol{\varepsilon}}_2'\hat{\boldsymbol{\varepsilon}}_2 = S(\mathbf{b}_{21}, \mathbf{b}_{22}) \leq S(\mathbf{b}_{11}, 0) = \hat{\boldsymbol{\varepsilon}}_1'\hat{\boldsymbol{\varepsilon}}_1$$

which implies the statement:

$$R_2^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}_2'\hat{\boldsymbol{\varepsilon}}_2}{\sum_{i=1}^n (y_i - \bar{y})^2} \geq 1 - \frac{\hat{\boldsymbol{\varepsilon}}_1'\hat{\boldsymbol{\varepsilon}}_1}{\sum_{i=1}^n (y_i - \bar{y})^2} = R_1^2 \quad \square$$

Because of this, the R^2 cannot be used as a criterion for model selection. Possible solutions are given by penalized criterions such as the so-called *adjusted* R^2 defined as

$$\begin{aligned} \overline{R}^2 &= 1 - \frac{\frac{1}{n-K}\sum_{i=1}^n \hat{u}_i^2}{\frac{1}{n-1}\sum_{i=1}^n (y-\bar{y})_i^2} \\ &= 1 - \frac{n-1}{n-K}\left(1 - R^2\right) \\ &= R^2 - \frac{K-1}{n-K}\left(1 - R^2\right) \;\leq\; R^2 \end{aligned}$$

The adjustment is in terms of degrees of freedom.

Partitioned regression model

Already in the first edition of Econometrica (1933) Frisch and Waugh pointed to an interesting property of multivariate linear regression analysis, which was later generalized to by Lovell (1963). The so-called Frisch-Waugh-Lovell (FWL) theorem points to a property of the OLS estimation method, which allows to gain a deeper understanding of the estimation method that is useful for the interpretation of the estimated coefficients.

$$\mathbf{y} \;=\; \mathbf{X}_1\,\mathbf{b}_1 + \mathbf{X}_2\,\mathbf{b}_2 + \hat{\boldsymbol{\varepsilon}} \;=\; (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{pmatrix} + \hat{\boldsymbol{\varepsilon}}, \qquad (4.9)$$

where $\operatorname{rank}(\mathbf{X}_j) = K_j$ for $j = 1, 2$.

A regression of $\mathbf{y}$ only on $\mathbf{X}_2$ (not on $\mathbf{X}_1$), which however *takes into account the effect of* $\mathbf{X}_1$, has to be done as following:

$$\mathbf{M}_1\mathbf{y} \;=\; \mathbf{M}_1\,\mathbf{X}_2\,\hat{\boldsymbol{\beta}}_2 + \hat{\mathbf{v}}, \qquad (4.10)$$

where $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\,\mathbf{X}_1)^{-1}\,\mathbf{X}_1'$. Note that (4.10) is a regression model full of residuals: The dependent variables $\mathbf{M}_1\mathbf{y}$ are the residuals from regressing $\mathbf{y}$ on $\mathbf{X}_1$ and the K_2 columns in the matrix of independent variables $\mathbf{M}_1\mathbf{X}_2$ are the residuals from the regressing $\mathbf{X}_2$ column-wise on $\mathbf{X}_1$. This means that the variables $\mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\,\mathbf{X}_2$ contain only those parts of $\mathbf{y}$ and $\mathbf{X}_2$, which are orthogonal to $\mathbf{X}_1$; the effect of $\mathbf{X}_1$ is "*partialled out*". By the FWL theorem we have that:

Proposition 4.1.5 (Frisch-Waugh-Lovell theorem) *For the equations* (4.9) *and* (4.10) *it holds that:*

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{b}_2 \quad \textit{and} \quad \hat{\boldsymbol{\varepsilon}} = \hat{\mathbf{v}}.$$

Proof:
The OLS estimator $\hat{\boldsymbol{\beta}}_2$ is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= ((\mathbf{M}_1\,\mathbf{X}_2)'(\mathbf{M}_1\,\mathbf{X}_2))^{-1}\,(\mathbf{M}_1\,\mathbf{X}_2)'\mathbf{y}\\ &= (\mathbf{X}_2'\,\mathbf{M}_1\,\mathbf{X}_2)^{-1}\,\mathbf{X}_2'\,\mathbf{M}_1\mathbf{y} \qquad (4.11)\end{aligned}$$

In the following, we show that $\hat{\boldsymbol{\beta}}_2 = \mathbf{b}_2$:
From the normal equations for $\mathbf{b}$, we have that (using the partition $X = [\mathbf{X}_1, \mathbf{X}_2]$):

$$\left(\mathbf{X}\,\mathbf{X}\right)^{-1}\mathbf{b} = \mathbf{X}'\,\mathbf{y}$$
$$\begin{pmatrix} \mathbf{X}_1'\,\mathbf{X}_1 & \mathbf{X}_1'\,\mathbf{X}_2 \\ \mathbf{X}_2'\,\mathbf{X}_1 & \mathbf{X}_2'\,\mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\,\mathbf{y} \\ \mathbf{X}_2'\,\mathbf{y} \end{pmatrix},$$

which is an equation system with two equations:

$$\mathbf{X}_1'\,\mathbf{X}_1\,\mathbf{b}_1 + \mathbf{X}_1'\,\mathbf{X}_2\,\mathbf{b}_2 = \mathbf{X}_1'\,\mathbf{y} \tag{4.12}$$
$$\mathbf{X}_2'\,\mathbf{X}_1\,\mathbf{b}_1 + \mathbf{X}_2'\,\mathbf{X}_2\,\mathbf{b}_2 = \mathbf{X}_2'\,\mathbf{y} \tag{4.13}$$

From (4.12):

$$\mathbf{b}_1 = \left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1'\,\mathbf{y} - \mathbf{X}_1'\,\mathbf{X}_2\,\mathbf{b}_2\right) \tag{4.14}$$

Plugging (4.14) into (4.13) yields,

$$\begin{aligned}
\mathbf{X}_2'\,\mathbf{X}_1\left\{\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1'\,\mathbf{y} - \mathbf{X}_1'\,\mathbf{X}_2\,\mathbf{b}_2\right)\right\} + \mathbf{X}_2'\,\mathbf{X}_2\,\mathbf{b}_2 &= \mathbf{X}_2'\,\mathbf{y} \\
-\mathbf{X}_2'\,\mathbf{X}_1\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\,\mathbf{X}_2\,\mathbf{b}_2 + \mathbf{X}_2'\,\mathbf{X}_2\,\mathbf{b}_2 &= \mathbf{X}_2'\,\mathbf{y} - \mathbf{X}_2'\,\mathbf{X}_1\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\,\mathbf{y} \\
\left(\mathbf{X}_2'\,\mathbf{X}_2 - \mathbf{X}_2'\,\mathbf{X}_1\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\,\mathbf{X}_2\right)\mathbf{b}_2 &= \mathbf{X}_2'\left(I - \mathbf{X}_1\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{y} \\
\mathbf{X}_2'\left(I - \mathbf{X}_1\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{X}_2\,\mathbf{b}_2 &= \mathbf{X}_2'\left(I - \mathbf{X}_1\left(\mathbf{X}_1'\,\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{y} \\
\mathbf{X}_2'\,\mathbf{M}_1\,\mathbf{X}_2\,\mathbf{b}_2 &= \mathbf{X}_2'\,\mathbf{M}_1\mathbf{y} \\
\Leftrightarrow \mathbf{b}_2 &= \left(\mathbf{X}_2'\,\mathbf{M}_1\,\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\,\mathbf{M}_1\mathbf{y}
\end{aligned} \tag{4.15}$$

From (4.11) and (4.15) it follows that $\hat{\boldsymbol{\beta}}_2 = \mathbf{b}_2$ as stated by the proposition.
It remains to show that $\hat{\boldsymbol{\varepsilon}} = \hat{\mathbf{v}}$:
Observe that

$$\hat{\mathbf{v}} = \mathbf{M}_1\mathbf{y} - \mathbf{M}_1\,\mathbf{X}_2\,\hat{\boldsymbol{\beta}}_2.$$

But, using (4.14)

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \mathbf{y} - \mathbf{X}_1\,\mathbf{b}_1 - \mathbf{X}_2\,\mathbf{b}_2 \\ &= \mathbf{y} - \mathbf{X}_1\,(\mathbf{X}_1'\,\mathbf{X}_1)^{-1}\,(\mathbf{X}_1'\,\mathbf{y} - \mathbf{X}_1'\,\mathbf{X}_2\,\mathbf{b}_2) - \mathbf{X}_2\,\mathbf{b}_2 \\ &= \mathbf{y} - \mathbf{P}_1\mathbf{y} - (\mathbf{X}_2\,\mathbf{b}_2 - \mathbf{P}_1\,\mathbf{X}_2\,\mathbf{b}_2) \\ &= \mathbf{M}_1\mathbf{y} - \mathbf{M}_1\,\mathbf{X}_2\,\mathbf{b}_2 \\ &= \hat{\mathbf{v}} \quad \square\end{aligned}$$

4.1.3 Finite-Sample Properties of OLS

Notice that, by contrast to (the true but unknown) parameter vector $\boldsymbol{\beta}$, $\mathbf{b}$ is a stochastic quantity, since it depends on $\boldsymbol{\varepsilon}$ through $\mathbf{y}$. The stochastic difference $\mathbf{b}-\boldsymbol{\beta}$ is termed the **sampling error**:

$$\begin{aligned}\mathbf{b}-\boldsymbol{\beta} &= (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{y} - \boldsymbol{\beta} \quad \text{(By Eq. (4.7))} \\ &= (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'(\mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \quad \text{(By Assumption 1)} \\ &= (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{X}\,\boldsymbol{\beta} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\boldsymbol{\varepsilon}\end{aligned}$$

The distribution of $\mathbf{b}$ depends (among others) on the sample size n, although this is not made explicitly by our notation. In this section, we focus on the case of a fix, finite sample size n.

Theorem 4.1.6 (Finite-sample properties)
The OLS estimator $\mathbf{b}$

(i) is an unbiased estimator: $\mathbb{E}(\mathbf{b}\,|\,\mathbf{X}) = \boldsymbol{\beta}$

(ii) has variance: $\mathbb{V}(\mathbf{b}\,|\,\mathbf{X}) = \sigma^2(\mathbf{X}'\,\mathbf{X})^{-1}$

(iii) (Gauss-Markov Theorem) is efficient in the class of all linear unbiased estimators. That is, for any unbiased estimator $\tilde{\mathbf{b}}$ *that is linear in* $\mathbf{y}$*, we have:* $\mathbb{V}(\tilde{\mathbf{b}}|\,\mathbf{X}) \geq \mathbb{V}(\mathbf{b}\,|\,\mathbf{X})$ *in the matrix sense.*

While part (ii) and (iii) need all of the classical Assumptions 1.1-1.4, part (i) needs only the Assumptions 1.1-1.3.

Note that, by saying: "$\mathbb{V}(\tilde{\mathbf{b}}|\,\mathbf{X}) \geq \mathbb{V}(\mathbf{b}\,|\,\mathbf{X})$ in the matrix sense", we mean that $\mathbb{V}(\tilde{\mathbf{b}}|\,\mathbf{X}) - \mathbb{V}(\mathbf{b}\,|\,\mathbf{X}) = \mathbf{D}$, where $\mathbf{D}$ is a *positive semidefinite* $K \times K$ matrix, i.e., $\mathbf{a}'\mathbf{D}\mathbf{a} \geq 0$ for any K-dimensional vector $\mathbf{a}$. Observe that this implies that $\mathbb{V}(\tilde{\mathrm{b}}_k|\,\mathbf{X}) \geq \mathbb{V}(\mathrm{b}_k|\,\mathbf{X})$ for any $k = 1, \ldots, K$.

Proof:

Part (i):

$$\begin{aligned}
\mathbb{E}(\mathbf{b}\,|\,\mathbf{X}) &= \mathbb{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\,\mathbf{X}\right)\\
&= \mathbb{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\,\mathbf{X}\right)\\
&= \mathbb{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)\\
&= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}\,(\boldsymbol{\varepsilon}\,|\,\mathbf{X}) \;=\; \boldsymbol{\beta},
\end{aligned}$$

where the last step follows from the strict exogeneity assumption.

Part (ii):

$$\begin{aligned}
\mathbb{V}(\mathbf{b}\,|\,\mathbf{X}) &= \mathbb{V}(\mathbf{b}-\boldsymbol{\beta}\,|\,\mathbf{X}) \quad \text{(Since } \boldsymbol{\beta} \text{ is not random)}\\
&= \mathbb{V}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)\\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{V}(\boldsymbol{\varepsilon}\,|\,\mathbf{X})\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\\
&= \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,I_n\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\\
&= \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Part (iii), Gauss-Markov:

Since $\tilde{\mathbf{b}}$ is assumed to be linear in $\mathbf{y}$, we can write

$$\tilde{\mathbf{b}} = \mathbf{C}\mathbf{y},$$

where $\mathbf{C}$ is some $K \times n$ matrix, which is a function of $\mathbf{X}$ and/or nonrandom components.

Adding a $K \times n$ zero matrix $\mathbf{0}$ yields

$$\tilde{\mathbf{b}} = \Bigl(\mathbf{C}\overbrace{-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}^{=\mathbf{0}}\Bigr)\mathbf{y}.$$

Let now $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, then

$$\begin{aligned}
\tilde{\mathbf{b}} &= \mathbf{D}\mathbf{y} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\
\tilde{\mathbf{b}} &= \mathbf{D}\,(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\\
\tilde{\mathbf{b}} &= \mathbf{D}\,\mathbf{X}\,\boldsymbol{\beta} + \mathbf{D}\,\boldsymbol{\varepsilon} + \mathbf{b} \qquad (4.16)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad \mathbb{E}(\tilde{\mathbf{b}}|\,\mathbf{X}) &= \mathbb{E}(\mathbf{D}\,\mathbf{X}\,\boldsymbol{\beta}\,|\,\mathbf{X}) + \mathbb{E}(\mathbf{D}\,\boldsymbol{\varepsilon}\,|\,\mathbf{X}) + \mathbb{E}(\mathbf{b}\,|\,\mathbf{X})\\
&= \mathbf{D}\,\mathbf{X}\,\boldsymbol{\beta} + \mathbf{0} + \boldsymbol{\beta} \qquad (4.17)
\end{aligned}$$

Since $\tilde{\mathbf{b}}$ is (by assumption) unbiased, we have that $\mathbb{E}(\tilde{\mathbf{b}}|\,\mathbf{X}) = \boldsymbol{\beta}$. The latter, together with (4.17), implies that $\mathbf{D}\,\mathbf{X} = \mathbf{0}$.

Plugging $\mathbf{D}\,\mathbf{X} = \mathbf{0}$ into (4.16) yields,

$$\begin{aligned}
\tilde{\mathbf{b}} &= \mathbf{D}\,\boldsymbol{\varepsilon} + \mathbf{b} \\
\tilde{\mathbf{b}} - \boldsymbol{\beta} &= \mathbf{D}\,\boldsymbol{\varepsilon} + (\mathbf{b} - \boldsymbol{\beta}) \quad \text{(Adding a zero vector } \boldsymbol{\beta} - \boldsymbol{\beta}\text{)} \\
\tilde{\mathbf{b}} - \boldsymbol{\beta} &= \mathbf{D}\,\boldsymbol{\varepsilon} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\boldsymbol{\varepsilon} \quad \text{(Sampling error expression)} \\
\tilde{\mathbf{b}} - \boldsymbol{\beta} &= \left(\mathbf{D} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\right)\boldsymbol{\varepsilon} \qquad (4.18)
\end{aligned}$$

So,

$$\begin{aligned}
\mathbb{V}(\tilde{\mathbf{b}}|\,\mathbf{X}) &= \mathbb{V}(\tilde{\mathbf{b}} - \boldsymbol{\beta}\,|\,\mathbf{X}) \quad \text{(Since } \boldsymbol{\beta} \text{ is not random)} \\
&= \mathbb{V}((\mathbf{D} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}')\,\boldsymbol{\varepsilon}\,|\,\mathbf{X}) \quad \text{(using (4.18))} \\
&= (\mathbf{D} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}')\,\mathbb{V}(\boldsymbol{\varepsilon}\,|\,\mathbf{X})(\mathbf{D}' + \mathbf{X}(\mathbf{X}'\,\mathbf{X})^{-1}) \\
&= \sigma^2(\mathbf{D} + (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}')I_n(\mathbf{D}' + \mathbf{X}(\mathbf{X}'\,\mathbf{X})^{-1}) \\
&= \sigma^2\left(\mathbf{D}\mathbf{D}' + (\mathbf{X}'\,\mathbf{X})^{-1}\right) \quad \text{(using that } \mathbf{D}\,\mathbf{X} = \mathbf{0}\text{)} \\
&\geq \sigma^2(\mathbf{X}'\,\mathbf{X})^{-1} = \mathbb{V}(\mathbf{b}\,|\,\mathbf{X}) \quad \text{(Since } \mathbf{D}\mathbf{D}' \text{ is pos. semidef.)}
\end{aligned}$$

Showing that $\mathbf{D}\mathbf{D}'$ is positive definite:

$$\mathbf{a}'\mathbf{D}\mathbf{D}'\mathbf{a} = (\mathbf{D}'\mathbf{a})'(\mathbf{D}'\mathbf{a}) = \tilde{\mathbf{a}}'\tilde{\mathbf{a}} \geq 0,$$

where $\tilde{\mathbf{a}}$ is a K dimensional column-vector.

Remember:

- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{A}\mathbf{B})' = \mathbf{B}'\mathbf{A}'$
- $\mathbf{A}' = \mathbf{A} \;\Leftrightarrow\; \mathbf{A}$ is a symmetric matrix

Proposition 4.1.7 (Unbiasedness of s^2) *Under Assumptions 1.1-1.4, we have that:*

$$\mathbb{E}(s^2 |\, \mathbf{X}) = \sigma^2,$$

and hence $\mathbb{E}(s^2) = \sigma^2$*, provided that* $n > K$ *(otherwise* s^2 *isn't well defined).*

Proof:
In the following we show that $\mathbb{E}(s^2 |\, \mathbf{X}) = \sigma^2$, where

$$s^2 = \frac{1}{n-K} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{n-K}.$$

In fact, it will be convenient to show the following equivalent statement:

$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} |\, \mathbf{X}) = \sigma^2 \cdot (n-K).$$

Note that

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}} &= (\mathbf{M}\mathbf{y})'\mathbf{M}\mathbf{y} \\ &= (\mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}))'\mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (\mathbf{M}\,\boldsymbol{\varepsilon})'\mathbf{M}\,\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}'\,\mathbf{M}\,\boldsymbol{\varepsilon}\,. \end{aligned}$$

First, we show that $\mathbb{E}(\boldsymbol{\varepsilon}'\,\mathbf{M}\,\boldsymbol{\varepsilon}\,|\,\mathbf{X}) = \sigma^2\,\text{trace}(\mathbf{M})$, second, we show that $\text{trace}(\mathbf{M}) = n - K$.

1st Part:

$$\boldsymbol{\varepsilon}'\,\mathbf{M}\,\boldsymbol{\varepsilon} = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j \quad \text{(All } m_{ij}\text{'s are functions of } \mathbf{X}\text{)}$$

$$\Rightarrow \mathbb{E}(\boldsymbol{\varepsilon}'\,\mathbf{M}\,\boldsymbol{\varepsilon}\,|\,\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\,\mathbb{E}(\varepsilon_i\varepsilon_j|\,\mathbf{X})$$
$$= \sum_{i=1}^{n} m_{ii}\sigma^2 = \sigma^2\,\text{trace}(\mathbf{M}).$$

2nd Part:

$$\begin{aligned} \text{trace}(\mathbf{M}) &= \text{trace}(I_n - P) \\ &= \text{trace}(I_n) - \text{trace}(P) \quad \text{(By linearity of trace(.))} \\ &= n - \text{trace}(P) \\ &= n - \text{trace}(\mathbf{X}(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}') \\ &= n - \text{trace}(\mathbf{X}'\,\mathbf{X}(\mathbf{X}'\,\mathbf{X})^{-1}) \\ &= n - \text{trace}(I_K) \\ &= n - K. \end{aligned}$$

Such that

$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}|\,\mathbf{X}) = \sigma^2(n-K). \quad \square$$

Remember (trace-trick):

- trace(AB) = trace(BA)

4.1.4 Hypothesis Testing under Normality

Assumption 1.5: Normality

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

Strictly, speaking, the only aspect of this assumption is that ε is normally distributed. The assumption immediately implies that

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

which inspires our test statistics. E.g., if we would know σ^2, we have

$$z_k = \frac{\mathrm{b}_k - \bar{\beta}_k}{\left(\sigma^2 \left[(\mathbf{X}' \mathbf{X})^{-1}\right]_{kk}\right)^{1/2}} \sim N(0, 1),$$

where $\bar{\beta}_k$ is some known value specified by the null hypothesis: H_0: $\mathrm{b}_k = \bar{\beta}_k$.

Usually, we do not know the value of σ^2 and have to estimate it. In this case σ^2 is termed a **nuisance parameter**. Plugging in the OLS estimate s^2 leads to

$$\text{t-ratio}_k = \frac{\mathrm{b}_k - \bar{\beta}_k}{\left(s^2 \left[(\mathbf{X}' \mathbf{X})^{-1}\right]_{kk}\right)^{1/2}} \sim t_{n-K},$$

where t_{n-K} is the (Student) t-distribution with $n - K$ degrees of freedom.

Of course, **confidence intervals** for the single estimators $\hat{\beta}_k$ can also be directly derived using the normality assumption:

$$CI_{1-\alpha} \;=\; \left[\mathbf{b}_k \pm t_{1-\frac{\alpha}{2}, n-K} \; s^2 \sqrt{\left[(\mathbf{X}\,\mathbf{X})^{-1}\right]_{kk}}\right],$$

where $CI_{1-\alpha}$ contains the true unknown β_k with probability $1-\alpha$.

Testing linear combinations of hypotheses (so-called **linear restrictions**) on $\beta_1, \dots, \beta_K$:

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

where the $(\#\mathbf{r} \times K)$ dimensional matrix $\mathbf{R}$ and the vector $\mathbf{r}$ are known and specified by the hypothesis, and $\#\mathbf{r}$ is the number of elements in $\mathbf{r}$ (i.e., the number of linear equations in the nullhypothesis). To make sure that there are no redundant equations it is required that $\text{rank}(\mathbf{R}) = \#\mathbf{r}$.

Based on the normality assumption we can test the nullhypothesis using the χ^2-distributed test statistic

$$\mathrm{W} = \frac{(\mathbf{R}\,\mathbf{b} - \mathbf{r})'(\mathbf{R}(\mathbf{X}'\,\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\,\mathbf{b} - \mathbf{r})}{\sigma^2} \sim \chi^2_{\#\mathbf{r}},$$

where $\chi^2_{\#\mathbf{r}}$ denotes the χ^2-distribution with $\#\mathbf{r}$ degrees of freedom. If σ^2 is unknown we have to plug-in its estimator s^2, which then changes the distribution of the test statistic:

$$\mathrm{F} = \frac{(\mathbf{R}\,\mathbf{b} - \mathbf{r})'(\mathbf{R}(\mathbf{X}'\,\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\,\mathbf{b} - \mathbf{r})}{s^2\#\mathbf{r}} \sim F_{\#\mathbf{r},n-K},$$

where $F_{\#\mathbf{r},n-K}$ is the F-distribution with $\#\mathbf{r}, n-K$ degrees of freedom.

4.1.5 Asymptotics under the Classic Regression Model

In this section we proof that the OLS estimators $\mathbf{b}$ and s^2 applied to the classic regression model (defined by Assumptions 1.1

to 1.4) are consistent estimators as $n \to \infty$. Even better, we can show that it is possible to drop the unrealistic normality assumption (Assumption 1.5.), but still to use the usual test statistics as long as the sample size n is large. Though, before we can formally state the asymptotic properties, we first need to adjust the rank assumption (Assumption 1.3), such that the full column rank of $\mathbf{X}$ is guaranteed for the limiting case as $n \to \infty$, too. Second, we need to assume that the sample $(y_i, \mathbf{x}_i)$ is iid, which allows us to apply Kolmogorov's strong LLN and Lindeberg-Levy's CLT.

Assumption 1.3*: $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') = \mathbf{\Sigma}_{\mathbf{xx}}$,
such that the $(K \times K)$ matrix $\mathbf{\Sigma}_{\mathbf{xx}}$ has full rank K (i.e., is nonsingular).

Assumption 1.5*: The sample $(\mathbf{x}_i, \varepsilon_i)$ (equivalently $(y_i, \mathbf{x}_i)$) is iid for all $i = 1, \dots, n$, with existing and finite first, second, third, and fourth moments.

Note that existence and finiteness of the first two moments of $\mathbf{x}_i$ is actually already implied by Assumption 1.3*.

Under the Assumptions 1.1, 1.2, 1.3*, 1.4, and, 1.5* we can show the following results.

Proposition 4.1.8 (Consistency of $\mathbf{S}_{\mathbf{xx}}^{-1}$)

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = \mathbf{S}_{\mathbf{xx}}^{-1} \quad \xrightarrow{\text{p}} \quad \mathbf{\Sigma}_{\mathbf{xx}}^{-1}$$

Proof:
1st Part: Let define

$$[\mathbf{S_{xx}}]_{kl} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{x_{ik}x_{il}}_{z_{i,kl}} = \bar{z}_{kl}.$$

From:

$\mathbb{E}[z_{i,kl}] = [\mathbf{S_{xx}}]_{kl}$ (By Assumption 1.3*)
and
$z_{i,kl}$ is iid and has four moments (By Assumption 1.5*)

it follows by Kolmogorov's strong law of large numbers (LINK) that

$$\bar{z}_{kl} \xrightarrow{\text{a.s.}} [\boldsymbol{\Sigma}_{\mathbf{xx}}]_{kl}, \quad \text{for any} \quad 1 \leq k, l \leq K.$$

Consequently, $\mathbf{S_{xx}} \xrightarrow{\text{a.s.}} \boldsymbol{\Sigma}_{\mathbf{xx}}$ element-wise.

2nd Part: By the Continuous Mapping Theorem (LINK) we have that also

$$(\mathbf{S_{xx}})^{-1} \xrightarrow{\text{a.s.}} (\boldsymbol{\Sigma}_{\mathbf{xx}})^{-1}.$$

3rd Part: Almost-Sure-Convergence implies Convergence-in-Probability ($\xrightarrow{\text{a.s.}} \Rightarrow \xrightarrow{\text{p}}$); see relations among modes of convergence (LINK). □

Proposition 4.1.9 (Consistency of b)

$$\mathbf{b} \xrightarrow{\text{p}} \boldsymbol{\beta}$$

Proof: We show the equivalent result that $\mathbf{b} - \boldsymbol{\beta} \xrightarrow{\text{p}} \mathbf{0}$. Remember:

$$\begin{aligned} \mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\boldsymbol{\varepsilon} \\ &= (n^{-1}\,\mathbf{X}'\,\mathbf{X})^{-1}\frac{1}{n}\,\mathbf{X}'\,\boldsymbol{\varepsilon} \\ &= (\mathbf{S}_{\mathbf{xx}})^{-1}\;\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \end{aligned}$$

From Proposition 4.1.8: $(\mathbf{S}_{\mathbf{xx}})^{-1} \xrightarrow{\text{p}} (\boldsymbol{\Sigma}_{\mathbf{xx}})^{-1}$.

Let us focus on element-by-element asymptotics of $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i$: Define

$$\frac{1}{n}\sum_{i=1}^{n}\underbrace{x_{ik}\varepsilon_i}_{z_{ik}} = \bar{z}_{n,k}.$$

From:

$$\begin{aligned} &\mathbb{E}[z_{ik}] = \mathbb{E}[x_{ik}\varepsilon_i] = 0 && \text{(By Str. Exog. Ass 1.2)} \\ &\text{and} \\ &z_{ik} \quad \text{is iid and has four moments} && \text{(By Assumption 1.5*)} \end{aligned}$$

it follows by Kolmogorov's strong law of large numbers (LINK) that

$$\bar{z}_{n,k} = \sum_{i=1}^{n} x_{ik}\varepsilon_i \xrightarrow{\text{a.s.}} 0 \quad \text{for any} \quad 1 \leq k \leq K.$$

Consequently, also

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{\text{a.s.}} \underset{(K\times 1)}{\mathbf{0}} \quad \text{(element-wise).}$$

Almost-Sure-Convergence implies Convergence-in-Probability ($\xrightarrow{\text{a.s.}} \Rightarrow \xrightarrow{\text{p}}$); see relations among modes of convergence (LINK):

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{\text{p}} \underset{(K\times 1)}{\mathbf{0}} \quad \text{(element-wise).}$$

Final step: From

$$(\mathbf{S}_{\mathbf{xx}})^{-1} \xrightarrow{\text{p}} (\mathbf{\Sigma}_{\mathbf{xx}})^{-1}$$

and

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

it follows by Slutsky's Theorem (LINK) that

$$\mathbf{b}-\boldsymbol{\beta} = (\mathbf{S}_{\mathbf{xx}})^{-1}\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{\text{p}} \mathbf{0}. \quad \square$$

Furthermore, we can show that the appropriately scaled (by $\sqrt{n}$) sampling error $\mathbf{b}-\boldsymbol{\beta}$ of the OLS estimator is asymptotically normal distributed.

Proposition 4.1.10 (Sampling error limiting normality)

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

In order to show Proposition 4.1.10, we will need to use the so-called Cramér Wold Device on multivariate convergence in distribution:

Cramér Wold Device: Let $\mathbf{z}_n, \mathbf{z} \in \mathbf{R}^K$, then

$$\mathbf{z}_n \xrightarrow{\text{d}} \mathbf{z} \quad \text{if and only if} \quad \boldsymbol{\lambda}'\mathbf{z}_n \xrightarrow{\text{d}} \boldsymbol{\lambda}'\mathbf{z}$$

for any $\boldsymbol{\lambda} \in \mathbb{R}^K$.

The Cramér Wold Device is needed, since $\mathbf{z}_n \xrightarrow{\text{d}} \mathbf{z}$ implies convergence in distribution element-by-element, **but** convergence in distribution element-by-element does not imply $\mathbf{z}_n \xrightarrow{\text{d}} \mathbf{z}$.

Proof:

Let's start with some rearrangements:

$$\begin{aligned} \mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\boldsymbol{\varepsilon} \\ &= (n^{-1}\,\mathbf{X}'\,\mathbf{X})^{-1}\frac{1}{n}\,\mathbf{X}'\,\boldsymbol{\varepsilon} \\ &= (\mathbf{S}_{\mathbf{xx}})^{-1}\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \\ \Leftrightarrow \sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) &= (\mathbf{S}_{\mathbf{xx}})^{-1}\ \left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right) \end{aligned}$$

From Proposition 4.1.8, we already know that

$$\left(\frac{1}{n}\mathbf{X}'\,\mathbf{X}\right)^{-1} = \mathbf{S}_{\mathbf{xx}}^{-1} \quad \xrightarrow{\text{p}} \quad \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\,.$$

What happens with

$$\sqrt{n}\,\underbrace{\frac{1}{n}\sum_{i=1}^{n}\overbrace{\mathbf{x}_i\varepsilon_i}^{\mathbf{z}_i}}_{\bar{\mathbf{z}}_n} = \sqrt{n}\,\bar{\mathbf{z}}_n \quad ?$$

In the following we show that $\sqrt{n}\,\bar{\mathbf{z}}_n \xrightarrow{\text{d}} N(\mathbf{0}, \sigma^2\ \boldsymbol{\Sigma}_{\mathbf{xx}})$ using the Cramér Wold Device:

1st Moment:

$$\mathbb{E}(\boldsymbol{\lambda}'\mathbf{z}_i) = \boldsymbol{\lambda}' \underbrace{\begin{pmatrix} \mathbb{E}(\mathbf{x}_{i1}\varepsilon_i) \\ \vdots \\ \mathbb{E}(\mathbf{x}_{iK}\varepsilon_i) \end{pmatrix}}_{\substack{\mathbf{0} \\ \text{(By Str. Exog. Ass 1.2)}}} = \boldsymbol{\lambda}'\mathbf{0} = 0,$$

for any $\boldsymbol{\lambda} \in \mathbb{R}^K$ and for all $i = 1, 2, \ldots$

2nd Moment:

$$\begin{aligned} \mathbb{V}(\boldsymbol{\lambda}'\mathbf{z}_i) &= \boldsymbol{\lambda}'\,\mathbb{V}(\mathbf{z}_i)\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'\,\mathbb{E}(\varepsilon_i\mathbf{x}_i\mathbf{x}_i')\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'\,\mathbb{E}(\mathbb{E}(\varepsilon_i\mathbf{x}_i\mathbf{x}_i'|\,\mathbf{X}))\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'\,\mathbb{E}(\mathbf{x}_i\mathbf{x}_i'\underbrace{\mathbb{E}(\varepsilon_i|\,\mathbf{X})}_{\substack{=\sigma^2 \\ \text{(Ass 1.4)}}})\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'\sigma^2\underbrace{\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')}_{\substack{\boldsymbol{\Sigma}_{\mathbf{xx}} \\ \text{(Ass 1.3*)}}}\boldsymbol{\lambda} = \sigma^2\boldsymbol{\lambda}'\,\boldsymbol{\Sigma}_{\mathbf{xx}}\,\boldsymbol{\lambda}, \end{aligned}$$

for any $\boldsymbol{\lambda} \in \mathbb{R}^K$ and for all $i = 1, 2, \ldots$

From $\mathbb{E}(\boldsymbol{\lambda}'\mathbf{z}_i) = 0$, $\mathbb{V}(\boldsymbol{\lambda}'\mathbf{z}_i) = \sigma^2\boldsymbol{\lambda}'\,\boldsymbol{\Sigma}_{\mathbf{xx}}\,\boldsymbol{\lambda}$, and $\mathbf{z}_i = (\mathbf{x}_i\varepsilon_i)$ being iid (Ass 1.5*), it follows by the Lindeberg-Levy's CLT

(LINK) and the Cramér Wold Device that

$$\sqrt{n}\boldsymbol{\lambda}'\bar{\mathbf{z}}_n \xrightarrow{\text{d}} N(0, \sigma^2 \boldsymbol{\lambda}' \,\boldsymbol{\Sigma}_{\mathbf{xx}}\, \boldsymbol{\lambda}) \quad \text{(By Lindeberg-Levy's CLT)}$$

$$\Leftrightarrow \quad \underbrace{\sqrt{n}\bar{\mathbf{z}}_n}_{=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i} \xrightarrow{\text{d}} N(\mathbf{0}, \sigma^2 \,\boldsymbol{\Sigma}_{\mathbf{xx}}) \quad \text{(Cramér Wold Device)}$$

Now, we can conclude the proof:
From $\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{\text{p}} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$ (by Proposition 4.1.8) and $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \xrightarrow{\text{d}} N(\mathbf{0}, \sigma^2 \,\boldsymbol{\Sigma}_{\mathbf{xx}})$ it follows by Slutsky's Theorem (LINK) that

$$\underbrace{(\mathbf{S}_{\mathbf{xx}})^{-1}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i\right)}_{\sqrt{n}(\mathbf{b}-\boldsymbol{\beta})} \xrightarrow{\text{d}} N\left(\mathbf{0}, \underbrace{(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1})\,(\sigma^2\,\boldsymbol{\Sigma}_{\mathbf{xx}})\,(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1})'}_{\sigma^2\,\boldsymbol{\Sigma}_{\mathbf{xx}}}\right) \quad \square$$