Econometrics M1 TSE
Lecture Notes

Instructors:

Koen Jochmans

Pascal Lavergne

François Poinas

2021 – 2022

- These lecture notes have been prepared for you as TSE M1 students. They contain the core material of the course. These lecture notes are completed by additional material (such as quizzes on Moodle and lecture slides).

- EXERCICES in these notes are aimed to help thinking about the different concepts. They should be easy enough to be solved by yourself (alone or in group) and are not to be corrected in class or during tutorials.

- The appendices contain two types of material:
  - Appendix A contains material related to concepts considered as prerequisites for this course.
  - Appendix B contains material to go further into concepts presented in the main text. This will not serve as a basis for the exam. Students who intend to enter a doctoral program are advised to study this material.

Contents

Chapter 1

Ordinary Least Squares

Learning Objectives of this Chapter

At the end of this chapter, you should:

- Be at ease with matrix notations in linear regression model.
- Understand the geometric interpretation of the Least Squares and its link to orthogonal projection.
- Know the assumptions of the linear model and their interpretation.
- Know the asymptotic properties of the OLS estimator and be able to show them. This implies that you should be able to:
  - understand the main definitions and theorems of asymptotic theory and know in which conditions the results hold,
  - assess the role of the assumptions of the linear model to get the properties of the OLS estimator,
  - prove the consistency and find the asymptotic distribution of related estimators.

1.1 The Ordinary Least Squares Estimator

1.1.1 Linear Regression

Linear Regression Model

We want to explain a variable y by a linear function of other variables $x_1, x_2, ..., x_K$. We assume the following linear relationship between the variables:

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon \,.$$

This relationship will be completed with some assumptions to constitute a **model**.

Economists are interested in explaining many relationships. Here are some examples:

- Impact of fertilizer on yield,
- Wage explained by education, experience, occupation, . . .
- Gasoline consumption explained by income, prices, household characteristics, . . .

The variables included in the model have to be understood as **random variables**. As a consequence, they have a probability distribution, summarized by the Cumulative Distribution Function (CDF). A given variable can be continuous (its realizations are taken in an infinite set of possible values), and has a Probability Density Function (PDF). Alternatively, it can be discrete (its realizations are taken in a finite set of possible values).

Usually $x_1 = 1$ (that is, not random), so that

- β_1 is the **intercept**,
- other β's are the **slope** coefficients.

Linear regression means linear in parameter, not in variables.

- Here are some simple examples:
  - The model $y = \beta_1 + \beta_2 \log(x) + \varepsilon$ is linear in parameters.
  - The model $y = \frac{1}{1+\exp(\beta_1+\beta_2 x)} + \varepsilon$ is non linear in parameters (i.e. $\frac{\partial^2 y}{\partial \beta_1^2} \neq 0$).
- Introducing explanatory variables in a non-linear form can be driven by considerations coming from economic reasoning.
  - As an example, in the following wage equation,

$$Wage = \beta_1 + \beta_2 Education + \beta_3 Experience + \beta_4 Experience^2 + \varepsilon \,,$$

    $experience^2$ accounts for concavity of the wage-experience profile. An increasing and concave wage-experience profile corresponds to $\beta_3 > 0$ and $\beta_4 < 0$.
- Assuming linearity in parameters might be less restrictive than it sounds.
  - As an example, the Cobb-Douglas production function, $Q = AK^{\alpha_1} L^{\alpha_2} \exp(\varepsilon)$, can be transformed by taking the logarithm transformation into

$$\ln Q = \ln A + \alpha_1 \ln K + \alpha_2 \ln L + \varepsilon \,.$$

    Therefore, it gives a model linear in its parameters.

Vector Notation

The linear model can be written more compactly using the vector notation. It gives

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon\,,$$

where

- y is the dependent random variable (it is a scalar),
- $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \end{pmatrix}'$ is the $(K \times 1)$ vector of explanatory variables,
- $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \end{pmatrix}'$ is the $(K \times 1)$ vector of unknown parameters,
- ε is the disturbance or error term.

Notation conventions:

- *Vectors are written in bold and in column, each element placed vertically one after the other.*
- *The prime sign $'$ denotes the transpose of a vector.*

Mean Independence

We make the following assumption about the error term: $\mathrm{E}\,(\varepsilon|x_1,\ldots,x_K) = 0$.

- This means we assume **mean independence** between the regressors and the error.
- Interpretation: all the factors omitted from the specification (that are summarized in ε) are orthogonal to the factors that are included (appearing in the vector $\boldsymbol{x}$).

The mean independence assumption is equivalent to $\mathrm{E}\,(y|x_1,\ldots,x_K) = \boldsymbol{x'\beta}$.

- This means that if we know $\boldsymbol{x}$ (i.e. conditionally on $\boldsymbol{x}$), the best predictor for y is $\boldsymbol{x'\beta}$.
- As a consequence, the β's are **marginal effects** of the x's on y, on average and *ceteris paribus*:

$$\frac{\partial \mathrm{E}\,(y|x_1,\ldots,x_K)}{\partial x_k} = \beta_k\,.$$

Consider the following examples to get the sense of the mean independence assumption:

- Example 1: We consider the following model to assess the impact of fertilizer on yield:

$$Yield = \beta_1 + \beta_2 Fertilizer + \varepsilon\,.$$

  - ε captures unobserved factors that affect the yield, like climate, exposition, fertility of the soil, and others.

  - $Fertilizer$ is the quantity of fertilizer used in the field.
  - The mean independence assumption writes $\mathrm{E}\left(\varepsilon | Fertilizer\right) = 0$. It means that the unobserved factors are orthogonal to the quantity of fertilizer used in the field. Can we believe in this assumption? It depends on the context.
    - Consider an experiment in which different quantities of fertilizer have been put in different fields, and the quantity of fertilizer assigned to each field has been chosen randomly. In this case, the assumption is likely to be true, as the chosen quantity of fertilizer is not related to the fields' characteristics that affect yields.
    - Consider now the case in which the researcher has access to survey data of farmers in Occitanie. Each farmer declares the yields and the quantity of fertilizer he/she used. In that case, assuming that the mean-independence assumption is true would require to assume that the farmers do not use the quantity of fertilizer based on unobserved characteristics. This is unlikely to be true, as a the quantity of fertilizer used might be different for different climate exposure or might depend on the nutriment composition of the soil.

- Example 2: we want to assess the impact of years of education on wages. Let's assume that the true model is the following

$$Wage = \beta_1 + \beta_2 Education + \beta_3 Experience + \beta_4 Ability + u\,.$$

  - Assume *Ability* is not observed. In that case, we would estimate the following model

$$Wage = \beta_1 + \beta_2 Education + \beta_3 Experience + \varepsilon\,.$$

    Ability is part of the unobserved factors, i.e. $\varepsilon = \beta_4 Ability + u$.
  - Assuming the mean independence assumption in the model we estimate implies assuming that ability is orthogonal to education. Indeed, if the level of education is affected itself by the ability level, i.e. if $\mathrm{E}\left(Ability | Education\right) \neq 0$, then $\mathrm{E}\left(\varepsilon | Education\right) \neq 0$

Here we assume that $\mathrm{E}\left(\varepsilon | x_1, \dots, x_K\right) = 0$. Later on, we will see what has to be done if the assumption does not hold.

Random Sampling

Assume we have data from a random sample of size n. For each observation i from the random sample, we observe $(y_i, x_{1i}, x_{2i}, \dots, x_{Ki})$ and we assume that it follows the linear regression model:

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_K x_{Ki} + \varepsilon_i \qquad i = 1, \dots, n \qquad (1.1) \\ &= \boldsymbol{x_i'\beta} + \varepsilon_i \end{aligned}$$

- We will work from a sample drawn from a population. By assumption, **the sample is random**, that is representative of the population.

- Observations on $\boldsymbol{x}$ and y are thus **random** (but x_1 if we set $x_1 = 1$).
- Observations i can be people (wages), households (consumption), firms (production), or countries (GDP).
- ε stands for error, including all **unobserved factors** (it is a random variable). Restrictions (assumptions) on ε lead to identify the unknown parameters (i.e. make their estimation possible).
- Unknown parameters $\boldsymbol{\beta}$ are deterministic (fixed).
- Parameters will be estimated using the sample.

We will mostly consider the case of a random sample (typical in microeconometrics), but we may also have

- time series: replace $i = 1, \ldots, n$ by $t = 1, \ldots, T$ to obtain similar notations, but we may have dependence across time,
- panel data: indexed both by individual i and time t.

Matrix Notations

The n equations in (1.1) can be gathered as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{1.2}$$

where

- $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}'$ is a $(n \times 1)$ vector.
- $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \end{pmatrix}'$ is a $(n \times K)$ matrix.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} x_{k=1,i=1} & x_{21} & \cdots & x_{K1} \\ x_{12} & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{Kn} \end{pmatrix}$$

- $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \end{pmatrix}'$ is (still) the $(K \times 1)$ vector of unknown parameters,
- $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{pmatrix}'$ is a $(n \times 1)$ vector.

Notation conventions:

- *Matrices are written in bold and in capital letters.*
- *In the matrix of covariates, $\boldsymbol{X}$, each row i collects the K variables for individual i and each column k collects the n observations for covariate x_k.*

1.1.2 Least Squares

The **Ordinary Least Squares (OLS) estimator** is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} s(\boldsymbol{\beta}),$$

$$\text{where} \qquad s(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 = \parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \parallel^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\,.$$

Hence $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes the Euclidean distance between the vector $\boldsymbol{y}$ and $\boldsymbol{X\beta}$, denoted $\parallel \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \parallel^2$.

NB: There are (many) other possible estimators, e.g. the *Minimum Absolute Distance* (MAD) estimator, that minimizes $\sum_{i=1}^{n} |y_i - \boldsymbol{x_i'}\boldsymbol{\beta}|$.

The first order condition gives

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\,,$$

if $(\boldsymbol{X}'\boldsymbol{X})$ is not singular, i.e. $\boldsymbol{X}$ is of full rank $\text{Rank}(\boldsymbol{X}) = K$. See Appendix A.4 for details.

Some definitions:

- **Fitted values** are the elements of $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.
- **Residuals** are the elements of $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. They are defined as the differences between the observed explanatory variables and the fitted values.
  - Note: do not confuse "residuals" with "errors".
    - Errors, ε_i, $i = 1, \ldots, n$, are the idiosyncratic errors in the model. They refer to the "theoretical" model.
    - Residuals, $\widehat{\varepsilon}_i$, $i = 1, \ldots, n$, are the differences between observed explanatory variables y_i and and fitted values $\widehat{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$.
- By definition $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\varepsilon}}$.

1.1.3 Geometric Interpretation of Least Squares

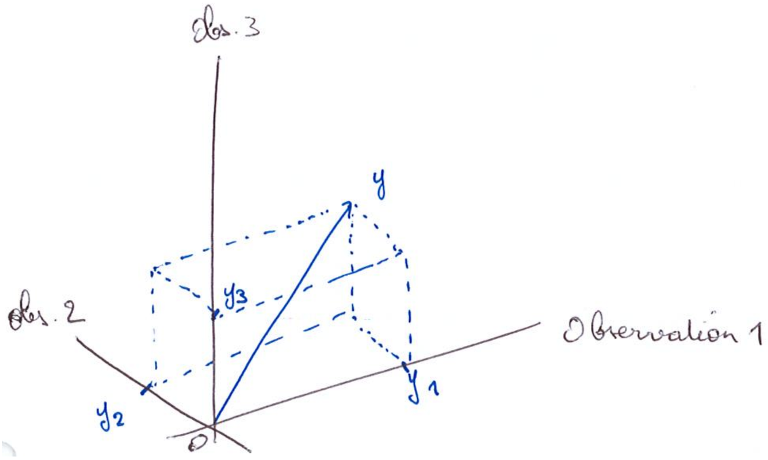
Least Squares can be viewed as a geometrical problem, which allows to interpret it using algebra. This section gives the main intuition and introduces the notations that will be used later in the course. See Appendix B.1 for a complete treatment of the geometry of least squares.

$\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes the Euclidean distance between the vector $\boldsymbol{y}$ and $\boldsymbol{X\beta}$. Therefore, $\widehat{\boldsymbol{\beta}}$ is the vector $\boldsymbol{\beta}$ such that $\boldsymbol{X}\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\mu}}$ is the **orthogonal projection** of $\boldsymbol{y}$ on $\text{L}(\boldsymbol{X})$, the vector linear subspace generated by the columns of $\boldsymbol{X}$ (or *spanned* by $\boldsymbol{X}$).

Graphical representation:

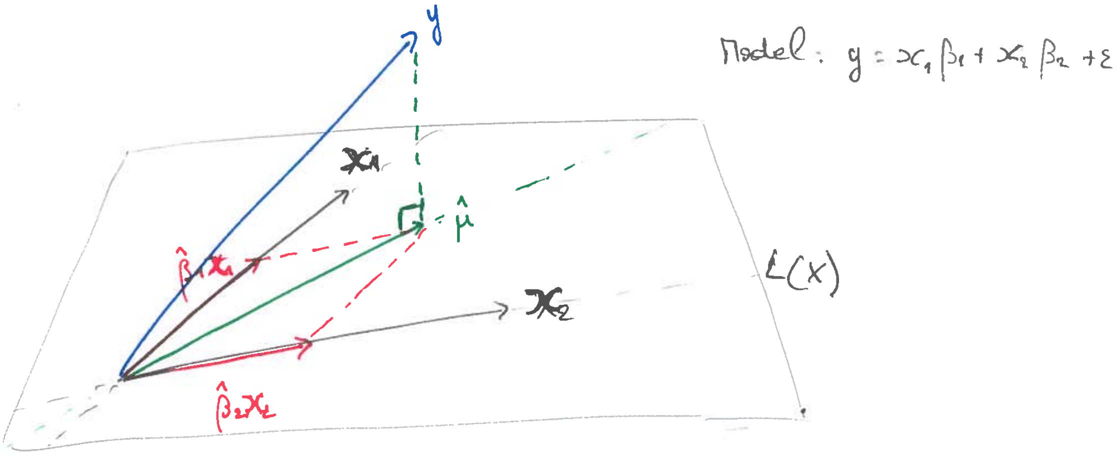
- Each variable observed on n units is represented by a vector in $\mathbb{R}^n$. Axes represent observations, not variables.
  Example: representation of a vector in $\mathbb{R}^3$, i.e. for 3 observations:

Figure 1.1: Three Observations of Data



- LS problem with 2 explanatory variables (2 vectors in $\boldsymbol{X}$, i.e. $\boldsymbol{X} = (\boldsymbol{x}_{k=1} \quad \boldsymbol{x}_{k=2})$).

Figure 1.2: Orthogonal Projection of $\boldsymbol{y}$ on L($\boldsymbol{X}$)



① Draw $x_1, x_2 \longrightarrow L(X)$

② Draw vector y

③ Projection of y on $L(X) \longrightarrow \hat{\mu}$

④ $\hat{\beta}_1$ and $\hat{\beta}_2$ such that $\hat{\mu} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

- If we have more than 2 variables in $\boldsymbol{X}$, L($\boldsymbol{X}$) is not a plane anymore, but a linear subspace of dimension K.

The operation that gives $\widehat{\boldsymbol{\mu}}$ is called an **orthogonal projection** on $\mathrm{L}(\boldsymbol{X})$. The operator is called an **orthogonal projector** (or orthogonal projection).

- If $\mathrm{Rank}(\boldsymbol{X}) = K$, the orthogonal projector on $\mathrm{L}(\boldsymbol{X})$ is $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.
- The orthogonal projector on $\mathrm{L}^{\perp}(\boldsymbol{X})$, the **orthogonal complement** of $\mathrm{L}(\boldsymbol{X})$, is $\boldsymbol{M_X} = \boldsymbol{I_n} - \boldsymbol{P_X} = \boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

As a consequence:

- $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{P_X}\boldsymbol{y}$
- $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{M_X}\boldsymbol{y}$

Properties of the matrices $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$:

- $\boldsymbol{P_X}$ is idempotent and symmetric:

$$\boldsymbol{P_X}\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{P_X}$$

$$\boldsymbol{P}'_{\boldsymbol{X}} = \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{P_X}$$

  - Intuition: the projector $\boldsymbol{P_X}$ projects any vector $\boldsymbol{y}$ on $\boldsymbol{L(X)}$. When we project again the projected vector, this gives the same vector, i.e. $\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{P_X}\boldsymbol{P_X}\boldsymbol{y}$, as the vector $\boldsymbol{P_X}\boldsymbol{y}$ belongs to $\boldsymbol{L(X)}$.

- $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are complementary: they add up and annihilate each other.

$$\boldsymbol{P_X} + \boldsymbol{M_X} = \boldsymbol{P_X} + \boldsymbol{I_n} - \boldsymbol{P_X} = \boldsymbol{I_n}$$

$$\boldsymbol{P_X}\boldsymbol{M_X} = \boldsymbol{P_X}(\boldsymbol{I_n} - \boldsymbol{P_X}) = \boldsymbol{P_X} - \boldsymbol{P_X}\boldsymbol{P_X} = \boldsymbol{0}$$

  - $\boldsymbol{y} = \boldsymbol{P_X}\boldsymbol{y} + \boldsymbol{M_X}\boldsymbol{y}$: this is the "**orthogonal decomposition**" of $\boldsymbol{y}$ because $\boldsymbol{P_X}\boldsymbol{y}$ and $\boldsymbol{M_X}\boldsymbol{y}$ lie in 2 orthogonal subspaces.

- Hence we can write $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{\varepsilon}}$ with $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{\varepsilon}}$ orthogonal, and by Pythagora

$$\|\boldsymbol{y}\|^2 = \|\widehat{\boldsymbol{y}}\|^2 + \|\widehat{\boldsymbol{\varepsilon}}\|^2 .$$

1.1.4 Goodness of Fit

The R^2 of a regression is the fraction of the sample variance of y explained by the regressors.

The (centered) R^2 writes

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} ,$$

where:

- $TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$ is the Total Sum of Squares → Variance in y
- $ESS = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$ is the Explained Sum of Squares → Variance explained by the model
- $RSS = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$ is the Sum of Squared Residuals → Variance not explained by the model

Some comments:

- $R^2 = 1$: perfect fit. $R^2 = 0$: null fit.
- If the model does not include an intercept, we could get R^2 outside $(0, 1)$!
- We always use an intercept in linear regression, we almost never interpret it.
- Alternative definition of R^2: squared correlation between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$, i.e.

$$R^2 = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y})(\widehat{y}_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}\,.$$

If defined that way, R^2 is then always in $(0, 1)$ (why?)

EXERCISE 1.1. *Check that R^2 is the squared correlation between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$.*

Beware of R^2!

- R^2 measures correlation, not causality.
- A high R^2 indicates correlation, not a direct link.
- A high / low R^2 does not mean a good / bad model.
- Often in microeconometrics, we don't care much about R^2.

1.2 Statistical Properties of OLS

The objective in estimating a model is to understand the process that generated the data. That is, we want to know what can be inferred about the world (the population) from particular observations (the sample). Therefore, we need to know under what conditions OLS are useful for such inference.

The statistical analysis is built on assumptions on the way the data are generated (data generating process). We will start by listing and explaining these assumptions. Then, we will remind the properties of the OLS estimator when the sample size is small. Finally, we will study the asymptotic properties of the OLS estimator.

1.2.1 Standard Assumptions

Assumption 1 (IID). $(y_i, x_{1i}, \ldots, x_{Ki}), i = 1, \ldots, n$, are independent and identically distributed.

We consider a random sample, or cross section data. Hence $(y_i, x_{1i}, \ldots, x_{Ki}), i = 1, \ldots, n$, are statistically independent (across individuals) and all have the same distribution.

Assumption 2 (ERRORS). $\mathrm{E}(\varepsilon | x_1, \ldots, x_k) = 0$.

This is the mean-independence assumption that was discussed in Section 1.1.1.

- This assumption is equivalent to $\mathrm{E}\ (y|x_1,\ldots,x_k) = \boldsymbol{x'\beta}$.

$$\mathrm{E}\ (y|\boldsymbol{x}) = \mathrm{E}\ (\boldsymbol{x'\beta}|\boldsymbol{x}) + \mathrm{E}\ (\varepsilon|\boldsymbol{x}) = \boldsymbol{x'\beta}\,.$$

- Assumptions Errors and Iid imply $\mathrm{E}\ (\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X\beta}$.

$$\mathrm{E}\ (\boldsymbol{y}|\boldsymbol{X}) = \mathrm{E}\ (\boldsymbol{X\beta}|\boldsymbol{X}) + \mathrm{E}\ (\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{X\beta}\,.$$

Law of Iterated Expectation

The non-conditional law of iterated expectations (L.I.E.) is a useful result that will be used extensively in the course. It writes:

$$\mathrm{E}\ (\mathrm{E}\ (y|x)) = \mathrm{E}\ (y)\ .$$

This expression involves two random variables y and x. If we were to condition on a realization of x, e.g. $x = x_1$, $\mathrm{E}\ (y|x = x_1)$ would give directly a number. But, by conditioning on x, i.e. the random variable, $\mathrm{E}\ (y|x)$ is a function of x (because x is a random variable that can take many realizations), so it is itself a random variable. The L.I.E. says that, by taking the expected value of this random variable over x, it gives the expectation of y.

The conditional version of the L.I.E. writes

$$\mathrm{E}\ (\mathrm{E}\ (y|x,z)\,|x) = \mathrm{E}\ (y|x)\ .$$

In this expression, the expectation of y taken on x and z gives a function of z and x. Then, by taking the expectation of it (with respect to z), we get the expectation of y, not conditional on z (but conditional on x).

Assumption 3 (Moments)**.** All variables have a bounded moment of order 4, i.e. $\mathrm{E}\,(x_k^4) < \infty$, $k = 1,\ldots K$, $\mathrm{E}\,(y^4) < \infty$.

This assumption means that atypical values (large outliers) are unlikely. This is because of Markov's inequality

$$\Pr\left[|U| > M\right] \leq \frac{\mathrm{E}\,|U|}{M}\,.$$

So for instance

$$\Pr\left[|X - \mathrm{E}\,(X)| > M\right] = \Pr\left[(X - \mathrm{E}\,(X))^2 > M^2\right] \leq \frac{\mathrm{Var}\,(X)}{M^2}\,,$$

which is Chebychev's inequality. Hence the probability that X is at more than M from the mean decreases as $1/M^2$. Similarly

$$\Pr\left[|X| > M\right] = \Pr\left[X^4 > M^4\right] \leq \frac{\mathrm{E}\,X^4}{M^4}\,.$$

Hence the probability that X is greater than M in absolute value decreases as $1/M^4$. The more moments you have, the faster this probability decreases.
Note: for some results, moments of smaller order are sufficient.

Assumption 4 (RANK)**.** No perfect collinearity among the explanatory variables.

Assumption 5 (VARIANCE)**.** Error terms are homoskedastic, i.e. $\text{Var}\,(\varepsilon_i|\boldsymbol{x_i}) = \sigma^2$, $i = 1, \ldots, n$.

Since IID implies $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$, it gives that the variance-covariance matrix of the vector $\boldsymbol{\varepsilon}$ is $\text{Var}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\mathbf{I}$.

As a consequence, $\text{Var}\,(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2\mathbf{I}$

- $\text{Var}\,(\boldsymbol{y}|\boldsymbol{X}) = \text{Var}\,(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}|\boldsymbol{X}) = \text{Var}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\mathbf{I}$

1.2.2 Small Sample Properties (Reminder)

Expectation of OLS

Theorem 1.2.1. *Under Assumptions* IID*,* ERRORS*,* RANK*,*

$$\text{E}\,(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{\beta}\,.$$

The OLS estimator is unbiased, conditionnally on $\boldsymbol{X}$.

Proof. See Appendix A.5. □

Moreover, $\text{E}\,(\widehat{\boldsymbol{y}}|\boldsymbol{X}) = \boldsymbol{X\beta}$.

- $\widehat{\boldsymbol{y}} = \boldsymbol{P_X y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, so $\text{E}\,(\widehat{\boldsymbol{y}}|\boldsymbol{X}) = \boldsymbol{X}\text{E}\,(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \boldsymbol{X\beta}$.

EXERCISE 1.2. *Show that* $\text{E}\,(\widehat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ *and* $\text{E}\,(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$*.*

Variance of OLS

Theorem 1.2.2. *Under Assumptions* IID*,* ERRORS*,* RANK*,* VARIANCE*,*

$$\text{Var}\,\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \sigma^2\left(\boldsymbol{X'X}\right)^{-\mathbf{1}}\,.$$

Proof. See Appendix A.6. □

Moreover $\text{Var}\,(\widehat{\boldsymbol{y}}|\boldsymbol{X}) = \sigma^2\boldsymbol{P_X}$.

- $\text{Var}\,(\widehat{\boldsymbol{y}}|\boldsymbol{X}) = \text{Var}\,(\boldsymbol{P_X y}|\boldsymbol{X}) = \boldsymbol{P_X}\text{Var}\,(\boldsymbol{y}|\boldsymbol{X})\,\boldsymbol{P'_X} = \boldsymbol{P_X}\sigma^2\mathbf{I}\boldsymbol{P_X} = \sigma^2\boldsymbol{P_X}$.

EXERCISE 1.3. *If we regress y on a constant, what is the OLS estimator $\widehat{\beta}$? Determine its expectation* $\text{E}\,(\widehat{\beta})$ *and variance* $\text{Var}\,(\widehat{\beta})$*.*

EXERCISE 1.4. *What is* $\text{Var}\,(\widehat{\boldsymbol{\varepsilon}}|\boldsymbol{X})$*? Compare with* $\text{Var}\,(\boldsymbol{\varepsilon}|\boldsymbol{X})$.

Efficiency

An estimator is said to be more **efficient** than another if it yields more accurate estimates i.e. it utilizes the information available in the sample more "efficiently". The following theorem states the conditions under which the OLS estimator is more efficient than any other linear unbiased estimator.

Theorem 1.2.3 (Gauss-Markov (OLS is BLUE))**.** *Under Assumptions* Iid*,* Errors*,* Rank*,* Variance*, OLS is the most efficient estimator among all the unbiased linear estimators.*

The estimator is said to be linear since it is a linear combination of the y_i, $i = 1, \ldots, n$.

Proof. See Appendix A.7. □

Exercise 1.5. *Consider the estimation of $\mu = \mathrm{E}\, y$ and a linear estimator $\sum_{i=1}^{n} w_i y_i$. Show by calculus that the empirical mean is the most efficient unbiased linear estimator.*

- **Linear Combination of Parameters**:
  If we consider the linear combination $\boldsymbol{\gamma} = \boldsymbol{a'\beta}$, then $\widehat{\boldsymbol{\gamma}} = \boldsymbol{a'}\widehat{\boldsymbol{\beta}}$ is an unbiased estimator and $\mathrm{Var}\,(\widehat{\gamma}) = \boldsymbol{a'}\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{a}$. Therefore, it is also the best estimator of $\boldsymbol{\gamma}$ among all the unbiased linear estimators.

- **Prediction**:
  Given $\boldsymbol{x_{n+1}}$, we can predict y_{n+1} by $\widehat{y}_{n+1} = \boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}$. The Mean Squared Error (MSE) of the prediction is

  $$\mathrm{E}\,\left[(y_{n+1} - \widehat{y}_{n+1})^2\,|\boldsymbol{X}, \boldsymbol{x_{n+1}}\right] = \sigma^2 + \sigma^2 \boldsymbol{x'_{n+1}(X'X)^{-1}x_{n+1}}\,.$$

  Proof. See Appendix B.2. □

Estimation of the Variance of the Errors

- **Errors / Residuals**:
  - The *errors* ε_i are i.i.d., with $\mathrm{E}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}$, and possibly $\mathrm{Var}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\mathbf{I}$.
  - The *residuals* $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{M_X y} = \boldsymbol{M_X(X\beta + \varepsilon)} = \boldsymbol{M_X\varepsilon}$ are linear combinations of all the errors, so

    $$\mathrm{E}\,(\widehat{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \mathrm{E}\,(\boldsymbol{M_X\varepsilon}|\boldsymbol{X}) = \boldsymbol{M_X}\mathrm{E}\,(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}$$

    $$\mathrm{Var}\,(\widehat{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \mathrm{Var}\,(\boldsymbol{M_X\varepsilon}|\boldsymbol{X}) = \boldsymbol{M_X}\mathrm{Var}\,(\boldsymbol{\varepsilon}|\boldsymbol{X})\,\boldsymbol{M'_X} = \sigma^2\boldsymbol{M_X M'_X} = \sigma^2\boldsymbol{M_X}$$

    Hence residuals are generally heteroskedastic and correlated (and hence not independent).

- **Estimation of** σ^2:

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

  is an unbiased estimator of σ^2

  Proof. See Appendix A.8. □

- **Standard error**:
  - It is an estimator of the standard deviation of each element of $\widehat{\boldsymbol{\beta}}$. It permits to assess the accuracy of $\widehat{\boldsymbol{\beta}}$.
  - The variance is estimated by $s^2\left(\boldsymbol{X'X}\right)^{\boldsymbol{-1}}$ and the standard errors are the square roots of the diagonal elements. In summary:
    - Variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$: $\operatorname{Var}\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \sigma^2\left(\boldsymbol{X'X}\right)^{\boldsymbol{-1}}$.
    - σ^2 is unobserved and is estimated by s^2.
    - Estimator of the variance of $\boldsymbol{\beta}$: $\widehat{\operatorname{Var}}\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) = s^2\left(\boldsymbol{X'X}\right)^{\boldsymbol{-1}}$.
    - Standard error: $s.e.\left(\widehat{\beta}_k\right) = \sqrt{\left(s^2\left(\boldsymbol{X'X}\right)^{\boldsymbol{-1}}\right)_{(k,k)}}$

- **The** $\bar{R}^2$:

  The R^2 does not use unbiased estimators. Therefore, it necessarily increases (or stays the same) as we add regressors.

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

  When K increases, $RSS = \|\widehat{\boldsymbol{\varepsilon}}\|^2$ decreases mechanically (or stays the same). So, we correct for this by using $n-K$ in the numerator. Hence $\bar{R}^2$ increases only if the decrease in RSS more than compensates the increase in K.

Properties of the OLS Estimator with Normal Errors

Theorem 1.2.4. *If* $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ *and under assumptions* IID *and* ERRORS*,* $\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}\left(\boldsymbol{X\beta}, \boldsymbol{\sigma^2}\mathbf{I}\right)$ *and under Assumption* RANK*,*

1. $\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{\beta}, \boldsymbol{\sigma^2}(\boldsymbol{X'X})^{-1})$.
2. *OLS estimator is efficient in the class of unbiased estimators.*
3. $\widehat{\boldsymbol{y}}|\boldsymbol{X} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{X\beta}, \boldsymbol{\sigma^2 P_X})$ *and is independent from* $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{M_X y}$ *given* $\boldsymbol{X}$.
4. $s^2/\sigma^2 \sim \chi^2_{n-K}/(n-K)$ *and is independent of* $\widehat{\boldsymbol{y}}$.

Proof. See Appendix A.9. □

Adopting a distribution for $\boldsymbol{\varepsilon}$ permits to characterize results about the exact distributions of $\widehat{\boldsymbol{y}}$, $\widehat{\boldsymbol{\varepsilon}}$, $\widehat{\boldsymbol{\beta}}$ and s^2, which allows to perform tests on parameters.

But without assuming normality of the errors, we can use results from asymptotic theory to obtain approximate distributions for $\widehat{\boldsymbol{y}}$, $\widehat{\boldsymbol{\varepsilon}}$, $\widehat{\boldsymbol{\beta}}$ and s^2.

1.2.3 Convergence

To study asymptotic properties of the OLS estimator (convergence and asymptotic normality), we need to use definitions and results from asymptotic theory. These concepts are reminded in what follows, and are considered as a prerequisite for the course. Students not at ease with such concepts should refer to the prerequisite material.

We start by defining how a random vector that depends on the sample size, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, is considered as being arbitrarily close to a vector $\boldsymbol{\theta}$ when n is sufficiently large (i.e. $n \to +\infty$). We say that $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ **converges** to $\boldsymbol{\theta}$. There are different ways to define asymptotic convergence and we give now two convergence criteria.

Definition. $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ **converges in quadratic mean** to $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \xrightarrow{qm} \boldsymbol{\theta}$, if

$$\lim_{n\to+\infty} \mathrm{E}\, \|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}\|^2 = \mathbf{0}\,.$$

Definition. $\widehat{\boldsymbol{\theta}}_n$ **converges in probability (or weakly)** to $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{\theta}$, if for all $\epsilon > 0$

$$\lim_{n\to+\infty} \Pr\left(\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}\| > \epsilon\right) = 0 \Leftrightarrow \lim_{n\to+\infty} \Pr\left(\|\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}\| < \epsilon\right) = 1\,.$$

From Chebychev's inequality, convergence in quadratic mean implies convergence in probability. In this course, we will use extensively the convergence in probability criterion.

The next theorem is a result that permits to establish convergence in probability of a sample average.

Theorem 1.2.5 (Khinchine's Weak law of large numbers)**.** *Let $y_1, \ldots, y_n$, be n i.i.d. with finite mean $E(y_i) = \mu < \infty$. Then*

$$\bar{y} \xrightarrow{p} \mu\,,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean.

This theorem states the conditions under which a sample average, $\frac{1}{n}\sum_{i=1}^{n} y_i$, which can be considered as the mean of the *empirical* distribution $(y_1, \ldots, y_n)$, converges to the expected value of the random variable y, which is the mean of the *population* distribution. For this convergence to happen, it has to be that the n realizations $(y_1, \ldots, y_n)$ should be independent and taken from the same distribution as y, whose mean is finite.

EXERCISE 1.6. *Show that for a vector* $\boldsymbol{u}$*,* $\bar{\boldsymbol{u}} \xrightarrow{p} \mathrm{E}(\boldsymbol{u})$ *as soon as each components of* $\bar{\boldsymbol{u}}$ *converges weakly to the corresponding component of* $\mathrm{E}\,\boldsymbol{u}$*. Show that the converse is also true.*

As will become clearer soon, the OLS estimator can be written as a function of a sample average, so we will apply the LLN to this sample average to determine its probability limit. To go from this result to the convergence of the OLS estimator, we will need to consider the convergence of a function of a sample average. This is given by the next theorem.

Theorem 1.2.6 (CONTINUOUS MAPPING THEOREM (I))**.** *If* $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \xrightarrow{p} \boldsymbol{\theta}$ *and* $g(\cdot)$ *is continuous at* $\boldsymbol{\theta}$*, then* $g(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}) \xrightarrow{p} g(\boldsymbol{\theta})$*.*

Application: If $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \xrightarrow{\boldsymbol{p}} \boldsymbol{\theta}$, then $\boldsymbol{a'}\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} \xrightarrow{\boldsymbol{p}} \boldsymbol{a'\theta}$ for all $\boldsymbol{a}$.

We now have all the tools needed to show the convergence of the OLS estimator.

Theorem 1.2.7. *Under Assumptions* IID*,* ERRORS*,* MOMENTS*,* RANK*,*

$$n^{-1}\boldsymbol{X'X} \xrightarrow{p} \boldsymbol{Q} = \mathrm{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right),$$

and

$$n^{-1}\boldsymbol{X'\varepsilon} \xrightarrow{p} \boldsymbol{0},$$

so that

$$\widehat{\boldsymbol{\beta}} \xrightarrow{\boldsymbol{p}} \boldsymbol{\beta}.$$

If, in addition, we assume VARIANCE*, then also*

$$s^2 \xrightarrow{p} \sigma^2.$$

Proof. Intuition: we will apply the LLN. In order to do so, we need, first, to write the OLS estimator as an average of random variables and, then, apply the continuous mapping theorem.

We first show the consistency of $\widehat{\boldsymbol{\beta}}$:

- $\widehat{\boldsymbol{\beta}}$ can be written this way:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}.$$

- We consider the element $\boldsymbol{X'X}$ (matrix of dimension $(K \times K)$):
  - Each element (k,l) of $\boldsymbol{X'X}$ is of the form:

$$\left(\boldsymbol{X'X}\right)_{(k,l)} = \sum_{i=1}^{n} x_{ki}x_{li} \qquad \text{for } k,l = 1,\ldots,K.$$

  - Using Iɪᴅ, we assume that the expected value of $x_{ki}x_{li}$ is $\mathrm{E}\left(x_{ki}x_{li}\right) = \mathrm{E}\left(x_k x_l\right)$ for all $i = 1, \ldots, n$. Under the Mᴏᴍᴇɴᴛs assumption, we have $\mathrm{E}\left(x_k x_l\right) < +\infty$. Therefore, by the WLLN:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ki}x_{li} \xrightarrow{p} \mathrm{E}\left(x_k x_l\right).$$

    Implication: there is a plim (probability limit) for the $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ matrix: $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{Q}$ where $\boldsymbol{Q}$ is a finite, deterministic, positive definite matrix because of Rᴀɴᴋ. This is a result that will also be used later.

    Note here that we only require moments of order 2 to be finite, which is less restrictive than the Mᴏᴍᴇɴᴛs assumption.

- Now, we consider the element $\boldsymbol{X}'\boldsymbol{\varepsilon}$ (vector of dimension $(K \times 1)$):
  - $\boldsymbol{X}'\boldsymbol{\varepsilon}$ writes

$$\boldsymbol{X}'\boldsymbol{\varepsilon} = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}\varepsilon_i \\ \sum_{i=1}^{n} x_{2i}\varepsilon_i \\ \vdots \\ \sum_{i=1}^{n} x_{Ki}\varepsilon_i \end{pmatrix}.$$

  - Let's consider a particular element $\sum_{i=1}^{n} x_{ki}\varepsilon_i$:
    * Assuming Iɪᴅ, we get $\mathrm{E}\left(x_{ki}\varepsilon_i\right) = \mathrm{E}\left(x_k\varepsilon\right)$ for all $i = 1, \ldots, n$.
    * Using the law of iterated expectations, we get $\mathrm{E}\left(x_k\varepsilon\right) = \mathrm{E}\left(\mathrm{E}\left(x_k\varepsilon | x_k\right)\right) = \mathrm{E}\left(x_k \mathrm{E}\left(\varepsilon | x_k\right)\right)$.
    * Assumption Eʀʀᴏʀs gives $\mathrm{E}\left(\varepsilon | x_k\right) = 0$, which implies $\mathrm{E}\left(x_k\varepsilon\right) = 0$.
    * Therefore, by the WLLN:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ki}\varepsilon_i \xrightarrow{p} \mathrm{E}\left(x_k\varepsilon\right) = 0$$

  - Given this reasoning applies to all elements $k = 1, \ldots, K$ of the vector $\boldsymbol{X}'\boldsymbol{\varepsilon}$, this implies $\frac{1}{n}\boldsymbol{X}'\boldsymbol{\varepsilon} \xrightarrow{p} \boldsymbol{0}$.
- Conclusion: by the continuous mapping theorem, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} + \boldsymbol{Q} \cdot \boldsymbol{0}$, which implies

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}.$$

for the consistency of s^2, see Appendix B.3. □

This convergence theorem shows that the OLS estimator is a **consistent estimator** of the parameters of the model under the assumptions Iɪᴅ, Eʀʀᴏʀs, Mᴏᴍᴇɴᴛs and Rᴀɴᴋ. It means that the estimator ($\widehat{\boldsymbol{\beta}}$) tends to the true value of the parameter ($\boldsymbol{\beta}$) as the sample size tends to infinity.

1.2.4 Asymptotic Distribution

In this section, we will show that the OLS estimator's distribution can be approximated when the sample size is large enough.

Definition. Let $\boldsymbol{S_n}$ be a sequence of random vectors with cumulative distribution functions F_{S_n}. This sequence **converges in law (or in distribution)** to $\boldsymbol{S}$, with cumulative distribution function F_S, if

$$F_{S_n}(t) \to F_S(t)$$

for all t at which $F_S(\cdot)$ is continuous. This is denoted $\boldsymbol{S_n} \xrightarrow{d} \boldsymbol{S}$ or $\boldsymbol{S_n} \xrightarrow{L} \boldsymbol{S}$.

We also say that "$\boldsymbol{S_n}$ is distributed asymptotically as $\boldsymbol{S}$", or, "As $n \to +\infty$ the random variable $\boldsymbol{S_n}$ tends to a random variable that follows a F_S distribution".

To study the asymptotic distribution of the OLS estimator, we will apply the following theorem, that establishes the conditions under which a sample average converges asymptotically to a Normal distribution.

Theorem 1.2.8 (Central Limit Theorem)**.** *Let $y_1, \ldots, y_n$ be i.i.d. of expected value μ and variance σ^2. Then*

$$\sqrt{n}\frac{\bar{y} - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)\,.$$

Remark that $\frac{\bar{y}-\mu}{\sigma}$ is the **normalized sample average**, i.e. the sample average centered around the expected value μ of the random variables y_i and divided by their common standard deviation.

We say that $\bar{y}$ converges to μ at rate $\sqrt{n}$, and that the *rate of convergence* of $\bar{y}$ to a normal distribution is $\sqrt{n}$.

In some references you will see the equivalent notation

$$\bar{y} \overset{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

which means that asymptotically $\bar{y}$ is distributed as a normal with mean μ and variance $\frac{\sigma^2}{n}$. You can use this notation if it suits you better. Be careful not to confuse the notations and don't write

$$\bar{y} \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

because what is on the right side of the arrow should be a limit distribution, and thus cannot depend on n.

As we work with vectors of random variables, we need to work with a multivariate version of the CLT.

Theorem 1.2.9 (Multivariate Central Limit Theorem)**.** *Let $\boldsymbol{w_1} \dots \boldsymbol{w}_n$ be i.i.d. vectors of $\mathbb{R}^p$ with expected value $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Then*

$$\sqrt{n}\,(\bar{\boldsymbol{w}} - \boldsymbol{\mu}) \xrightarrow{d} \boldsymbol{\mathcal{N}(0, \Sigma)}\,.$$

As was the case for the consistency result, we also need a result to be able to work with a function of a sample average.

Theorem 1.2.10 (Continuous Mapping Theorem (II))**.** *If $\boldsymbol{S_n} \xrightarrow{d} \boldsymbol{S}$ and $g(\cdot)$ is continuous, then $g(\boldsymbol{S_n}) \xrightarrow{d} g(\boldsymbol{S})$.*

Exercise 1.7. *Show that if $\boldsymbol{Z_n} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$ of dimension p, then $\|\boldsymbol{Z_n}\|^2 \xrightarrow{d} \chi^2_p$.*

We also need a result that permits to "combine" convergence in distribution and convergence in probability.

Theorem 1.2.11 (Slutsky)**.** *If $\boldsymbol{S_n} \xrightarrow{d} \boldsymbol{S}$ and $\boldsymbol{A_n} \xrightarrow{p} \boldsymbol{A}$, then*

- $\boldsymbol{A_n} + \boldsymbol{S_n} \xrightarrow{d} \boldsymbol{A} + \boldsymbol{S}$
- $\boldsymbol{A_n S_n} \xrightarrow{d} \boldsymbol{AS}$
- *For a scalar $\boldsymbol{A} \neq 0$, $\boldsymbol{S_n/A_n} \xrightarrow{d} \boldsymbol{S/A}$.*

Exercise 1.8. *For a random sample $(y_1, \dots y_n)$ from y with $\mathrm{E}\, y = \mu$ and $\mathrm{Var}\, y = \sigma^2$, show that the t-statistic is such that*

$$\sqrt{n}\frac{\overline{y} - \mu}{s} \xrightarrow{d} \mathcal{N}(0, 1)\,.$$

Theorem 1.2.12. *Under Assumptions* Iid, Errors, Moments, Rank, Variance

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \boldsymbol{\mathcal{N}(0, \sigma^2 Q^{-1})}\,.$$

Proof. Intuition: we will apply the CLT. In order to do so, we need, first, to write the OLS estimator as an average of random variables and, then, apply Slutsky.

- Rewrite $\widehat{\boldsymbol{\beta}}$ as:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{\varepsilon}$$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \sqrt{n}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{\varepsilon}$$

- We have already shown that using IID, RANK and MOMENTS, the weak law of large numbers gives $\frac{1}{n}\boldsymbol{X'X} \xrightarrow{p} \boldsymbol{Q} = \mathrm{E}\left(\boldsymbol{xx'}\right)$
  So: $n\left(\boldsymbol{X'X}\right)^{-1} \xrightarrow{p} \boldsymbol{Q^{-1}}$
- For $\boldsymbol{X'\varepsilon}$:
  - Express $\boldsymbol{X'\varepsilon}$ as a sum of individual elements:

$$\boldsymbol{X'\varepsilon} = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}\varepsilon_i \\ \sum_{i=1}^{n} x_{2i}\varepsilon_i \\ \vdots \\ \sum_{i=1}^{n} x_{Ki}\varepsilon_i \end{pmatrix} = \sum_{i=1}^{n} \boldsymbol{x_i}\varepsilon_i \,.$$

  - We want to apply the CLT to the vector $\boldsymbol{z_i} = \boldsymbol{x_i}\varepsilon_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{pmatrix} \varepsilon_i$.
    * $\mathrm{E}\left(\boldsymbol{x_i}\varepsilon_i\right) = \begin{pmatrix} \mathrm{E}\left(x_{1i}\varepsilon_i\right) \\ \mathrm{E}\left(x_{2i}\varepsilon_i\right) \\ \vdots \\ \mathrm{E}\left(x_{Ki}\varepsilon_i\right) \end{pmatrix} = \boldsymbol{0}$, because of IID and ERRORS.
    * $\mathrm{Var}\left(\boldsymbol{x_i}\varepsilon_i\right) = \sigma^2\boldsymbol{Q}$ since $\boldsymbol{Q} = \mathrm{E}\left(\boldsymbol{x_i x_i'}\right)$.
      Indeed,

$$\begin{aligned}
\mathrm{Var}\left(\boldsymbol{x_i}\varepsilon_i\right) &= \mathrm{E}\left[(\boldsymbol{x_i}\varepsilon_i - \mathrm{E}\left[\boldsymbol{x_i}\varepsilon_i\right])(\boldsymbol{x_i}\varepsilon_i - \mathrm{E}\left[\boldsymbol{x_i}\varepsilon_i\right])'\right] \\
&= \mathrm{E}\left[(\boldsymbol{x_i}\varepsilon_i)(\boldsymbol{x_i}\varepsilon_i)'\right] \\
&= \mathrm{E}\left[\varepsilon_i^2\boldsymbol{x_i x_i'}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\varepsilon_i^2\boldsymbol{x_i x_i'}|\boldsymbol{x_i}\right]\right] \\
&= \mathrm{E}\left[\boldsymbol{x_i x_i'}\mathrm{E}\left[\varepsilon_i^2|\boldsymbol{x_i}\right]\right] \\
&= \mathrm{E}\left[\sigma^2\boldsymbol{x_i x_i'}\right] \quad \text{using VARIANCE} \\
&= \sigma^2\mathrm{E}\left[\boldsymbol{x_i x_i'}\right]
\end{aligned}$$

  - Apply the Central Limit Theorem:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x_i}\varepsilon_i - \mathrm{E}\left(\boldsymbol{x_i}\varepsilon_i\right)\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \mathrm{Var}\left(\boldsymbol{x_i}\varepsilon_i\right)\right)$$

$$\frac{1}{\sqrt{n}}\left(\boldsymbol{X'\varepsilon}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{Q}\right)\,.$$

- Then, apply the Slutsky theorem:

$$\left.\begin{array}{c} n\left(\boldsymbol{X'X}\right)^{-1} \xrightarrow{p} \boldsymbol{Q^{-1}} \\ \frac{1}{\sqrt{n}}\left(\boldsymbol{X'\varepsilon}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{Q}\right) \end{array}\right\} \Rightarrow \sqrt{n}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'\varepsilon} \xrightarrow{d} \boldsymbol{Q^{-1}}\mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{Q}\right)$$

$$\boldsymbol{Q^{-1}}\mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{Q}\right) = \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{Q^{-1}}\sigma^2\boldsymbol{QQ^{-1}}\right)$$

  So:

$$\sqrt{n}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'\varepsilon} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{Q^{-1}}\right)$$

□

EXERCISE 1.9. *Show that for any element $\widehat{\beta}_k$ with standard error se_k,*

$$\frac{\widehat{\beta}_k - \beta_k}{se_k} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Appendix A

Prerequisites

A.1 Concepts of Matrix Algebra

- **Vector** of dimension n (n-vector): set of n elements:

$$\boldsymbol{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}' \in \mathbb{R}^n$$

  ($\mathbb{R}^n$: Euclidian space)
- **Vector space**: set of elements (vectors) that can be combined by the elementary operations (vector addition and scalar multiplication) to yield other elements of the same set. Example: $\mathbb{R}^n$
- **Scalar product** (between 2 vectors): $<\boldsymbol{x}, \boldsymbol{y}>= \boldsymbol{x}'\boldsymbol{y}$
- **Norm** of a vector: $\|\boldsymbol{y}\| = \sqrt{\boldsymbol{y}'\boldsymbol{y}}$. Recall that $\|\boldsymbol{y}\|^2 = \boldsymbol{y}'\boldsymbol{y} = \sum_{i=1}^n y_i^2$
- A vector linear **subspace** S of a vector space V is called a subspace of V if $\forall \boldsymbol{u}, \boldsymbol{v} \in S$ and $\forall a, b \in \mathbb{R}$, $a \cdot \boldsymbol{u} + b \cdot \boldsymbol{v} \in S$.
- The **spanned subspace** of $\boldsymbol{X}$ (or the vector subspace spanned by $\boldsymbol{X}$), $\mathrm{L}(\boldsymbol{X}) = \mathrm{L}(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_K})$, is the set containing all the vectors that can be formed as a linear combination of the $\boldsymbol{x_k}$, $k = 1, \ldots, K$.

$$\mathrm{L}(\boldsymbol{X}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n | \boldsymbol{z} = \sum_{k=1}^K \lambda_k \boldsymbol{x_k} \qquad \forall \boldsymbol{x_1}, \ldots, \boldsymbol{x_K} \in \boldsymbol{X} \text{ and } \lambda_1, \ldots, \lambda_K \in \mathbb{R} \right\}$$

- The **orthogonal complement** of $\mathrm{L}(\boldsymbol{X})$ in $\mathbb{R}^n$ is the set of all vectors $\boldsymbol{\omega} \in \mathbb{R}^n$ that are orthogonal to any vector $\boldsymbol{z} \in \mathrm{L}(\boldsymbol{X})$.

$$\mathrm{L}^{\perp}(\boldsymbol{X}) = \left\{ \boldsymbol{\omega} \in \mathbb{R}^n | \boldsymbol{\omega}'\boldsymbol{z} = 0 \qquad \forall \boldsymbol{z} \in \mathrm{L}(\boldsymbol{X}) \right\}$$

- A matrix $\boldsymbol{X}$ is of **full rank** if the K vectors that form it are linearly independent
- A symmetric matrix $\boldsymbol{A}$ is **positive definite** if $\forall \boldsymbol{x} \neq \boldsymbol{0}$ $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$.
- A symmetric matrix $\boldsymbol{A}$ is **positive semi-definite** if $\forall \boldsymbol{x} \neq \boldsymbol{0}$ $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$.
  - A matrix $\boldsymbol{A}'\boldsymbol{A}$ is always positive semi-definite as $\boldsymbol{x}'(\boldsymbol{A}'\boldsymbol{A})\boldsymbol{x} = \|\boldsymbol{A}\boldsymbol{x}\|^2 \geq 0 \quad \forall \boldsymbol{x} \neq \boldsymbol{0}$
  - Moreover, if $\boldsymbol{A}$ is full rank, $\boldsymbol{A}'\boldsymbol{A}$ is positive definite (since $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0} \Rightarrow \boldsymbol{x} = \boldsymbol{0}$)
- A system of vectors $(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_K})$ is said to be linearly independent if

$$\forall \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K) \qquad \sum_{k=1}^K \lambda_k \boldsymbol{x_k} = \boldsymbol{0} \Rightarrow \lambda_1 = \lambda_2 = \ldots = \lambda_K = 0$$

  In contrast if a vector is a linear combination of other vectors, we have multicollinearity.

- **Derivatives with vectors and matrices**:
  - For vectors $\boldsymbol{u}$ and $\boldsymbol{a}$ of size p,

$$\frac{\partial \boldsymbol{a}'\boldsymbol{u}}{\partial \boldsymbol{u}} = \frac{\partial \boldsymbol{u}'\boldsymbol{a}}{\partial \boldsymbol{u}} = \begin{pmatrix} \frac{\partial \boldsymbol{u}'\boldsymbol{a}}{\partial \boldsymbol{u}_1} \\ \dots \\ \frac{\partial \boldsymbol{u}'\boldsymbol{a}}{\partial \boldsymbol{u}_p} \end{pmatrix} = \boldsymbol{a}\,.$$

  - For a *symmetric* $p \times p$ matrix $\boldsymbol{A}$

$$\frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}} = \begin{pmatrix} \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_1} \\ \dots \\ \frac{\partial \boldsymbol{a}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_p} \end{pmatrix} = 2\boldsymbol{A}\boldsymbol{u}\,.$$

  - **Hessian**: symmetric matrix of second order derivatives. For instance,

$$\begin{aligned} \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}\partial \boldsymbol{u}'} &\equiv \left(\frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_j}; i = 1, \dots p \quad j = 1, \dots p\right) \\ &= \begin{pmatrix} \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial^2 \boldsymbol{u}_1} & \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_1 \partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_1 \partial \boldsymbol{u}_p} \\ \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_2 \partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial^2 \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_2 \partial \boldsymbol{u}_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_p \partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial \boldsymbol{u}_p \partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}}{\partial^2 \boldsymbol{u}_p} \end{pmatrix} \\ &= 2\boldsymbol{A}\,. \end{aligned}$$

- **Pythagora's Theorem:**
  Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors of $\mathbb{R}^n$. If $\boldsymbol{a}$ and $\boldsymbol{b}$ are orthogonal, i.e. $\boldsymbol{a'b} = 0$, then $\|\boldsymbol{a} + \boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2$.

A.2 Concepts of Probability

Expectations, Variance, Covariance of Random Vectors

- The **expectation of a random vector** is the vector containing the expectations of its individual elements, i.e.

$$\mathrm{E}\,(\boldsymbol{x}) = \begin{pmatrix} \mathrm{E}\,(x_1) \\ \mathrm{E}\,(x_2) \\ \vdots \\ \mathrm{E}\,(x_K) \end{pmatrix}.$$

- The **expectation of a random matrix** is the matrix containing the corresponding expectations of its individual elements.
- **Linearity of expectation**: For a random vector $\boldsymbol{x}$ and constant (non-random) matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\mathrm{E}\,(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\mathrm{E}\,(\boldsymbol{x}) \qquad \text{and} \qquad \mathrm{E}\,(\boldsymbol{x}\boldsymbol{B}) = \mathrm{E}\,(\boldsymbol{x})\boldsymbol{B}$$

- Covariance matrix of two random vectors: collects the covariances of each pair of elements from the two random vectors.

The **covariance matrix** of $\boldsymbol{y}$ and $\boldsymbol{z}$ is

$$\begin{aligned}
\operatorname{Cov}(\boldsymbol{y},\boldsymbol{z}) &\equiv (\operatorname{Cov}(y_i,z_j); i=1,\dots,m, j=1,\dots,n)\\
&= (\mathrm{E}\left[(y_i-\mathrm{E}\,(y_i))(z_i-\mathrm{E}\,(z_i))\right]; i=1,\dots,m, j=1,\dots,n)\\
&= \mathrm{E}\left[(\boldsymbol{y}-\mathrm{E}\,(\boldsymbol{y}))(\boldsymbol{z}-\mathrm{E}\,(\boldsymbol{z}))'\right]\\
&= \begin{pmatrix}
\operatorname{Cov}(y_1,z_1) & \operatorname{Cov}(y_1,z_2) & \dots & \operatorname{Cov}(y_1,z_n)\\
\operatorname{Cov}(y_2,z_1) & \operatorname{Cov}(y_2,z_2) & \dots & \operatorname{Cov}(y_2,z_n)\\
\vdots & \vdots & \ddots & \vdots\\
\operatorname{Cov}(y_m,z_1) & \operatorname{Cov}(y_m,z_2) & \dots & \operatorname{Cov}(y_m,z_n)
\end{pmatrix}
\end{aligned}$$

- **Bilinearity of covariance**: For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\operatorname{Cov}(\boldsymbol{Ay},\boldsymbol{Bz}) = \boldsymbol{A}\operatorname{Cov}(\boldsymbol{y},\boldsymbol{z})\boldsymbol{B'}$.
  Proof:

$$\begin{aligned}
\operatorname{Cov}(\boldsymbol{Ay},\boldsymbol{Bz}) &= \mathrm{E}\left[(\boldsymbol{Ay}-\mathrm{E}\,(\boldsymbol{Ay}))(\boldsymbol{Bz}-\mathrm{E}\,(\boldsymbol{Bz}))'\right]\\
&= \mathrm{E}\left[\boldsymbol{A}(\boldsymbol{y}-\mathrm{E}\,(\boldsymbol{y}))(\boldsymbol{z}-\mathrm{E}\,(\boldsymbol{z}))'\boldsymbol{B}'\right]\\
&= \boldsymbol{A}\mathrm{E}\left[(\boldsymbol{y}-\mathrm{E}\,(\boldsymbol{y}))(\boldsymbol{z}-\mathrm{E}\,(\boldsymbol{z}))'\right]\boldsymbol{B}'
\end{aligned}$$

- Variance of a random vector, called "*variance-covariance matrix*" or simply "*variance matrix*": is the covariance matrix of this vector with itself.
  The **variance-covariance matrix** of $\boldsymbol{y}$ is

$$\begin{aligned}
\operatorname{Var}(\boldsymbol{y}) &\equiv \operatorname{Cov}(\boldsymbol{y},\boldsymbol{y})\\
&= \mathrm{E}\left[(\boldsymbol{y}-\mathrm{E}\,(\boldsymbol{y}))(\boldsymbol{y}-\mathrm{E}\,(\boldsymbol{y}))'\right]\\
&= \begin{pmatrix}
\operatorname{Var}(y_1) & \operatorname{Cov}(y_1,y_2) & \dots & \operatorname{Cov}(y_1,y_m)\\
\operatorname{Cov}(y_1,y_2) & \operatorname{Var}(y_2) & \dots & \operatorname{Cov}(y_2,y_m)\\
\vdots & \vdots & \ddots & \vdots\\
\operatorname{Cov}(y_1,y_m) & \operatorname{Cov}(y_2,y_m) & \dots & \operatorname{Var}(y_m)
\end{pmatrix}
\end{aligned}$$

- Variance of a linear combination of random variables: $\operatorname{Var}(\boldsymbol{Ay}) = \boldsymbol{A}\operatorname{Var}(\boldsymbol{y})\boldsymbol{A'}$ using bilinearity.
  - Proof: $\operatorname{Var}(\boldsymbol{Ay}) = \operatorname{Cov}(\boldsymbol{Ay},\boldsymbol{Ay}) = \boldsymbol{A}\operatorname{Var}(\boldsymbol{y})\boldsymbol{A'}$
  - Remark: this property for a two-dimensional vector is is

$$\operatorname{Var}(ax+by) = a^2\operatorname{Var}(x) + b^2\operatorname{Var}(y) + 2ab\operatorname{Cov}(x,y)$$

- Var $(\boldsymbol{y})$ is a symmetric positive semidefinite matrix: because of the above property, $\boldsymbol{a}'\operatorname{Var}(\boldsymbol{y})\boldsymbol{a} = \operatorname{Var}(\boldsymbol{a'y}) \geq 0$ for any $\boldsymbol{a}$.

Normal Vector

- $\boldsymbol{y}$ is a normal vector $\boldsymbol{\mathcal{N}(\mu,\Sigma)}$ if and only if $\boldsymbol{a'y}$ is univariate normal for all $\boldsymbol{a}$. Then $\boldsymbol{a'y} \sim \boldsymbol{\mathcal{N}(a'\mu, a'\Sigma a)}$
- If $\boldsymbol{y}$ is a normal vector $\boldsymbol{\mathcal{N}(\mu,\Sigma)}$, $\boldsymbol{\Sigma}$ positive definite, then $\boldsymbol{y = Az + \mu}$ with $\boldsymbol{z} \sim \boldsymbol{\mathcal{N}(0,\mathrm{I})}$, $\boldsymbol{A}$ square matrix of full rank, and $\boldsymbol{AA' = \Sigma}$.
- For normal vectors, independence and null covariance are equivalent.

Chi-Square and Student Distribution

Definition 1 (Chi-square with p degrees of freedom)**.** If $\boldsymbol{z} \sim \boldsymbol{\mathcal{N}(0,\mathrm{I})}$ of dimension p, then $\|\boldsymbol{z}\|^{\boldsymbol{2}} \sim \chi^2_p$

- If $\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\Sigma})$ of dimension p, then $\boldsymbol{y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \sim \chi^2_p$
- If $\boldsymbol{z} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})$ of dimension p and $\boldsymbol{P}$ is an orthogonal projection of rank m, then $\boldsymbol{z}'\boldsymbol{P}\boldsymbol{z} \sim \chi^2_m$

Definition 2 (Student T with p degrees of freedom)**.** If $\boldsymbol{z} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})$ of dimension p and $w \sim \mathcal{N}(0,1)$ independent of $\boldsymbol{z}$, then

$$\frac{w}{\|z\|^2} \sim T_p \,.$$

A.3 Concepts of Estimation and Inferential Statistics

Estimator and estimated value

We are interested in an unknown parameter $\boldsymbol{\theta}$, say sample mean, and we denote its true value as $\boldsymbol{\theta}^*$. An **estimator** $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is a function of the sample observations (statistic), or formula to get a "guess" for $\boldsymbol{\theta}^*$. Given a particular sample of observations, the value given by the estimator is the **estimated value**.

Bias

- **Bias** of $\widehat{\boldsymbol{\theta}}$ is $\mathrm{E}\,(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}^*$: expectation of $\widehat{\boldsymbol{\theta}}$ minus true value
- $\widehat{\boldsymbol{\theta}}$ is an **unbiased estimator** if $\mathrm{E}\,(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}^*$ (bias is zero).

Quality of an estimator

- For a scalar estimator $\widehat{\gamma}$, the quality is often measured by the **Mean Squared Error** (MSE),
  - Lower MSE ⇔ better quality.

  $$\mathrm{E}\,(\widehat{\gamma} - \gamma)^2 = \mathrm{E}\,(\widehat{\gamma} - \mathrm{E}\,\widehat{\gamma})^2 + (\mathrm{E}\,\widehat{\gamma} - \gamma)^2 = \text{Variance} + \text{Bias}^2$$

  When the estimator is unbiased, MSE = Variance. The best estimator is the one with the lowest variance.
- For a vector estimator $\widehat{\boldsymbol{\theta}}$, we consider the MSE of any linear combination $\boldsymbol{a}'\widehat{\boldsymbol{\theta}}$

  $$\begin{aligned} \mathrm{E}\left[\left(\boldsymbol{a}'\widehat{\boldsymbol{\theta}} - \boldsymbol{a}'\boldsymbol{\theta}\right)^2\right] &= \boldsymbol{a}'\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\right]\boldsymbol{a} \\ &= \boldsymbol{a}'\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{a} + \boldsymbol{a}'\left(\mathrm{E}\,(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right)\left(\mathrm{E}\,(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}\right)'\boldsymbol{a}\,. \end{aligned}$$

  When the estimator is unbiased, $\mathrm{E}\left[\left(\boldsymbol{a}'\widehat{\boldsymbol{\theta}} - \boldsymbol{a}'\boldsymbol{\theta}\right)^2\right] = \mathrm{Var}\left(\boldsymbol{a}\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{a}'\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)\boldsymbol{a}$.

Definition. Let $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}}$ be two different estimators (possibly conditionally) unbiased of $\boldsymbol{\theta} \in \mathbb{R}^p$. Then $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$ is more **efficient** than $\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}}$ if $\mathrm{Var}\,\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}} - \mathrm{Var}\,\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$ (where possibly variances are conditional) is positive semi-definite (but not equal to zero).

The definition thus entails that if $\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$ is more efficient than $\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}}$, any linear combination $\boldsymbol{a}'\boldsymbol{\theta}$ is more accurately estimated with $\boldsymbol{a}'\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}$, because

$$\mathrm{Var}\,\boldsymbol{a}'\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}} - \mathrm{Var}\,\boldsymbol{a}'\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}} = \boldsymbol{a}'\left(\mathrm{Var}\,\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}} - \mathrm{Var}\,\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}\right)\boldsymbol{a} \geq 0 \Leftrightarrow \mathrm{Var}\,\boldsymbol{a}'\widehat{\boldsymbol{\theta}}_{\boldsymbol{B}} \geq \mathrm{Var}\,\boldsymbol{a}'\widehat{\boldsymbol{\theta}}_{\boldsymbol{A}}\,.$$

Moreover, because the matrices are different, there is at least one $\boldsymbol{a}$ such that the inequality is strict.

A.4 Derivation of the OLS Estimator

The *Ordinary Least Squares (OLS)* estimator is

$$\widehat{\boldsymbol{\beta}} = \arg\min s(\boldsymbol{\beta}),$$

$$\begin{aligned} \text{where} \qquad s(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i'}\boldsymbol{\beta}\right)^2 \qquad \text{(A.1)} \\ &= \| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y'y} - 2\boldsymbol{y'X\beta} + \boldsymbol{\beta' X' X \beta} \end{aligned}$$

- Note: since $\boldsymbol{y'X\beta}$ is a scalar, $(\boldsymbol{y'X\beta})' = \boldsymbol{y'X\beta} \Leftrightarrow \boldsymbol{y'X\beta} = (\boldsymbol{X\beta})'\boldsymbol{y} = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$.
- This is a quadratic optimization problem in $\boldsymbol{\beta}$.
- To get $\widehat{\boldsymbol{\beta}}$, look at first order conditions

  $$\frac{\partial}{\partial \boldsymbol{\beta}} s(\widehat{\boldsymbol{\beta}}) = -2\boldsymbol{X'y} + 2\boldsymbol{X'X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$$

  so that

  $$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}.$$

  if $(\boldsymbol{X'X})$ is not singular, i.e. $\boldsymbol{X}$ of full rank $\text{Rank}(\boldsymbol{X}) = K$.
- Check that the problem is convex

  $$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} s(\boldsymbol{\beta}) = 2\boldsymbol{X'X}$$

  For $\boldsymbol{X}$ of full rank, the hessian is positive definite, which is a sufficient condition for $\widehat{\boldsymbol{\beta}}$ to be a unique minimum.

A.5 Proof of Theorem 1.2.1 (the OLS Estimator is Unbiased)

$$\begin{aligned} \mathrm{E}\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) &= \mathrm{E}\left((\boldsymbol{X'X})^{-1}\boldsymbol{X'y}|\boldsymbol{X}\right) \quad \text{because of } \textsc{Rank} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathrm{E}\left(\boldsymbol{y}|\boldsymbol{X}\right) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta} \quad \text{because of } \textsc{Iid} \text{ and } \textsc{Errors} \\ &= \boldsymbol{\beta} \end{aligned}$$

A.6 Proof of Theorem 1.2.2 (Variance of the OLS Estimator)

$$\begin{aligned} \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}\right) &= \mathrm{Var}\left((\boldsymbol{X'X})^{-1}\boldsymbol{X'y}|\boldsymbol{X}\right) \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\mathrm{Var}\left(\boldsymbol{y}|\boldsymbol{X}\right)\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\sigma^2\mathbf{I}\boldsymbol{X}(\boldsymbol{X'X})^{-1} \\ &= \sigma^2\left(\boldsymbol{X'X}\right)^{\boldsymbol{-1}} \end{aligned}$$

A.7 Proof of Theorem 1.2.3 (Gauss-Markov)

- Consider another linear estimator $\tilde{\boldsymbol{\beta}}$. We can write:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{Ay} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} + \boldsymbol{Cy} = \widehat{\boldsymbol{\beta}} + \boldsymbol{Cy}$$

- $\tilde{\boldsymbol{\beta}}$ has to be unbiased:

$$\mathrm{E}\left(\tilde{\boldsymbol{\beta}}|\boldsymbol{X}\right) = \boldsymbol{\beta} + \boldsymbol{C}E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\beta} + \boldsymbol{CX\beta}$$

  $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ if $\boldsymbol{CX} = \mathbf{0}$.

- Covariance between $\widehat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right) &= \mathrm{E}\left(\left(\widehat{\boldsymbol{\beta}} - \mathrm{E}\,(\widehat{\boldsymbol{\beta}})\right)\left(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} - \mathrm{E}\,(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\right)'\right) \\
&= \mathrm{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)'\right) \qquad \text{since } \tilde{\boldsymbol{\beta}} \text{ is unbiased}
\end{aligned}$$

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\left(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}\right) - \beta \\
&= (\boldsymbol{X'X})^{-1}\boldsymbol{X'\varepsilon}
\end{aligned}$$

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} &= \boldsymbol{Cy} = \boldsymbol{CX\beta} + \boldsymbol{C\varepsilon} \\
&= \boldsymbol{C\varepsilon} \qquad \text{since } \boldsymbol{CX} = \mathbf{0}
\end{aligned}$$

  Therefore:

$$\begin{aligned}
\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right) &= \mathrm{E}\left((\boldsymbol{X'X})^{-1}\boldsymbol{X'\varepsilon\varepsilon'C'}\right) \\
&= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\mathrm{E}\left(\boldsymbol{\varepsilon\varepsilon'}\right)\boldsymbol{C'} \\
&= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\sigma^2\mathbf{I}\boldsymbol{C'} \\
&= \sigma^2(\boldsymbol{X'X})^{-1}\boldsymbol{X'C'} \\
&= \mathbf{0} \qquad \text{since } \boldsymbol{CX} = \mathbf{0}
\end{aligned}$$

- Let's write the variance of $\tilde{\boldsymbol{\beta}}$:

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\right) &= \mathrm{Var}\left(\widehat{\boldsymbol{\beta}} + (\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\right) \\
&= \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) + \mathrm{Var}\left(\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right) \qquad \text{since } \mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right) = \mathbf{0} \\
&= \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) + \mathrm{Var}\left(\boldsymbol{Cy}\right)
\end{aligned}$$

  So:

$$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}}\right) - \mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left(\boldsymbol{Cy}\right)$$

  Since Var $(\boldsymbol{Cy})$ is a variance-covariance matrix, it is a positive semidefinite matrix. Therefore, $\widehat{\boldsymbol{\beta}}$ is more efficient than $\tilde{\boldsymbol{\beta}}$.

A.8 Proof of Unbiasedness of s^2

- First, consider as an estimator of σ^2 the sample moment, i.e. the average of the n squared residuals:
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 \,.$$
We show that $\widehat{\sigma}^2$ is a biased (though consistent, more on this later) estimator of σ^2:
$$\mathrm{E}\left(\widehat{\sigma}^2|\boldsymbol{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\widehat{\varepsilon}_i^2|\boldsymbol{X}\right) \,.$$
Since Var $(\widehat{\boldsymbol{\varepsilon}}|\boldsymbol{X}) = \sigma^2\boldsymbol{M_X} = \sigma^2(\boldsymbol{I}-\boldsymbol{P_X})$, we have for a particular element i:
$$\mathrm{Var}\,(\widehat{\varepsilon}_i|\boldsymbol{X}) = \mathrm{E}\left(\widehat{\varepsilon}_i^2|\boldsymbol{X}\right) = (1-h_i)\sigma^2 \,,$$
where h_i is the i^{th} diagonal element of matrix $\boldsymbol{P_X}$. Note that Var $(\widehat{\varepsilon}_i|\boldsymbol{X}) < \sigma^2$.
$\sum_{i=1}^{n} h_i$ is the trace of the matrix $\boldsymbol{P_X}$ (sum of its diagonal elements).
$$\begin{aligned}\sum_{i=1}^{n} h_i &= Tr(\boldsymbol{P_X}) = Tr(\boldsymbol{X(X'X)^{-1}X'}) \\ &= Tr(\boldsymbol{(X'X)^{-1}X'X}) \qquad \text{since } Tr(\boldsymbol{AB}) = Tr(\boldsymbol{BA}) \\ &= Tr(\boldsymbol{I_K}) \\ &= K\end{aligned}$$
Therefore:
$$\mathrm{E}\left(\widehat{\sigma}^2|\boldsymbol{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\widehat{\varepsilon}_i^2|\boldsymbol{X}\right) = \frac{1}{n}\sum_{i=1}^{n}(1-h_i)\sigma^2 = \frac{n-K}{n}\sigma^2 < \sigma^2 \,.$$
So, $\widehat{\sigma}^2$ is a downward biased estimator of σ^2.
- In order to get an unbiased estimator, we correct for the bias. A unbiased estimator is:
$$s^2 = \frac{1}{n-K}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$
$$\mathrm{E}\,(s^2|\boldsymbol{X}) = \sigma^2$$

A.9 Proof of Theorem 1.2.4 (Properties of the OLS Estimator with Normal Errors)

Proof.

1. $\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \boldsymbol{\mathcal{N}(\beta, \sigma^2(X'X)^{-1})}$:
   - Expectation and variance of $\widehat{\boldsymbol{\beta}}$ have been shown earlier.
   - $\widehat{\boldsymbol{\beta}}$ is a linear combination of $\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}$. So, $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{\mathcal{N}}$.
2. OLS estimator is efficient in the class of unbiased estimators:
   - To be shown later in the course.
3. $\widehat{\boldsymbol{y}}|\boldsymbol{X} \sim \boldsymbol{\mathcal{N}(X\beta, \sigma^2 P_X)}$ and is independent from $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{M_X y}$ given $\boldsymbol{X}$:
   (a) Expectation and variance of $\widehat{\boldsymbol{y}}$ have been shown earlier.
   (b) $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is a linear combination of $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{\mathcal{N}}$, so $\widehat{\boldsymbol{y}} \sim \boldsymbol{\mathcal{N}}$.

(c) $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is a linear combination of $\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}$ and $\widehat{\boldsymbol{y}} \sim \boldsymbol{\mathcal{N}}$, so $\widehat{\boldsymbol{\varepsilon}} \sim \boldsymbol{\mathcal{N}}$.

(d) We know that $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{\varepsilon}}$ are orthogonal. Here, we show their distributional independence:

- Covariance matrix of $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{\varepsilon}}$ (! Not variance-covariance matrix):

$$\begin{aligned}
\text{Cov}(\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{\varepsilon}}) &= \text{Cov}(\boldsymbol{P_X y}, \boldsymbol{M_X y}) \\
&= \boldsymbol{P_X}\text{Cov}(\boldsymbol{y}, \boldsymbol{y})\boldsymbol{M}'_{\boldsymbol{X}} \\
&= \boldsymbol{P_X}\text{Var}\,(\boldsymbol{y})\boldsymbol{M_X} \\
&= \boldsymbol{P_X}\sigma^2\boldsymbol{I}\boldsymbol{M_X} \\
&= \sigma^2\boldsymbol{P_X}\boldsymbol{M_X} \\
&= \boldsymbol{0}
\end{aligned}$$

- With normal distribution, a null covariance is equivalent to independence. Since $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{\varepsilon}}$ are Normally distributed, they are independent.

4. $s^2/\sigma^2 \sim \chi^2_{n-K}/(n-K)$ and is independent of $\widehat{\boldsymbol{y}}$:
   - Distribution of $\frac{(n-K)s^2}{\sigma^2} \sim \chi^2_{n-K}$:
     - We have already shown:

$$s^2 = \frac{1}{n-K}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = \frac{1}{n-K}\boldsymbol{\varepsilon}'\boldsymbol{M_X}\boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim \boldsymbol{\mathcal{N}}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right) \Rightarrow \frac{\boldsymbol{\varepsilon}}{\sigma} \sim \boldsymbol{\mathcal{N}}\left(\boldsymbol{0}, \boldsymbol{I}\right)$$

     - Since $\boldsymbol{M_X}$ is an orthogonal projector of rank $n-K$, we have:

$$\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)'\boldsymbol{M_X}\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi^2_{n-K}$$

     - Therefore:

$$\frac{(n-K)s^2}{\sigma^2} \sim \chi^2_{n-K}$$

   - Independence:
     - s^2 is a function on $\widehat{\boldsymbol{\varepsilon}}$ only.
     - Since $\widehat{\boldsymbol{\varepsilon}}$ and $\widehat{\boldsymbol{y}}$ are independent given $\boldsymbol{X}$, so are s^2 and $\widehat{\boldsymbol{y}}$ given $\boldsymbol{X}$. But the distribution of s^2 does not depend on $\boldsymbol{X}$ at all, so that independence also holds unconditionally.

□

Appendix B

Further and Deeper Details

B.1 The Geometry of Least Squares

The first order condition of the least squares problem writes the following way:

$$\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \mathbf{0} \Leftrightarrow \boldsymbol{X}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{0} \Leftrightarrow \boldsymbol{X}'\widehat{\boldsymbol{\varepsilon}} = \mathbf{0}$$

Least Squares residuals are orthogonal to each column of $\boldsymbol{X}$.
The LS problem can be decomposed in two steps:

1. **Step 1: optimization**: find the vector $\widehat{\boldsymbol{\mu}} \in \mathrm{L}(\boldsymbol{X})$ that is the closest to the vector $\boldsymbol{y}$ (minimization of the Euclidian distance (squared norm)).

   $$\widehat{\boldsymbol{\mu}} = \arg \min_{\boldsymbol{\mu} \in \mathrm{L}(\boldsymbol{X})} \parallel \boldsymbol{y} - \boldsymbol{\mu} \parallel^2$$

   → $\widehat{\boldsymbol{\mu}}$ is the orthogonal projection of $\boldsymbol{y}$ on $\mathrm{L}(\boldsymbol{X})$.
2. **Step 2: solve a linear equation**: find the scalars (β's) that have to multiply each vector in $\boldsymbol{X}$ to give $\widehat{\boldsymbol{\mu}}$.

   $$\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\mu}}.$$

Theorem B.1.1. $\widehat{\boldsymbol{\mu}}$ *exists and is unique.*
$\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu} \in \mathrm{L}(\boldsymbol{X})} \parallel \boldsymbol{y} - \boldsymbol{\mu} \parallel^2$ *iff* $\boldsymbol{y} - \widehat{\boldsymbol{\mu}} \in \mathrm{L}^{\perp}(X)$.

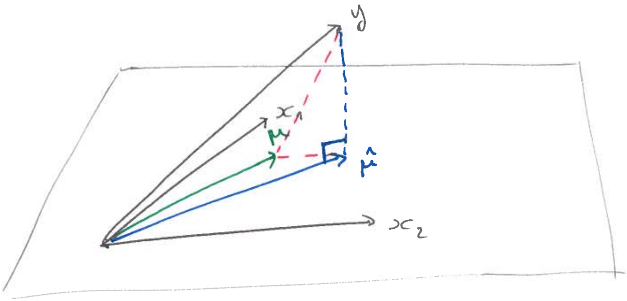
Proof. The proof is shown in two steps:

1. **Sufficiency**: $\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu} \in \mathrm{L}(\boldsymbol{X})} \parallel \boldsymbol{y} - \boldsymbol{\mu} \parallel^2 \Leftarrow \boldsymbol{y} - \widehat{\boldsymbol{\mu}} \in \mathrm{L}^{\perp}(\boldsymbol{X})$
   - $\boldsymbol{y} - \widehat{\boldsymbol{\mu}} \in \mathrm{L}^{\perp}(\boldsymbol{X})$
   - Let's take another vector $\boldsymbol{\mu} \in \mathrm{L}(\boldsymbol{X})$
   - Therefore, $\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}} \in \mathrm{L}(\boldsymbol{X})$, so $\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}} \perp \boldsymbol{y} - \widehat{\boldsymbol{\mu}}$
   - By Pythagoras' theorem, we have

     $$\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 + \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2$$

   - So: $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 \geq \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 \Rightarrow$ any vector $\boldsymbol{\mu} \neq \widehat{\boldsymbol{\mu}}$ is such that $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 \geq \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2$, with equality iff $\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 = 0$, that is $\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}}$.

Figure B.1: Theorem B.1.1: Sufficient Condition



① $\hat{\mu}$ is such that $y-\hat{\mu} \in L^{\perp}(X)$

② take $\mu \in L(X)$: μ is on the same subspace as $\hat{\mu} \Rightarrow \mu-\hat{\mu} \perp y-\hat{\mu}$

③ By Pythagoras' theorem :

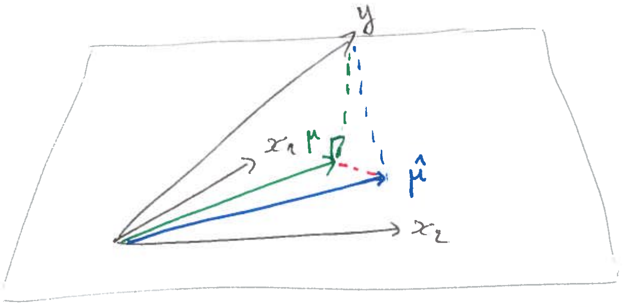
$\|y-\mu\|^2 = \|y-\hat{\mu}\|^2 + \|\mu-\hat{\mu}\|^2$

2. **Necessity**: $\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}\in \mathrm{L}(\boldsymbol{X})} \parallel \boldsymbol{y}-\boldsymbol{\mu} \parallel^2 \Rightarrow \boldsymbol{y}-\widehat{\boldsymbol{\mu}} \in \mathrm{L}^{\perp}(\boldsymbol{X})$
   - $\widehat{\boldsymbol{\mu}}$ is such that $\widehat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}\in \mathrm{L}(\boldsymbol{X})} \parallel \boldsymbol{y}-\boldsymbol{\mu} \parallel^2$ but let's assume that $(\boldsymbol{y}-\widehat{\boldsymbol{\mu}})$ is not orthogonal to $\mathrm{L}(\boldsymbol{X})$
   - Let's take a $\boldsymbol{\mu} \in \mathrm{L}(\boldsymbol{X})$, different from $\widehat{\boldsymbol{\mu}}$, such that $(\boldsymbol{y}-\boldsymbol{\mu})$ is orthogonal to $\mathrm{L}(\boldsymbol{X})$ (i.e. $\in \mathrm{L}^{\perp}(\boldsymbol{X})$) (such a vector always exists)
   - Therefore, $\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}} \in \mathrm{L}(\boldsymbol{X})$, so $\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}} \perp \boldsymbol{y}-\boldsymbol{\mu}$
   - By Pythagoras' theorem, we have

$$\|\boldsymbol{y}-\widehat{\boldsymbol{\mu}}\|^2 = \|\boldsymbol{y}-\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}\|^2$$

   - So $\|\boldsymbol{y}-\widehat{\boldsymbol{\mu}}\|^2 > \|\boldsymbol{y}-\boldsymbol{\mu}\|^2$, which contradicts the fact that $\widehat{\boldsymbol{\mu}}$ minimizes $\|\boldsymbol{y}-\boldsymbol{\mu}\|^2$
   - Conclusion: $(\boldsymbol{y}-\widehat{\boldsymbol{\mu}}) \in \mathrm{L}^{\perp}(\boldsymbol{X})$

Figure B.2: Theorem B.1.1: Necessary Condition



① $\hat{\mu}$ minimizes $\|y-\mu\|^2$ but $\|y-\hat{\mu}\| \not\perp L(X)$

② Take $\mu \in L(X)$ such that $(y-\mu) \in L^{\perp}(X)$

μ is in the same subspace as $\hat{\mu} \Rightarrow \mu-\hat{\mu} \perp y-\mu$

③ By Pythagoras' theorem :

$\|y-\hat{\mu}\|^2 = \|y-\mu\|^2 + \|\mu-\hat{\mu}\|^2$

□

Some consequences:

- The vector of residuals, $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ is orthogonal to $\mathrm{L}(\boldsymbol{X}) \Leftrightarrow \widehat{\boldsymbol{\varepsilon}}$ is in $\mathrm{L}^{\perp}(\boldsymbol{X})$. $\Leftrightarrow \widehat{\boldsymbol{\varepsilon}}$ is orthogonal to any vector in $\mathrm{L}(\boldsymbol{X})$.
- From Pythagora, for any $\boldsymbol{\mu} \in \mathrm{L}(\boldsymbol{X})$
$$\|y - \boldsymbol{\mu}\|^2 = \|y - \widehat{\boldsymbol{\mu}}\|^2 + \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \,.$$
True even for $\boldsymbol{\mu} = \boldsymbol{X\beta}$!
- The theorem implies in particular that residuals are orthogonal to all variables' columns.
- As the constant term is always included in the regressors (why?), this implies that the average (or sum) of residuals is zero.
Let $\boldsymbol{\iota}$ be the vector of ones (constant). Then $\boldsymbol{\iota}'\widehat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$.

Exercise B.1. *Show that the latter result implies that the average of fitted values is equal to $\bar{y}$.*

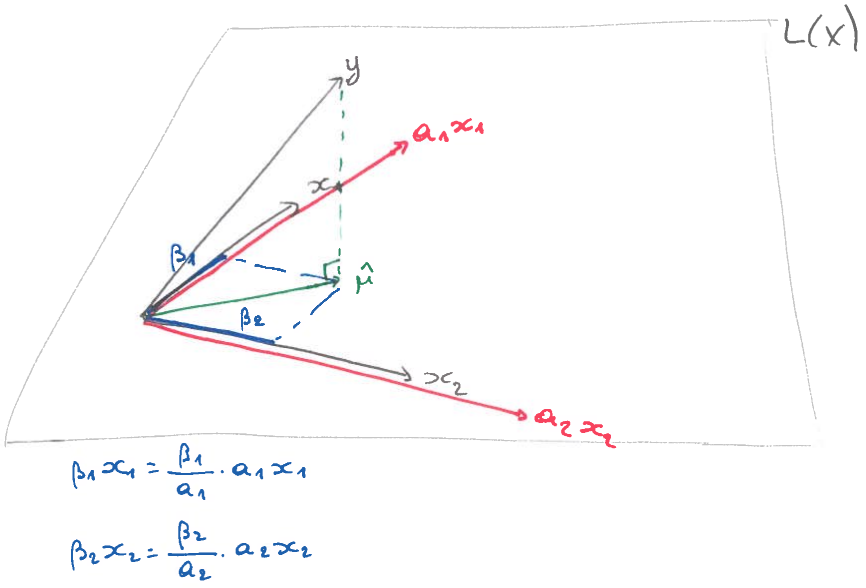
B.1.1 Orthogonal Projection

Definition 3. An **orthogonal projection** of any vector $\boldsymbol{y} \in \mathbb{R}^n$ on a subspace E is the vector $\boldsymbol{z} \in E$ which is the closest to $\boldsymbol{y}$:
$$\inf_{\boldsymbol{z} \in E} \|\boldsymbol{y} - \boldsymbol{z}\|^2$$

Properties and Consequences

- Uniqueness: see theorem
- Invariance of $\boldsymbol{P_X}$ to change of units:
If a square matrix $\boldsymbol{A}$ is non singular, then $\boldsymbol{P_{XA}} = \boldsymbol{P_X}$. So fitted values and residuals are the same, but OLS $\widehat{\boldsymbol{\beta}}$ changes.
  - Fitted values:
  $\boldsymbol{XA}$ is a linear combination of $\boldsymbol{X}$ so $\boldsymbol{XA}$ spans the same space than $\boldsymbol{X}$. So the projection (the fitted values) $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{y}}$ are the same, so $\boldsymbol{P_{XA}y} = \boldsymbol{P_X y} \Rightarrow \boldsymbol{P_{XA}} = \boldsymbol{P_X}$.
  - Residuals:
  $$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{P_{XA}y} = \boldsymbol{y} - \boldsymbol{P_X y} \,.$$
  - Parameter estimates:
  $$\widehat{\boldsymbol{y}}_{\boldsymbol{X}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{XAA}^{-1})\widehat{\boldsymbol{\beta}}$$
  $$\widehat{\boldsymbol{y}}_{\boldsymbol{XA}} = (\boldsymbol{XA})\widehat{\gamma}$$
  Since $\widehat{\boldsymbol{y}}_{\boldsymbol{X}} = \widehat{\boldsymbol{y}}_{\boldsymbol{XA}}$, we have $\widehat{\gamma} = (\boldsymbol{A}^{-1})\widehat{\boldsymbol{\beta}}$.

Figure B.3: Change in Measurement Units



$$\beta_1 x_1 = \frac{\beta_1}{a_1} \cdot a_1 x_1$$

$$\beta_2 x_2 = \frac{\beta_2}{a_2} \cdot a_2 x_2$$

- $\boldsymbol{P_X}$ idempotent (and symmetric), so eigenvalues are either 0 or 1. There are K 1's and $n - K$ 0's. Hence $\boldsymbol{P_X}$ is positive semi-definite.
- Conversely, every symmetric idempotent matrix is an orthogonal projector

Exercise B.2. *When there is only a constant in the model, show that the fitted values are $\bar{y}$ (how does $\boldsymbol{P_\iota}$ act on $\boldsymbol{y}$?). What are the residuals $\boldsymbol{M_\iota} y$?*

B.1.2 Collinearity

Case where Rank$(\boldsymbol{X}) < K$

- Columns vectors of $\boldsymbol{X}$ are linearly dependent: one of the variable is a linear function of others
- We can't compute $\widehat{\boldsymbol{\beta}}$: multiple solutions
- But we can determine $\widehat{\boldsymbol{\mu}}$, which is unique
- There is (at least) one matrix X^* of dimension $n \times K^*$, $K^* < K$, of full rank, such that $\boldsymbol{P_X} = \boldsymbol{P_{X^*}} = \boldsymbol{X}^* \left(\boldsymbol{X}^{*\prime} \boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^{*\prime}$
- In practice, you should likely get rid of one or more variables

In what follows we always assume a full rank design matrix $\boldsymbol{X}$.

B.1.3 Partitioned Fit

Consider two groups of regressors

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{\varepsilon}$$

- Goal: understand the decomposition of the fitted value of $\boldsymbol{y}$ into the 2 components $\boldsymbol{X_1}\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ and $\boldsymbol{X_2}\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$.
  - Since $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ vary at the same time, this is a priori difficult to determine the impact of $\boldsymbol{X_1}$ on $\boldsymbol{y}$ holding $\boldsymbol{X_2}$ fixed.
  - $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ is the OLS coefficient of the regression of $\boldsymbol{y}$ on $\boldsymbol{X_1}$ alone after the component of $\boldsymbol{X_1}$ that is collinear with $\boldsymbol{X_2}$ has been removed.

  - Refers to the interpretation of the estimated parameter associated to a particular variable: $\widehat{\beta}_k$ measures the estimated impact of x_k on $\boldsymbol{y}$ that cannot be explained by the other x variables.
- Do we get the same $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ whether we omit $\boldsymbol{X}_{\mathbf{2}}$?
- What is the formula for $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$?

Assume a full rank design matrix $\boldsymbol{X}$ and let

- $\widehat{\boldsymbol{\mu}} = \boldsymbol{P_X y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X_1}\widehat{\boldsymbol{\beta}}_{\mathbf{1}} + \boldsymbol{X_2}\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$ the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{X}$,
- $\widehat{\boldsymbol{\mu}}_{\mathbf{1}} = \boldsymbol{P_{X_1} y} = \boldsymbol{X_1}\tilde{\boldsymbol{\beta}}_{\mathbf{1}}$ the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{X}_1$ only,
- $\widehat{\boldsymbol{\mu}}_{\mathbf{2}} = \boldsymbol{P_{X_2} y} = \boldsymbol{X_2}\tilde{\boldsymbol{\beta}}_{\mathbf{2}}$ the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{X}_2$ only.

Then

- If we project orthogonally $\widehat{\boldsymbol{\mu}}$ on $\boldsymbol{X_1}$, we get $\widehat{\boldsymbol{\mu}}_{\mathbf{1}}$
  because $\boldsymbol{P_{X_1}P_{[X_1\,X_2]}} = \boldsymbol{P_{X_1}} = \boldsymbol{P_{[X_1\,X_2]}P_{X_1}}$.
- If $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are orthogonal, then $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_{\mathbf{1}} + \widehat{\boldsymbol{\mu}}_{\mathbf{2}}$ (with $\widehat{\boldsymbol{\mu}}_{\mathbf{1}}$ and $\widehat{\boldsymbol{\mu}}_{\mathbf{2}}$ orthogonal).
  That is, $\boldsymbol{P_{[X_1X_2]}} = \boldsymbol{P_{X_1}} + \boldsymbol{P_{X_2}}$. **Not true if no orthogonality**

Theorem B.1.2 (Frisch-Waugh). $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ *and residuals are the same for*

$$\boldsymbol{y} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \textbf{error}$$

and

$$\boldsymbol{M_{X_2}y} = \boldsymbol{M_{X_2}X_1\beta_1} + \textbf{error}\,.$$

Proof. The model is:

$$\begin{aligned} & \boldsymbol{y} = \boldsymbol{X_1\beta_1} + \boldsymbol{X_2\beta_2} + \boldsymbol{u} \\ \Rightarrow \quad & \boldsymbol{y} = \boldsymbol{X_1}\widehat{\boldsymbol{\beta}}_{\mathbf{1}} + \boldsymbol{X_2}\widehat{\boldsymbol{\beta}}_{\mathbf{2}} + \widehat{\boldsymbol{u}} \end{aligned}$$

Residuals are orthogonal to all regressors:

$$\begin{aligned} & \widehat{\boldsymbol{u}} \perp \begin{pmatrix} \boldsymbol{X_1} & \boldsymbol{X_2} \end{pmatrix} \\ \Rightarrow \quad & \boldsymbol{M_{X_1}}\widehat{\boldsymbol{u}} = \boldsymbol{M_{X_2}}\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{u}} \end{aligned}$$

So, by multiplying all elements by $\boldsymbol{M_{X_2}}$, the model becomes:

$$\boldsymbol{M_{X_2}y} = \boldsymbol{M_{X_2}X_1}\widehat{\boldsymbol{\beta}}_{\mathbf{1}} + \widehat{\boldsymbol{u}}$$

since $\boldsymbol{M_{X_2}X_2} = \mathbf{0}$.
We have now to check that $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ is the OLS estimator of $\boldsymbol{M_{X_2}y}$ on $\boldsymbol{M_{X_2}X_1}$. Therefore, we check that the orthogonality condition is satisfied:

$$(\boldsymbol{M_{X_2}X_1})'\,\widehat{\boldsymbol{u}} = \boldsymbol{X_1'M_{X_2}}\widehat{\boldsymbol{u}} = \boldsymbol{X_1'}\widehat{\boldsymbol{u}} = \mathbf{0}$$

so, the orthogonality condition is satisfied. Therefore, $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ is the OLS estimator of $\boldsymbol{M_{X_2}y}$ on $\boldsymbol{M_{X_2}X_1}$, which implies that $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ and residuals are identical.

□

Exercise B.3. *Show that* $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ *is the same when running*

$$\boldsymbol{y} = \boldsymbol{M_{X_2}X_1\beta_1} + \textbf{error}\,.$$

Particular Case: Deviation to the Mean

$$\boldsymbol{y} = \beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{x} + \boldsymbol{\varepsilon}$$

We know $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ with $\boldsymbol{X} = [\boldsymbol{\iota} \quad \boldsymbol{x}]$. We can write

$$\boldsymbol{y} = (\beta_0 + \beta_1\bar{x})\boldsymbol{\iota} + \beta_1(\boldsymbol{x} - \bar{x}\boldsymbol{\iota}) + \boldsymbol{\varepsilon} = \alpha_0 + \alpha_1(\boldsymbol{x} - \bar{x}\boldsymbol{\iota}) + \boldsymbol{\varepsilon}$$

We should have $\widehat{\alpha}_1 = \widehat{\beta}_1$ and $\widehat{\alpha}_0 = \bar{y}$.
So we obtain the same $\widehat{\beta}_1$ for

$$\boldsymbol{y} - \bar{y}\boldsymbol{\iota} = \beta_1(\boldsymbol{x} - \bar{x}\boldsymbol{\iota}) + \boldsymbol{\varepsilon}$$

Hence estimating

$$\boldsymbol{y} = \beta_0\boldsymbol{\iota} + \beta_1\boldsymbol{x} + \textbf{error}$$

and

$$\boldsymbol{M_\iota y} = \beta_1\boldsymbol{M_\iota x} + \textbf{error}\,,$$

we get the same value for $\widehat{\beta}_1$ and the same residuals.

Some consequences of Frisch-Waugh

- If we are only interested in $\widehat{\boldsymbol{\beta}}_\mathbf{1}$, we can get it running three “simpler” regressions
  - Regress $\boldsymbol{y}$ on $\boldsymbol{X_2}$ and save residuals ($\boldsymbol{M_{X_2}y}$),
  - Regress $\boldsymbol{X_1}$ on $\boldsymbol{X_2}$ and save residuals ($\boldsymbol{M_{X_2}X_1}$),
  - Run the *Double residual regression*: regress the first set of residuals ($\boldsymbol{M_{X_2}y}$) on the second set of residuals ($\boldsymbol{M_{X_2}X_1}$).
- FW can be used to plot in 2D a multiple regression line.
- The formula for $\widehat{\boldsymbol{\beta}}_\mathbf{1}$ is $(\boldsymbol{X_1'M_{X_2}X_1})^{-1}\boldsymbol{X_1'M_{X_2}y}$.
- If $\boldsymbol{X_1'X_2} = \mathbf{0}$, then $\widehat{\boldsymbol{\beta}}_\mathbf{1}$ is the same for the complete model and

$$\boldsymbol{y} = \boldsymbol{X_1\beta_1} + \textbf{error}\,.$$

  But residuals are not!
- Application: linear panel data model

$$y_{it} = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + ... + \beta_K x_{Kit} + \varepsilon_{it} \qquad i = 1,\dots n, \quad t = 1,\dots,T$$

$$\underset{(nT\times 1)}{\boldsymbol{y}} = \underset{(nT\times n)}{\boldsymbol{D}}\ \underset{(n\times 1)}{\boldsymbol{\alpha}} + \underset{(nT\times K)}{\boldsymbol{X}}\ \underset{(K\times 1)}{\boldsymbol{\beta}} + \underset{(nT\times 1)}{\boldsymbol{\varepsilon}}$$

  where $\boldsymbol{D}$ is a matrix with individual dummies. Then $\widehat{\beta}$ can be obtained by the regression

$$\boldsymbol{M_D y} = \boldsymbol{M_D X\beta} + \textbf{error}\,.$$

  How do we interpret this regression?

Exercise B.4. *Consider*

$$y_{it} = \gamma_t + \beta_1 x_{1it} + \beta_2 x_{2it} + ... + \beta_K x_{Kit} + \varepsilon_t \qquad i = 1,\dots n, \qquad t = 1,\dots,T$$

Apply the double regression to obtain the formula of the LS estimator of $\boldsymbol{\beta} = (\beta_1,\dots\beta_K)'$. Interpret.

B.1.4 Goodness of Fit

We know that

$$\|\boldsymbol{y}\|^2 = \|\widehat{\boldsymbol{y}}\|^2 + \|\widehat{\boldsymbol{\varepsilon}}\|^2 = \|\boldsymbol{P_X y}\|^2 + \|\boldsymbol{M_X y}\|^2$$

Uncentered R^2

Define the uncentered R-square as

$$R_u^2 = \frac{\|\widehat{\boldsymbol{y}}\|^2}{\|\boldsymbol{y}\|^2} = 1 - \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{\|\boldsymbol{y}\|^2} = \cos^2\theta$$

where θ is the angle between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$, so that $R_u^2 \in (0,1)$.
R_u^2 is invariant to units. But R_u^2 is not invariant to origin scaling: add a constant a to each y_i so that $\boldsymbol{y}$ becomes $\boldsymbol{z} = \boldsymbol{y} + \boldsymbol{a\iota}$, then if $\boldsymbol{X}$ includes the constant

$$\widehat{\boldsymbol{z}} = \boldsymbol{P_X z} = \boldsymbol{P_X y} + \boldsymbol{a\iota} = \widehat{\boldsymbol{y}} + \boldsymbol{a\iota}$$

and the R_u^2 becomes

$$\frac{\|\widehat{\boldsymbol{y}} + \boldsymbol{a\iota}\|^2}{\|\boldsymbol{y} + \boldsymbol{a\iota}\|^2}$$

We can choose a as large as we want and obtain R_u^2 as close to one as we want.

Centered R^2

Assume the model includes the intercept. Then by FW Theorem, we can express all variables as deviations from their means without changing parameter estimates or residuals. Since

$$\boldsymbol{M_\iota y} = \boldsymbol{M_\iota}\widehat{\boldsymbol{y}} + \widehat{\boldsymbol{\varepsilon}}$$

we have

$$\|\boldsymbol{M_\iota y}\|^2 = \|\boldsymbol{M_\iota}\widehat{\boldsymbol{y}}\|^2 + \|\widehat{\boldsymbol{\varepsilon}}\|^2 .$$

This is the *decomposition of variance formula.* Put differently

$$\text{TSS} = \text{ESS} + \text{RSS}$$

The centered R-square writes:

$$R^2 = \frac{\|\boldsymbol{M_\iota}\widehat{\boldsymbol{y}}\|^2}{\|\boldsymbol{M_\iota y}\|^2} = 1 - \frac{\|\widehat{\boldsymbol{\varepsilon}}\|^2}{\|\boldsymbol{M_\iota y}\|^2} = \cos^2\alpha$$

where α is the angle between $\boldsymbol{M_\iota y}$ et $\boldsymbol{M_\iota}\widehat{\boldsymbol{y}}$, so that $R^2 \in (0,1)$ and is invariant to scaling. This is the goodness-of-fit measure that is used in practice.

B.2 Proof of the mean squared error of the prediction $\widehat{y}_{n+1}$

$$\begin{aligned}
\mathrm{E}\left[(y_{n+1}-\widehat{y}_{n+1})^2|\boldsymbol{X},\boldsymbol{x_{n+1}}\right] &= \mathrm{E}\left[\left(\boldsymbol{x'_{n+1}\beta}+\varepsilon_{n+1}-\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}\right)^2|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]\\
&= \mathrm{E}\left[\varepsilon^2_{n+1}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]+\mathrm{E}\left[\left(\boldsymbol{x'_{n+1}\beta}-\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}\right)^2|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]+\\
&\quad 2\mathrm{E}\left[\varepsilon_{n+1}\left(\boldsymbol{x'_{n+1}\beta}-\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}\right)|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]\\
&= \mathrm{E}\left[\left(\varepsilon^2_{n+1}\right)|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]+\mathrm{E}\left[\left(\boldsymbol{x'_{n+1}\beta}-\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}\right)^2|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]\\
&= \mathrm{E}\left[\left(\varepsilon^2_{n+1}\right)|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]+\mathrm{Var}\left[\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]\\
&\qquad \text{since } \mathrm{E}\left[\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]=\boldsymbol{x'_{n+1}\beta}\\
&= \sigma^2+\boldsymbol{x'_{n+1}}\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]\boldsymbol{x_{n+1}}\\
&= \sigma^2+\sigma^2\boldsymbol{x'_{n+1}}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{x_{n+1}}
\end{aligned}$$

Note: we have $\mathrm{E}\left[\varepsilon_{n+1}\left(\boldsymbol{x'_{n+1}\beta}-\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}\right)|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]=0$ because:

- $\mathrm{E}\left[\varepsilon_{n+1}\boldsymbol{x'_{n+1}\beta}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]=\mathrm{E}\left[\varepsilon_{n+1}\boldsymbol{x'_{n+1}}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]\boldsymbol{\beta}=0$ by assumption
- $\mathrm{E}\left[\varepsilon_{n+1}\boldsymbol{x'_{n+1}}\widehat{\boldsymbol{\beta}}|\boldsymbol{X},\boldsymbol{x_{n+1}}\right]=0$ because ε_{n+1} and $\widehat{\boldsymbol{\beta}}$ are uncorrelated:
  - $\widehat{\boldsymbol{\beta}}=\boldsymbol{(X'X)^{-1}X'y}=\boldsymbol{\beta}+\boldsymbol{(X'X)^{-1}X'\varepsilon}$
  - The correlation between $\widehat{\boldsymbol{\beta}}$ and ε_{n+1} is null if errors are not correlated, i.e. if $\boldsymbol{\varepsilon}$ is uncorrelated with ε_{n+1}

B.3 Proof of consistency of s^2, Theorem 1.2.7

Recall that

$$\begin{aligned}
s^2 &= \frac{1}{n-K}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}=\frac{1}{n-K}\boldsymbol{\varepsilon' M_X \varepsilon}\\
&= \frac{1}{n-K}\boldsymbol{\varepsilon'}\left(\mathbf{I}-\boldsymbol{X(X'X)^{-1}X'}\right)\boldsymbol{\varepsilon}=\frac{1}{n-K}\boldsymbol{\varepsilon'\varepsilon}+\frac{1}{n-K}\boldsymbol{\varepsilon' X(X'X)^{-1}X'\varepsilon}\,.
\end{aligned}$$

- $\frac{1}{n}\boldsymbol{\varepsilon'\varepsilon}=\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2$. Therefore by the WLLN, $\frac{1}{n}\boldsymbol{\varepsilon'\varepsilon}=\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\xrightarrow{p}\mathrm{E}\left(\varepsilon^2\right)=\sigma^2$. Since $\frac{n}{n-K}\to 1$, we have by the continuous mapping theorem

$$\frac{1}{n-K}\boldsymbol{\varepsilon'\varepsilon}=\frac{n}{n-K}\frac{1}{n}\boldsymbol{\varepsilon'\varepsilon}\xrightarrow{p}\sigma^2\,.$$

- Write

$$\frac{1}{n-K}\boldsymbol{\varepsilon' X(X'X)^{-1}X'\varepsilon}=\frac{n}{n-K}\frac{\boldsymbol{\varepsilon' X}}{\boldsymbol{n}}\left(\frac{\boldsymbol{X'X}}{\boldsymbol{n}}\right)^{\boldsymbol{-1}}\frac{\boldsymbol{X'\varepsilon}}{\boldsymbol{n}}\,.$$

  Now by the WLLN $\frac{1}{n}\boldsymbol{X'\varepsilon}\xrightarrow{p}\boldsymbol{0}$. Moreover, $\frac{1}{n}\boldsymbol{X'X}\xrightarrow{p}\boldsymbol{Q}$. Hence, by the continuous mapping theorem,

$$\frac{1}{n-K}\boldsymbol{\varepsilon' X(X'X)^{-1}X'\varepsilon}\xrightarrow{p}0\,.$$

Remark: similarly one can check that $\frac{1}{n}\sum_{i=1}^n\widehat{\varepsilon}_i^2$ is consistent, even if biased (but asymptotically unbiased, i.e. its bias tends to zero).