# Regression analysis with cross-sectional data : Specification, estimation and inference

Bernard Lejeune

Supplemental lecture notes I

Advanced Econometrics
HEC-University of Liège
Academic year 2021-2022

- These lecture notes provide a synthesis of the basics of the linear regression model for cross-sectional data in matrix form. They recap the results obtained in Lejeune (2011), but here explicitly allowing for $X$ to be random (rather than assuming $X$ nonstochastic). They cover the main results of Chapter 1-5 of Wooldridge (2016).

# 1. Model specification

- The easiest way of apprehending the multiple linear regression model is to see it as a generic statistical model for evaluating :
  - how the mean value taken by some variable $y$ varies as a function of some other variables $(x_2, ..., x_k)$ in a given population,
  - based on a random sample of observations from the population.

- The mean value taken by some variable $y$ as a function of some other variables $(x_2, ..., x_k)$ is supposed to be the empirical counterpart of the theoretical relationship of interest.

- In terms of probability theory, the mean value taken by some variable $y$ as a function of some other variables $(x_2, ..., x_k)$ in a population is represented by the conditional expectation, also called the conditional mean :

$$E(y|x_2, ..., x_k) = g(x_2, ..., x_k)$$

where $(x_2, ..., x_k)$ and $y$ are random values obtained for an individual drawn at random from the population.

- Basically, the multiple linear regression model assumes that the conditional mean $E(y|x_2, ..., x_k)$ is linear (in parameters), i.e. that $g(x_2, ..., x_k) = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$.

- The classical linear regression model relies on further assumptions, such as constant conditional variance or conditional normality, but these further assumptions are not as essential as the conditional mean assumption, and may be relaxed if needed.

- The most important feature of the multiple linear regression model is that it allows to evaluate the causal effect – i.e., the effect, the other factors being held constant – of some explanatory variable of interest $x$ on another variable $y$ based on non-experimental data.

## 1.1. Model assumptions for cross-sectional data

- Following Wooldridge (2016), Chapter 3-4, the seminal statistical assumptions underlying the multiple regression model for cross-sectional data may be expressed as follows :

  – MLR.1 Linearity in parameters

  The population model, describing the relationship between the dependent variable $y$ and the explanatory variables $(x_2, ..., x_k)$ for a generic draw from the population, can be written as :

  $$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

  where $(\beta_1, ..., \beta_k)$ are unknown parameters and $u$ is an error term.

  – MLR.2 Random sampling

  The available data are realizations of a random sample of size $n$, $\{(x_{i2}, ..., x_{ik}, y_i) : i = 1, ..., n\}$, following the population model in assumption MLR.1

  – MLR.3 No perfect collinearity

  In the sample (and thus in the population), none of the explanatory variables $(x_2, ..., x_k)$ is constant, and there is no exact linear relationship among them.

  – MLR.4 Zero conditional mean

  The expected value of $u$ given any values of $(x_2, ..., x_k)$ is equal to zero, which is equivalent to say that the expected value of $y$ given any values of $(x_2, ..., x_k)$ is equal to $\beta_1 + \beta_2 x_2 + ... + \beta_k x_k$ :

  $$E(u|x_2, ..., x_k) = 0 \Leftrightarrow E(y|x_2, ..., x_k) = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$$

  – MLR.5 Homoskedasticity

  The variance of $u$ given any values of $(x_2, ..., x_k)$ is constant, which is

equivalent to say that the variance of $y$ given any values of $(x_2, ..., x_k)$ is constant :

$$Var(u|x_2, ..., x_k) = \sigma^2 \Leftrightarrow Var(y|x_2, ..., x_k) = \sigma^2$$

where $\sigma^2$ is a unknown parameter.

– MLR.6 Normality

The distribution of $u$ is the same given any values of $(x_2, ..., x_k)$ – i.e., $u$ is independent of $(x_2, ..., x_k)$ – and is normal with zero mean and variance $\sigma^2$, which is equivalent to say that the distribution of $y$ given any value of $(x_2, ..., x_k)$ is normal with mean equal to $\beta_1 + \beta_2 x_2 + ... + \beta_k x_k$ and variance $\sigma^2$ :

$$u|x_2, ..., x_k \sim N(0, \sigma^2) \Leftrightarrow y|x_2, ..., x_k \sim N(\beta_1 + \beta_2 x_2 + ... + \beta_k x_k, \sigma^2)$$

• The above set of assumptions deserves some comments :

– The linearity assumption MLR.1 requires the population model to be linear in parameters, not in variables. In practice, the original variables may be transformed, yielding a model which is nonlinear in variables, but still linear in parameters. The most widely used transformation is the natural logarithm, noted log(.).

– The random sampling assumption MLR.2 implies that the individual observations $(x_{i2}, ..., x_{ik}, y_i)$ are both identically and independently distributed across $i$.

– To be able to estimate the partial effect each of variable $x_2, ..., x_k$, we need each of these variables to vary, and to vary at least to some extent independently of the other variables. This is what is required by assumption MLR.3.

– The zero conditional mean assumption MLR.4 basically says that the systematic part of the population model in MLR.1 is the conditional mean of $y$ given $(x_2, ..., x_k)$. In other words, assumptions MLR.1 and MLR4 indicate that we are interested in estimating the conditional mean $E(y|x_2, ..., x_k)$, and that this conditional mean is assumed linear in parameters.

– By the law of iterated expectations[1], the zero conditional mean assumption MLR.4 implies that the unconditional mean of $u$ is zero (i.e., $E(u) = 0$), and that $u$ is uncorrelated (have zero covariance) with each explanatory variable (i.e., $E(x_j u) = 0$, $j = 2, ..., k$)).

– The homoskedasticity and normality assumptions MLR.5 and MLR.6 are auxiliary assumptions. They are less essential, and may be relaxed if needed (although at the price of additional complications).

[1] On the law of iterated expectations, and more generally on the properties of conditional expectation, see Wooldridge (2016), Appendix B4-f. A brief summary of these properties, and further of those of conditional variance, is provided in an appendix at the end of the present lecture notes.

## 1.2. Model and assumptions in matrix form

- The multiple linear regression model is more easily handled, and its properties studied, in matrix form[2].

- For any observation $i$ drawn from the population, we can write:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

$$\Leftrightarrow \quad y_i = X_i \beta + u_i$$

where $X_i = (1, x_{i2}, ..., x_{ik})$ is a $1 \times k$ (row) vector of explanatory variables (including a constant) and $\beta = (\beta_1, \beta_2, ..., \beta_k)$ is $k \times 1$ (column) vector. Further, stacking all observations of a random sample of size $n$, we can write:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\Leftrightarrow \quad Y = X\beta + u$$

where $Y$ and $u$ are $n \times 1$ vectors, and $X$ is a $n \times k$ matrix, whose the $i$-th row is equal to $X_i$.

- Following Wooldridge (2016), Appendix E-2, assumptions MLR.1 – MLR.6 can be rewritten in matrix form as follows[3]:

  – E.1 Linearity in parameters: $Y = X\beta + u$

  – E.2 No perfect collinearity: $\text{rank}(X) = k$

  – E.3 Zero conditional mean: $E(u|X) = 0 \Leftrightarrow E(Y|X) = X\beta$

  – E.4 Homoskedasticity & no correlation: $V(u|X) = \sigma^2 I \Leftrightarrow V(Y|X) = \sigma^2 I$

  – E.5 Normality: $u|X \sim N(0, \sigma^2 I) \Leftrightarrow Y|X \sim N(X\beta, \sigma^2 I)$

- Assumptions E.1 – E.5 are usually referred to as the 'classical linear model assumptions'. These assumptions are actually weaker than the seminal assumptions MLR.1 – MLR.6. They hold whenever MLR.1 – MLR.6 hold:

  – Assumption E.1 must be read: $y_i = X_i\beta + u_i$, $i = 1, ..., n$. It is exactly the same as assumption MLR.1.

  – Assumption E.2 means that $X$ is full rank, so that there is no exact exact linear relationship among its columns. It is exactly the same as assumption MLR.3.

[2] For a summary of matrix algebra, including moments and distributions of random vectors, see Wooldridge (2106), Appendix D.

[3] A brief summary of the properties of conditional mean and conditional variance of random vectors may be found in the appendix at the end of the present lecture notes.

– Assumption E.3 must be read: $E(u_i|X) = 0 \Leftrightarrow E(y_i|X) = X_i\beta$, $i = 1, ..., n$. It holds under assumptions MLR.2 and MLR.4. The random sampling assumption MLR.2 implies independence of the observations across $i$, so that $E(u_i|X) = E(u_i|X_i) \Leftrightarrow E(y_i|X) = E(y_i|X_i)$. On the other hand, the zero conditional mean MLR.4 states that $E(u_i|X_i) = 0 \Leftrightarrow E(y_i|X_i) = X_i\beta$.

– Assumption E.4 must be read: $Var(u_i|X) = \sigma^2 = Var(y_i|X)$, $i = 1, ..., n$, and $Cov(u_i, u_j|X) = 0 = Cov(y_i, y_j|X)$, for all $i \neq j$. It holds under assumptions MLR.2 and MLR.5. The no correlation (zero covariance) part of E.4 directly follows from the independence of the observations across $i$ implied by the random sampling assumption MLR.2. The variance part of E.4 follows from the random sampling assumption MLR.2, which implies that $Var(u_i|X) = Var(u_i|X_i) \Leftrightarrow Var(y_i|X) = Var(y_i|X_i)$, and from the homoskedasticity assumption MLR.5, which states that $Var(u_i|X_i) = \sigma^2 = Var(y_i|X_i)$.

– Assumption E.5 states that the joint distribution of $u$ given $X$ is (multivariate) normal with zero mean and variance-covariance matrix $\sigma^2 I$, which is equivalent to say that joint distribution of $Y$ given $X$ is (multivariate) normal with mean $X\beta$ and variance-covariance matrix $\sigma^2 I$ . It directly follows from the independence of the observations across $i$ implied by the random sampling assumption MLR.2 and the normality assumption MLR.6.

# 2. Model estimation

## 2.1. The ordinary least squares estimator

• The ordinary least squares (OLS) estimator of the unknown vector of parameters $\beta$ of the classical linear regression model is defined as:

$$\begin{aligned} \hat{\beta} &= \text{Argmin}_\beta \sum_{i=1}^{n} (y_i - X_i\beta)^2 \\ &= \text{Argmin}_\beta (Y - X\beta)'(Y - X\beta) \end{aligned}$$

• The first order condition of the above minimization problem is[4]:

$$\sum_{i=1}^{n} X_i'(y_i - X_i\hat{\beta}) = \sum_{i=1}^{n} X_i'\hat{u}_i = 0$$

$$\Leftrightarrow X'(Y - X\hat{\beta}) = X'\hat{u} = 0 \tag{1}$$

and the OLS estimator is given by:

[4] For a detailed derivation, see Wooldridge (2016) p. 720-722, Hayashi (2010) p. 15-18, or Lejeune (2011) p. 27-29 and p. 103-104.

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1} \sum_{i=1}^{n} X_i' y_i \\ &= \left(X'X\right)^{-1} X'Y\end{aligned}$$

- Some remarks :
  - The no perfect collinearity assumption (in matrix form, assumption E.2 : $\text{rank}(X) = k$) ensures that $X'X$ is full rank and thus invertible, so that $\hat{\beta}$ is well defined.
  - The first order condition (1) implies that the OLS residuals sum to zero (when as usual an intercept is included in the model) and that the sample covariance between each of the explanatory variables and the OLS residuals is zero.
  - For the simple linear regression model $y = \beta_1 + \beta_2 x + u$, the OLS estimators of $\beta_1$ and $\beta_2$ are given by[5] :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \quad \text{and} \quad \hat{\beta}_2 = \frac{Cov_{spl}(x_i, y_i)}{Var_{spl}(x_i)} \tag{2}$$

    where $Cov_{spl}(x_i, y_i)$ denotes the sample covariance between $x$ and $y$, and $Var_{spl}(x_i)$ is the sample variance of $x$[6].

## 2.2. Goodness-of-fit

- Once the model is estimated, each observation may be decomposed into two parts, a fitted (explained) value and a residual :

$$y_i = X_i \hat{\beta} + \hat{u}_i = \hat{y}_i + \hat{u}_i$$

- Provided that the model contains an intercept (so that $\bar{y} = \overline{\hat{y}}$), it may be shown that[7] :

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n} \hat{u}_i^2}_{\text{SSR}},$$

  where SST denotes the total sum of squares, SSE the explained sum of squares, and SSR the residual sum of squares of the regression.

- A goodness-of-fit measure of the regression is given by the so-called $R$-squared

[5] See Wooldridge (2016) p. 26 or Lejeune (2011) p. 15-16.

[6] $Cov_{spl}(x_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ and $Var_{spl}(x_i) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

[7] See Wooldridge (2016) p. 33-36 and p. 70-71, Hayashi (2000) p. 20-21, or Lejeune (2011) p. 82-84 and p. 117.

of the regression, also called the coefficient of determination of the regression :

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$$

$R^2$ measures the proportion of the sample variation in $y_i$ that is explained by the regression. By definition, $R^2$ is a number between zero and one : $0 \leq R^2 \leq 1$.

- It can be shown that the $R$-squared is also equal to the squared sample correlation coefficient $\rho_{spl}(y_i, \hat{y}_i)$ between the actual $y_i$ and the fitted values $\hat{y}_i$ :

$$R^2 = \left(\rho_{spl}(y_i, \hat{y}_i)\right)^2 = \left(\frac{Cov_{spl}(y_i, \hat{y}_i)}{\sqrt{Var_{spl}(y_i)}\sqrt{Var_{spl}(\hat{y}_i)}}\right)^2$$

# 3. Sampling properties of the OLS estimator

## 3.1. Finite sample properties of OLS

- Finite sample properties refer to sampling distribution properties which are valid for any finite sample size $n$. For both convenience and generality, we study these properties under the weakest set of assumptions E.1 – E.5.

### 3.1.1. Unbiasedness of $\hat{\beta}$

- Under assumptions E.1 and E.2, we have :

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y = (X'X)^{-1} X'(X\beta + u) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'u \\ &= \beta + (X'X)^{-1} X'u \end{aligned} \tag{3}$$

so that, under the zero conditional mean assumption E.3 $E(u|X) = 0$, we have :

$$\begin{aligned} E(\hat{\beta}|X) &= E\left[(\beta + (X'X)^{-1} X'u)|X\right] \\ &= \beta + (X'X)^{-1} X'E(u|X) \\ &= \beta \end{aligned} \tag{4}$$

- We have just established the following property[8] :

Property 1 Unbiasedness of $\hat{\beta}$

Under assumptions E.1 – E.3, the OLS estimator $\hat{\beta}$ is an unbiased estimator of $\beta$ :

$$E(\hat{\beta}|X) = \beta$$

[8] This property is the same as Theorem E.1 in Wooldridge (2016), Appendix E-2.

- Note that:
  - as $E(\hat{\beta}|X)$ does not depend on $X$, by the law of iterated expectations, we also have:
$$E(\hat{\beta}) = E\left[E(\hat{\beta}|X)\right] = E(\beta) = \beta$$
i.e., the unbiasedness also holds unconditionally.
  - as assumptions MLR.1 – MLR.4 imply assumptions E.1 – E.3, Property 1 also holds under the seminal assumptions MLR.1 – MLR.4.

### 3.1.2. Omitted variable bias

- Property 1 ensures that whenever the population model is correctly specified for the conditional mean of $y$ given $(x_2, ..., x_k)$, then OLS provides an unbiased estimator of the unknown parameters $(\beta_1, \beta_2, ..., \beta_k)$ of $E(y|x_2, ..., x_k) = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k$, i.e., an unbiased estimator of the partial effect of each variable, the other variables being held constant.

- If any of 'the other variables being held constant' is not included in the estimated regression, then OLS will usually no longer provide an unbiased estimator of the partial effects of interest, i.e., the partial effect of each variable, the other variables (including the omitted ones) being held constant. This is the omitted variable bias.

- The main features of the omitted variable bias may be outlined by looking at a simple case. Suppose we are interested in the parameter $\beta_2$ – i.e., the partial effect of $x_2$, $x_3$ being held constant – in the population model:
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u, \quad \text{with } E(u|x_2, x_3) = 0 \tag{5}$$
$$\Leftrightarrow \; E(y|x_2, x_3) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$$
Under assumption E.1 – E.3, Property 1 ensures that a regression[9] of $y_i$ on $x_{i2}, x_{i3}$ yields an unbiased estimator of $\beta_2$. Suppose that we run instead a (simple) regression of $y_i$ on $x_{i2}$, $x_{i3}$ being omitted. The OLS estimator associated with that regression with a omitted variable can be written:
$$\hat{\beta}^*_{12} = \begin{bmatrix} \hat{\beta}^*_1 \\ \hat{\beta}^*_2 \end{bmatrix} = (X'_{12}X_{12})^{-1} X'_{12}Y\,, \quad \text{where } Y = \begin{bmatrix} \vdots \\ y_i \\ \vdots \end{bmatrix} \text{ and } X_{12} = \begin{bmatrix} \vdots & \vdots \\ 1 & x_{i2} \\ \vdots & \vdots \end{bmatrix}$$
Let the population model (5) be written as:
$$Y = X\beta + u = X_{12}\beta_{12} + X_3\beta_3 + u\,, \quad \text{where } \beta_{12} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \text{ and } X_3 = \begin{bmatrix} \vdots \\ x_{i3} \\ \vdots \end{bmatrix}$$

[9] Unless explicitely stated, a constant is always included in the regression.

Under assumption E.1 – E.2, we have :

$$\begin{aligned}\hat{\beta}^*_{12} &= (X'_{12}X_{12})^{-1} X'_{12}Y \\ &= (X'_{12}X_{12})^{-1} X'_{12}(X_{12}\beta_{12} + X_3\beta_3 + u) \\ &= \beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}u\,,\end{aligned}$$

so that, under the zero conditional mean assumption E.3 $E(u|X) = 0$, we have :

$$\begin{aligned}E(\hat{\beta}^*_{12}|X) &= E\left[(\beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}u)|X\right] \\ &= \beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3 + (X'_{12}X_{12})^{-1} X'_{12}E(u|X) \\ &= \beta_{12} + (X'_{12}X_{12})^{-1} X'_{12}X_3\beta_3\end{aligned}$$

which can be written as :

$$E(\hat{\beta}^*_{12}|X) = \beta_{12} + \beta_3\hat{\delta}$$

$$\Leftrightarrow \quad \begin{bmatrix} E(\hat{\beta}^*_1|X) \\ E(\hat{\beta}^*_2|X) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \beta_3 \begin{bmatrix} \hat{\delta}_1 \\ \hat{\delta}_2 \end{bmatrix}$$

where $\hat{\delta} = (X'_{12}X_{12})^{-1} X'_{12}X_3$ is the OLS estimator of the (simple) regression of the omitted variable $x_{i3}$ on $x_{i2}$, $\hat{\delta}_1$ being the estimated intercept and $\hat{\delta}_2$ the estimated slope.

From (2) in Section 2.1, we have $\hat{\delta}_2 = \frac{Cov_{spl}(x_{i2},x_{i3})}{Var_{spl}(x_{i2})}$, so that $E(\hat{\beta}^*_2|X)$ can finally be written :

$$E(\hat{\beta}^*_2|X) = \beta_2 + \beta_3\frac{Cov_{spl}(x_{i2}, x_{i3})}{Var_{spl}(x_{i2})}$$

In other words, the estimator $\hat{\beta}^*_2$ from the regression omitting $x_{i3}$ will be an biased estimator of $\beta_2$ – i.e. of the partial effect of $x_2$, $x_3$ being held constant, in the population model (5) – unless one of the following conditions is fulfilled :

- $\beta_3 = 0$ in the population model (5). In this case, $E(y|x_2, x_3) = \beta_1 + \beta_2 x_2$, i.e., $E(y_i|x_{i2}, x_{i3})$ is a linear function of $x_{i2}$, and does not depend on $x_{i3}$. We have just omitted an irrelevant variable.
- The omitted variable $x_3$ is uncorrelated with $x_2$ (i.e., $Cov_{spl}(x_{i2}, x_{i3}) = 0$).

If none of these conditions is fulfilled, then the estimator $\hat{\beta}^*_2$ from the regression omitting $x_{i3}$ will be biased, and the direction of the bias $Bias(\hat{\beta}^*_2|X) = E(\hat{\beta}^*_2|X) - \beta_2 = \beta_3\frac{Cov_{spl}(x_{i2},x_{i3})}{Var_{spl}(x_{i2})}$ will depend on the sign of $\beta_3$ and the sign of $Cov_{spl}(x_{i2}, x_{i3})$[10].

It is worth noting that if $\hat{\beta}^*_2$ is biased (e.g., because $Cov_{spl}(x_{i2}, x_{i3}) \neq 0$), it is biased as an estimator of $\beta_2$, which is the partial effect of $x_2$ on $y$, $x_3$ being held constant, in the population model (5). But as an estimator of the effect of $x_2$ on $y$, regardless of $x_3$ ($x_3$ not being held constant), it may actually be an unbiased

[10] For a discussion, see Wooldridge (2016) p. 78-81.

estimator. As a matter of fact, if the population model (5) hold, we have:

$$\begin{aligned} E(y|x_2) &= E\left[(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + u)|x_2\right] \\ &= \beta_1 + \beta_2 x_2 + \beta_3 E\left(x_3|x_2\right) + E(u|x_2) \end{aligned}$$

By the law of iterated expectations, $E(u|x_2, x_3) = 0$ implies $E(u|x_2) = 0$, so that we have[11]:

$$E(y|x_2) = \beta_1 + \beta_2 x_2 + \beta_3 E\left(x_3|x_2\right)$$

If the conditional mean of $x_3$ given $x_2$ is linear – i.e., if $E\left(x_3|x_2\right) = \delta_1 + \delta_2 x_2$ –, we further have:

$$\begin{aligned} E(y|x_2) &= \beta_1 + \beta_2 x_2 + \beta_3(\delta_1 + \delta_2 x_2) \\ &= (\beta_1 + \beta_3\delta_1) + (\beta_2 + \beta_3\delta_2)x_2 \\ &= \beta_1^* + \beta_2^* x_2 \end{aligned} \tag{6}$$

If the above population model (6) holds – and it will if $E\left(x_3|x_2\right) = \delta_1 + \delta_2 x_2$ –, then, from Property 1, the OLS estimator $\hat{\beta}_2^*$ from the (simple) regression of $y_i$ on $x_{i2}$ is actually an unbiased estimator of $\beta_2^* = \beta_2 + \beta_3\delta_2$. Simply, the parameter $\beta_2^*$, which gives the effect of $x_2$ on $y$, regardless of $x_3$ ($x_3$ not being held constant), is not what we were looking for (unless $\beta_3 = 0$ and/or $\delta_2 = 0$)[12]. This highlights the fact that when discussing whether an estimator is or is not biased (or likewise consistent), one must carefully states biased (or consistent) for which quantity, and under which assumptions.

- The above analysis of omitted variables bias may be extended to the case where one or more explanatory variables are omitted from a regression with any number $k$ of explanatory variables. In this general case, it may likewise be shown that the OLS estimator $\hat{\beta}_{(.)}^*$ from the regression where one or more variables have been omitted will be a biased estimator of the parameters of interest of the original population model – i.e., the partial effects of each included variable, the other variables (including the omitted ones) being held constant – unless one of the following conditions is fulfilled:

  – In the original population model, the parameter of each omitted variable is equal to zero (i.e., we have just omitted irrelevant variables).

  – Each omitted variable is uncorrelated with each of the other variables of the original population model.

  If none of these conditions is fulfilled, then the estimator $\hat{\beta}_{(.)}^*$ from the regression where one or more variables have been omitted will be biased, but deriving the direction of the bias is now more difficult.

[11] Note that the same result may likewise directly be obtained from the law of iterated expectations: $E(y|x_2) = E\left[E(y|x_2, x_3)|x_2\right] = E\left[(\beta_1 + \beta_2 x_2 + \beta_3 x_3)|x_2\right] = \beta_1 + \beta_2 x_2 + \beta_3 E\left(x_3|x_2\right)$.

[12] Incidentaly, note that if $\delta_2 = 0$, then $E(x_3|x_2) = E(x_3) = \delta_1$, which implies that $Cov(x_2, x_3) = 0$.

### 3.1.3. Variance-covariance matrix of $\hat{\beta}$

- Under assumptions E.1 – E.3, from (3) and (4), we have :

$$\hat{\beta} = \beta + (X'X)^{-1} X'u \quad \text{and} \quad E(\hat{\beta}|X) = \beta$$

so that :

$$\begin{aligned} V(\hat{\beta}|X) &= E\left[(\hat{\beta} - E(\hat{\beta}|X))(\hat{\beta} - E(\hat{\beta}|X))'|X\right] \\ &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X\right] \\ &= E\left[(X'X)^{-1} X'uu'X (X'X)^{-1} |X\right] \\ &= (X'X)^{-1} X'E(uu'|X)X (X'X)^{-1} \end{aligned}$$

Using the homoskedasticity and no correlation assumption E.4 $V(u|X) = E(uu'|X) = \sigma^2 I$, we thus have :

$$\begin{aligned} V(\hat{\beta}|X) &= (X'X)^{-1} X'E(uu'|X)X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

- We have just established the following property[13] :

Property 2 Variance-covariance matrix of $\hat{\beta}$
Under assumptions E.1 – E.4, the variance-covariance matrix of the OLS estimator $\hat{\beta}$ is given by :

$$V(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$$

Note that as assumptions MLR.1 – MLR.5 imply assumptions E.1 – E.4, Property 2 also holds under the seminal assumptions MLR.1 – MLR.5.

- For the special case of the population model $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$, it may be shown that we have :

$$Var(\hat{\beta}_2|X) = \frac{\sigma^2}{SST_2 \left(1 - \rho_{spl}(x_{i2}, x_{i3})^2\right)}, \quad Var(\hat{\beta}_3|X) = \frac{\sigma^2}{SST_3 \left(1 - \rho_{spl}(x_{i2}, x_{i3})^2\right)}$$

and

$$Cov(\hat{\beta}_2, \hat{\beta}_3|X) = \frac{-\rho_{spl}(x_{i2}, x_{i3})\sigma^2}{\sqrt{SST_2}\sqrt{SST_3}\left(1 - \rho_{spl}(x_{i2}, x_{i3})^2\right)}$$

where $SST_2 = \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2$ and $SST_3 = \sum_{i=1}^{n}(x_{i3} - \bar{x}_3)^2$.

The above expressions allow to identify the components of the variance-covariance matrix of $\hat{\beta}$ :

[13] This property is the same as Theorem E.2 in Wooldridge (2016), Appendix E-2.

– The error variance $\sigma^2$ :

if $\sigma^2 \nearrow$ , then $Var(\hat{\beta}_2|X)$, $Var(\hat{\beta}_3|X)$ and $|Cov(\hat{\beta}_2, \hat{\beta}_3|X)| \nearrow$

– The variation of the explanatory variables around their mean :

if $SST_2 \nearrow$ and $SST_3 \nearrow$ , then $Var(\hat{\beta}_2|X)$, $Var(\hat{\beta}_3|X)$ and $|Cov(\hat{\beta}_2, \hat{\beta}_3|X)| \searrow$

– The sample size $n$ :

if $n \nearrow$ , $SST_2$ and $SST_3 \nearrow$ , then $Var(\hat{\beta}_2|X)$, $Var(\hat{\beta}_3|X)$ and $|Cov(\hat{\beta}_2, \hat{\beta}_3|X)| \searrow$

– The correlation between the explanatory variables :

if $|\rho_{spl}(x_{i2}, x_{i3})| \nearrow$ , then $Var(\hat{\beta}_2|X)$, $Var(\hat{\beta}_3|X)$ and $|Cov(\hat{\beta}_2, \hat{\beta}_3|X)| \nearrow$

### 3.1.4. Gauss-Markov theorem

- The Gauss-Markov theorem basically ensures that the OLS estimator $\hat{\beta}$ is, in some sense, the best estimator we can find for estimating the unknown parameter $\beta$ of a linear regression model satisfying assumptions E.1 – E.4. It may be expressed as follows[14] :

  Property 3 Gauss-Markov theorem
  Under assumptions E.1 – E.4, the OLS estimator $\hat{\beta}$ is the estimator which has the smallest (in a matrix sense) variance-covariance matrix among all linear unbiased estimators of $\beta$. It is the best linear unbiased estimator (BLUE) of $\beta$.

  Any linear estimator $\hat{\beta}^*$ of $\beta$ can be written as :

$$\hat{\beta}^* = AY$$

  where $A$ is a $k \times n$ matrix which can consist of constants or functions of $X$. The OLS estimator is simply obtained by setting $A = (X'X)^{-1} X'$. For the linear estimator $\hat{\beta}^*$, we have :

$$\begin{aligned} E(\hat{\beta}^*|X) &= E[AY|X]) = E[A(X\beta + u)|X] \\ &= AX\beta + AE(u|X) \\ &= AX\beta \end{aligned}$$

  so that $\hat{\beta}^*$ is an unbiased estimator of $\beta$ if, and only if, $AX = I$. For the OLS estimator, we indeed have $AX = (X'X)^{-1} X'X = I$.
  The Gauss-Markov theorem ensures that for any linear estimator $\hat{\beta}^*$ with $A$ such

[14] This property is the same as Theorem E.3 in Wooldridge (2016), Appendix E-2. For a detailed proof, see Wooldridge (2016) p. 725-726, Hayashi (2000) p. 29-30, or Lejeune (2011) p. 36-39 and p. 106 (under the assumption of $X$ nonstochastic).

that $AX = I$ (so that it is also unbiased), we have the matrix inequality:

$$V(\hat{\beta}^*|X) \geq V(\hat{\beta}|X) \tag{7}$$

which means that $V(\hat{\beta}^*|X) - V(\hat{\beta}|X) = D$, where $D$ a positive semi-definite matrix[15]. The matrix inequality (7) implies that, for any $k \times 1$ vector $a$ of constants, we have[16]:

$$Var(a'\hat{\beta}^*|X) \geq Var(a'\hat{\beta})$$

and in particular:

$$Var(\hat{\beta}_j^*|X) \geq Var(\hat{\beta}_j|X), \text{ for all } j = 1, ..., k$$

In other words, for estimating any parameter $\beta_j$, or any linear combination $a'\beta$ of the vector of parameters $\beta$, it is best to use the OLS estimator $\hat{\beta}$ rather than any other linear unbiased estimator $\hat{\beta}^*$.

- Note that:
  - as assumptions MLR.1 – MLR.5 imply assumptions E.1 – E.4, the Gauss-Markov theorem also holds under the seminal assumptions MLR.1 – MLR.5.
  - for obvious reason, assumptions E.1 – E.4 are often referred to as the 'Gauss-Markov assumptions'.

### 3.1.5. Distribution of $\hat{\beta}$ under normality

- Relying on assumptions E.1 – E.4, we obtained the first two (conditional) moments of the OLS estimator $\hat{\beta}$. If we add the normality assumption E.5, then the distribution of $\hat{\beta}$ is fully determined.

- For given $X$, the OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is a linear function of $Y$. Any linear function of random variables which are jointly normally distributed is itself normally distributed. This carries over conditional distributions. So, if the distribution of $Y$ conditional on $X$ is normal, then the distribution of $\hat{\beta}$ is also normal conditional on $X$[17]:

  Property 4 Distribution of $\hat{\beta}$ under normality
  Under assumptions E.1 – E.5, the distribution of the OLS estimator $\hat{\beta}$ is given by:

$$\hat{\beta}|X \sim N\left(\beta, \sigma^2(X'X)^{-1}\right) \tag{8}$$

  Note that as assumptions MLR.1 – MLR.6 imply assumptions E.1 – E.5, Property 4 also holds under the seminal assumptions MLR.1 – MLR.6.

---

[15] A $k \times k$ symmetric matrix $D$ is positive semi-definite if, for any $k \times 1$ vector $a$, we have $a'Da \geq 0$.

[16] Let $V(\hat{\beta}^*|X) - V(\hat{\beta}|X) = D$. We have: $Var(a'\hat{\beta}^*|X) = a'V(\hat{\beta}^*|X)a = a'\left[V(\hat{\beta}|X) + D\right]a \geq a'V(\hat{\beta}|X)a = Var(a'\hat{\beta}|X)$, because $a'Da \geq 0$.

[17] This property is the same as Theorem E.5 in Wooldridge (2016), Appendix E-3.

- The distributional result (8) is the basis for deriving exact in finite sample inference procedures (confidence interval, hypothesis testing).

- For inference, we will need an estimator of the variance-covariance matrix $V(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$ of the OLS estimator $\hat{\beta}$. This requires an estimator for $\sigma^2 = E(u^2)$. An unbiased estimator of $\sigma^2$ is given by:

$$\hat{s}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{\hat{u}'\hat{u}}{n-k}, \quad \text{where } \hat{u} = Y - X\hat{\beta}$$

This is formalized in the following property[18]:

Property 5 Unbiasedness of $\hat{s}^2$
Under assumptions E.1 – E.4, $\hat{s}^2$ is an unbiased estimator of $\sigma^2$:

$$E(\hat{s}^2|X) = \sigma^2$$

Note that:

– as $E(\hat{s}^2|X)$ does not depend on $X$, by the law of iterated expectations, we also have:

$$E(\hat{s}^2) = E\left[E(\hat{s}^2|X)\right] = E(\sigma^2) = \sigma^2$$

i.e., the unbiasedness also holds unconditionally.

– only the Gauss-Markov assumptions E.1 – E.4 are needed for $\hat{s}^2$ to be an unbiased estimator of $\sigma^2$.

– as assumptions MLR.1 – MLR.5 imply assumptions E.1 – E.4, Property 5 also holds under the seminal assumptions MLR.1 – MLR.5.

- With $\hat{s}^2$ as an estimator of $\sigma^2$, a natural estimator of $V(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$ is given by:

$$\hat{V}(\hat{\beta}) = \hat{s}^2 (X'X)^{-1} \tag{9}$$

As $\hat{s}^2$ is an unbiased estimator of $\sigma^2$, $\hat{V}(\hat{\beta})$ is also an unbiased estimator $V(\hat{\beta}|X)$:

$$\begin{aligned} E[\hat{V}(\hat{\beta})|X] &= E[\hat{s}^2 (X'X)^{-1}|X] = E(\hat{s}^2|X)(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} = V(\hat{\beta}|X) \end{aligned}$$

The diagonal elements $\hat{Var}(\hat{\beta}_j)$ of the $k \times k$ matrix estimator $\hat{V}(\hat{\beta})$ being the estimators of the variance $Var(\hat{\beta}_j|X)$ of the estimator $\hat{\beta}_j$ of the different parameters $\beta_j$ $(j = 1, ..., k)$, natural estimators of the standard error (also called

[18] This property is the same as Theorem E.4 in Wooldridge (2016), Appendix E-2. For a detailed proof, see Wooldridge (2016) p. 726, Hayashi (2000) p. 30-31, or Lejeune (2011) p. 42-43 and p. 107 (under the assumption of $X$ nonstochastic). Be aware: in Wooldridge (2016), $\hat{s}^2$ is noted $\hat{\sigma}^2$, and the degrees of freedom appearing in the expression of $\hat{\sigma}^2$ is $n-(k+1)$ rather than $(n-k)$ because he considers a model with $k$ explanatory variables + an intercept, while here the intercept is included in the set of the explanatory variables.

standard deviation) $s.e.(\hat{\beta}_j|X) = \sqrt{Var(\hat{\beta}_j|X)}$ of the estimator $\hat{\beta}_j$ of the different parameters $\beta_j$ are given by[19] :

$$s.\hat{e}.(\hat{\beta}_j) = \sqrt{V\hat{a}r(\hat{\beta}_j)}, \quad j = 1, ..., k$$

## 3.2. Asymptotic properties of OLS

- Asymptotic properties refer to sampling distribution properties which are valid when the sample size $n$ goes to infinity[20]. For studying the asymptotic properties of OLS, we switch back to the seminal set of assumptions MLR.1 – MLR.6, which assume random sampling.

- Before proceeding, it is worth summarizing what we already know of the finite sample properties of the OLS estimator $\hat{\beta}$ under the seminal set of assumptions MLR.1 – MLR.6 :

  – Under assumptions MLR.1 – MLR.4 (linearity in parameters, random sampling, no perfect colinearity and zero conditional mean), from Property 1, $\hat{\beta}$ is an unbiased estimator of $\beta$.

  – If assumption MLR.5 (homoskedasticity) is added to assumptions MLR.1 – MLR.4, from Property 2 and 3, the (conditional) variance-covariance matrix of $\hat{\beta}$ is given by $V(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$ and $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of $\beta$. Also, from Property 5, $\hat{s}^2$ is an unbiased estimator of $\sigma^2$.

  – If assumption MLR.6 (normality) is added to assumptions MLR.1 – MLR.5, the (conditional) distribution of $\hat{\beta}$ is normal and given by $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$.

- Hereafter, we show that :

  – under assumptions MLR.1 – MLR.4, $\hat{\beta}$ is a consistent estimator of $\beta$.

  – if assumption MLR.5 is added to assumptions MLR.1 – MLR.4, $\hat{\beta}$ is asymptotically normally distributed, and that regardless of whether or not assumption MLR.6 holds.

### 3.2.1. Consistency of $\hat{\beta}$

- An estimator $\hat{\theta}$ *converges in probability* to some constant $\theta$, which is noted

[19] Be aware : in Wooldridge (2016), $s.e.(\hat{\beta}_j|X)$ is noted $\text{sd}(\hat{\beta}_j)$, and $s.\hat{e}.(\hat{\beta}_j)$ is noted $\text{se}(\hat{\beta}_j)$.

[20] For a general discussion of asymptotic properties of estimators, including the concepts of consistency, asymptotic normality, law of large numbers and central limit theorem, see Wooldridge (2016), Appendix C-3.

$\hat{\theta} \xrightarrow{p} \theta$ or $\text{plim}(\hat{\theta}) = \theta$, if for any (arbitrarily small) $\varepsilon > 0$:

$$\lim_{n\to\infty} I\!P\left(|\hat{\theta} - \theta| < \varepsilon\right) = 1$$

In words, an estimator $\hat{\theta}$ converges in probability to $\theta$ if the probability that it yields a value as close as we wish from $\theta$ goes to 1 as $n \to \infty$. In terms of distribution, it means that the sampling distribution of $\hat{\theta}$ becomes more and more concentrated about $\theta$ as $n \to \infty$. If $\hat{\theta}$ is a vector, $\hat{\theta} \xrightarrow{p} \theta$ requires element-by-element convergence. When $\hat{\theta} \xrightarrow{p} \theta$, $\hat{\theta}$ is said to be a consistent estimator of $\theta$.

- We have the following property[21]:

  Property 6 Consistency of $\hat{\beta}$
  Under assumptions MLR.1 – MLR.4, the OLS estimator $\hat{\beta}$ is a consistent estimator of $\beta$:

$$\hat{\beta} \xrightarrow{p} \beta$$

  A sketch of the proof is as follows. Under assumptions MLR.1 – MLR.3, which imply assumptions E.1 and E.2, from (3), we have:

$$\begin{aligned}
\hat{\beta} &= \beta + (X'X)^{-1} X'u \\
&= \beta + \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i'u_i\right) \\
&= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i'X_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i'u_i\right) \qquad (10)
\end{aligned}$$

  Because under random sampling the observations $(X_i, y_i)$ are identically and independently distributed (i.i.d.) across $i$, both $X_i'X_i$ and $X_i'u_i$ are likewise i.i.d. across $i$, so that $\frac{1}{n}\sum_{i=1}^{n} X_i'X_i$ and $\frac{1}{n}\sum_{i=1}^{n} X_i'u_i$ are both sample average to which the law of large numbers (LLN) can be applied.
  If $\{Z_i\colon\ i = 1, ..., n\}$ are i.i.d. random variables with $E(Z_i) = m$, then by the LLN we have[22]:

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} m$$

  Under the zero conditional mean assumption MLR.4 $E(u_i|X_i) = 0$, by the law of iterated expectations, we have:

$$E(X_i'u_i) = E[E(X_i'u_i|X_i)] = E[X_i'E(u_i|X_i)] = E[X_i' \cdot 0] = 0 \qquad (11)$$

[21] This property is the same as Theorem 5.1 in Wooldridge (2016), Chapter 5. It is a special case of Theorem 11.1 in Wooldridge (2016), Chapter 11. For a sketch of the proof similar to the one developed below, see Wooldridge (2016) p. 728-729. For a more complete and rigorous treatment, see Wooldridge (2010).

[22] This holds for $Z_i$ being a scalar, a vector or a matrix.

Noting $E(X_i'X_i) = A$, from the LLN, we thus have :

$$\frac{1}{n}\sum_{i=1}^{n} X_i'X_i \xrightarrow{p} A \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i'u_i \xrightarrow{p} 0$$

so that, from (10), we finally have :

$$\hat{\beta} \xrightarrow{p} \beta + A^{-1} \cdot 0 = \beta$$

- Equation (11) in the above proof suggests that $\hat{\beta}$ would be consistent for $\beta$ under a weaker assumption than the zero conditional mean assumption MLR.4. It is actually sufficient to have $E\left(X_i'u_i\right) = 0$, i.e., the assumption MLR.4' :

  MLR.4' Zero mean and zero correlation
  The expected value of $u$ is zero and $u$ is uncorrelated with each explanatory variable $(x_2, ..., x_k)$ :

  $$E(u) = 0 \quad \text{and} \quad Cov(x_j, u) = 0, \text{ for } j = 2, ..., k$$

  Property 6' Consistency of $\hat{\beta}$ (bis)
  Under assumptions MLR.1 – MLR.3 and assumption MLR.4', the OLS estimator $\hat{\beta}$ is a consistent estimator of $\beta$ :

  $$\hat{\beta} \xrightarrow{p} \beta$$

  Note however that under assumption MLR.4', the OLS estimator $\hat{\beta}$ is no longer unbiased for $\beta$. It is only consistent for $\beta$.

### 3.2.2. Asymptotic normality of $\hat{\beta}$

- The consistency of $\hat{\beta}$ means that its sampling distribution becomes more and more concentrated about $\beta$ as $n \to \infty$. But as its sampling distribution is collapsing around $\beta$, the shape of this sampling distribution also becomes closer and closer to the normal distribution as $n \to \infty$[23]. And that happens regardless of the population distribution of $y$ given $(x_2, ..., x_k)$, i.e., without assuming normality. Technically, while the sampling distribution of $\hat{\beta}$ becomes degenerate at $\beta$ as $n \to \infty$, the sampling distribution of its scaled version $\sqrt{n}(\hat{\beta} - \beta)$ converges to a normal distribution.

- We have the following property[24] :

  Property 7 Asymptotic normality of $\hat{\beta}$
  Under assumptions MLR.1 – MLR.5, the OLS estimator $\hat{\beta}$ is asymptotically nor-

[23] For an illuminating graphical illustration (for the simple case of the estimation of a population mean), see Goldberger (1990), p. 94-97.

[24] This property is basically the same (but more general) as Theorem 5.2 in Wooldridge (2016), Chapter 5. It is a special case of Theorem 11.2 in Wooldridge (2016), Chapter 11. For a sketch of the proof similar to the one developed below, see Wooldridge (2016) p. 729-730. For a more complete and rigorous treatment, see Wooldridge (2010).

mally distributed:

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\sigma^2 A^{-1}), \quad \text{where } A = E(X_i'X_i) \tag{12}$$

The notation $\xrightarrow{d}$ means convergence in distribution.

A sketch of the proof is as follows. Under assumptions MLR.1 – MLR.3, which imply assumptions E.1 and E.2, from (10), we have:

$$\hat{\beta} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i'X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i'u_i\right)$$

$$\Leftrightarrow \sqrt{n}(\hat{\beta}-\beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i'X_i\right)^{-1}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n} X_i'u_i\right) \tag{13}$$

As already outlined, under random sampling, both $X_i'X_i$ and $X_i'u_i$ are i.i.d. across $i$. From the law of large numbers (LLN), we have $\frac{1}{n}\sum_{i=1}^{n} X_i'X_i \xrightarrow{p} E(X_i'X_i) = A$, and we can write:

$$\sqrt{n}(\hat{\beta}-\beta) \overset{as}{=} A^{-1}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n} X_i'u_i\right) \tag{14}$$

where $\overset{as}{=}$ means 'asymptotically equivalent', so that $\sqrt{n}(\hat{\beta}-\beta)$ has asymptotically the same distribution as $A^{-1}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n} X_i'u_i\right)$.
Now, if $\{Z_i\colon\ i = 1,...,n\}$ are i.i.d. $(k\times 1)$ random vectors with $E(Z_i) = m$ and $V(Z_i) = \Sigma$, then by the central limit theorem (CLT) we have:

$$\sqrt{n}(\bar{Z}_n - m) = n^{-\frac{1}{2}}\sum_{i=1}^{n}(Z_i - m) \xrightarrow{d} N(0,\Sigma)$$

As already outlined, under the zero conditional mean assumption MLR.4 $E(u_i|X_i) = 0$, from (11), we have $E(X_i'u_i) = 0$. From the CLT, we thus have:

$$n^{-\frac{1}{2}}\sum_{i=1}^{n} X_i'u_i \xrightarrow{d} N(0,B), \quad \text{where } B = V(X_i'u_i)$$

so that[25]:

$$A^{-1}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n} X_i'u_i\right) \xrightarrow{d} N(0, A^{-1}BA^{-1})$$

On the other hand, under the homoskedasticity assumption MLR.5 $Var(u_i|X_i) = E(u_i^2|X_i) = \sigma^2$, by the law of iterated expectations, we have:

$$\begin{aligned} B &= V(X_i'u_i) = E(u_i^2X_i'X_i) = E\left[E\left(u_i^2X_i'X_i|X_i\right)\right] \\ &= E\left[E\left(u_i^2|X_i\right)X_i'X_i\right] = E\left(\sigma^2X_i'X_i\right) = \sigma^2E\left(X_i'X_i\right) = \sigma^2A \end{aligned}$$

[25] because a linear function of jointly normally distributed random variables is itself normally distributed, and $A^{-1}$ is a symmetric matrix.

so that $A^{-1}BA^{-1} = \sigma^2 A^{-1}$. From (14), we thus finally have :

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 A^{-1})$$

- The limiting distributional result (12) provides an approximate finite sample distribution for the OLS estimator $\hat{\beta}$, which can be used – when $n$ is sufficiently large – for performing inference (confidence interval, hypothesis testing) without having to rely on the normality assumption MLR.6. From (12), in terms of approximation, we have :

$$\sqrt{n}(\hat{\beta} - \beta) \approx N(0, \sigma^2 A^{-1})$$

so that :

$$\hat{\beta} \approx N(\beta, \sigma^2 A^{-1}/n) \tag{15}$$

i.e., for $n$ sufficiently large[26], $\hat{\beta}$ can be treated as approximately normal with mean $\beta$ and variance-covariance matrix $\sigma^2 A^{-1}/n$. The variance-covariance matrix $\sigma^2 A^{-1}/n$ is usually called the 'asymptotic variance of $\hat{\beta}$' and noted $Avar(\hat{\beta})$. Note that $Avar(\hat{\beta}) \to 0$ as $n \to \infty$. Note further that if we replace $A = E(X_i' X_i)$ by its consistent estimator $\frac{1}{n}\sum_{i=1}^{n} X_i' X_i = X'X/n$, then $Avar(\hat{\beta})$ becomes :

$$Avar(\hat{\beta}) = \frac{\sigma^2 A^{-1}}{n} \approx \frac{\sigma^2 (X'X/n)^{-1}}{n} = \sigma^2 (X'X)^{-1}$$

This is the same as the exact in finite sample variance-covariance matrix $V(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$ obtained in Property 2.

- The approximate distributional result (15) is the basis for showing that the exact in finite sample inference procedures (confidence interval, hypothesis testing) derived under assumptions E.1 – E.5, which are thus likewise exact in finite sample under the seminal assumptions MLR.1 – MLR.6, remain asymptotically valid – i.e., approximately valid for $n$ sufficiently large – under assumptions MLR.1 – MLR.5, i.e., without having to rely on the normality assumption MLR.6.

- For inference based on the limiting distributional result (12), or equivalently on the approximate distributional result (15), we will need an estimator of $Avar(\hat{\beta}) = \sigma^2 A^{-1}/n$. This requires consistent estimators of $\sigma^2$ and $A$. We already outlined that $X'X/n$ is a consistent estimator of $A$ (this follows from the LLN). A consistent estimator of $\sigma^2$ is simply given by its unbiased estimator $\hat{s}^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{\hat{u}'\hat{u}}{n-k}$. This is formalized in the following property[27] :

  Property 8 Consistency of $\hat{s}^2$
  Under MLR.1 – MLR.5, $\hat{s}^2$ is a consistent estimator of $\sigma^2$ :

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

[26] The larger $n$, the better the approximation. An absolute minimum is $n > 30$.

[27] This property is outlined in Theorem 5.2 in the Wooldridge (2016), Chapter 5. For a detailed proof (under weaker assumptions than MLR.1 – MLR.5), see Hayashi (2000) p. 115-116.

Here is the intuition of this property: from the LLN, $\frac{1}{n}\sum_{i=1}^{n} u_i^2 \xrightarrow{p} E(u_i^2) = \sigma^2$. As $\hat{u}_i$ converges to $u_i$ (because $\hat{\beta} \xrightarrow{p} \beta$), we also have $\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{p} \sigma^2$. The final result follows from the fact that $\hat{s}^2 = (\frac{n}{n-k})\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$, and that $\frac{n}{n-k} \to 1$ as $n \to \infty$.

- With $\hat{s}^2$ and $X'X/n$ as consistent estimators of $\sigma^2$ and $A$, an estimator of $Avar(\hat{\beta}) = \sigma^2 A^{-1}/n$ is given by:

$$\hat{V}(\hat{\beta}) = \hat{s}^2 \left(X'X\right)^{-1}$$

This is the same as the unbiased estimator of $V(\hat{\beta}|X) = \sigma^2 \left(X'X\right)^{-1}$ outlined in (9). As in the case of the estimation of $V(\hat{\beta}|X) = \sigma^2 \left(X'X\right)^{-1}$, the diagonal elements $V\hat{a}r(\hat{\beta}_j)$ of the $k \times k$ matrix estimator $\hat{V}(\hat{\beta})$ being the estimators of the variance $Avar(\hat{\beta}_j)$ of the estimator $\hat{\beta}_j$ of the different parameters $\beta_j$ $(j = 1, ..., k)$, natural estimators of the asymptotic standard error $As.e.(\hat{\beta}_j) = \sqrt{Avar(\hat{\beta}_j)}$ of the estimator $\hat{\beta}_j$ of the different parameters $\beta_j$ are likewise given by:

$$s.\hat{e}.(\hat{\beta}_j) = \sqrt{V\hat{a}r(\hat{\beta}_j)}, \quad j = 1, ..., k$$

# 4. Inference

## 4.1. Exact confidence interval and hypothesis testing about a single parameter

- We first consider exact in finite sample inference about a single parameter[28], i.e., inference about a single parameter when the Gauss-Markov assumptions E.1 – E.4 as well as the normality assumption E.5 hold, which is the case if the seminal assumptions MLR.1 – MLR.5 as well as the normality assumption MLR.6 hold.

- From Property 4, under assumptions E.1 – E.5, and thus also under the seminal assumptions MLR.1 – MLR.6, we have:

$$\hat{\beta}|X \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$$

which implies that, for each parameter $\beta_j$ $(j = 1, ..., k)$, we have:

$$\hat{\beta}_j|X \sim N(\beta_j, Var(\hat{\beta}_j|X))$$

so that, conditional on $X$, we have:

$$\hat{z} = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j|X)} \sim N(0,1) \tag{16}$$

[28] For a general discussion about the basic concepts underlying confidence intervals and hypothesis testing, see Wooldridge (2016), Appendix C-5 and C-6.

where $s.e.(\hat{\beta}_j|X) = \sqrt{Var(\hat{\beta}_j|X)} = \sqrt{\sigma^2 q_{jj}}$, with[29] $q_{jj} = [(X'X)^{-1}]_{jj}$.

In particular, if $\beta_j = \beta_j^o$, still conditional on $X$, we have:

$$\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j|X)} \sim N(0,1) \tag{17}$$

while if $\beta_j = \beta_j^* \neq \beta_j^o$, we have:

$$\hat{z}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.e.(\hat{\beta}_j|X)} \sim N\left(\frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j|X)}, 1\right) \tag{18}$$

- It may be shown that, under the same assumptions, conditional on $X$, we have that $\hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)$ and that $\hat{z}$ and $\hat{v}$ are independent[30,31], so that from the definition of the Student distribution[32], still conditional on $X$, we have:

$$\hat{t} = \frac{\hat{z}}{\sqrt{\frac{\hat{v}}{n-k}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 q_{jj}}}}{\sqrt{\frac{\hat{s}^2}{\sigma^2}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{s}^2 q_{jj}}} \sim t(n-k)$$

i.e.:

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-k) \tag{19}$$

where $s.\hat{e}.(\hat{\beta}_j) = \sqrt{V\hat{a}r(\hat{\beta}_j)} = \sqrt{\hat{s}^2 q_{jj}}$.

In particular, if $\beta_j = \beta_j^o$, still conditional on $X$, we have:

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \sim t(n-k) \tag{20}$$

while if $\beta_j = \beta_j^* \neq \beta_j^o$, it may be checked[33] that we have:

$$\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \sim t(\delta^*, n-k)\,, \tag{21}$$

---

[29] i.e., $q_{jj}$ denotes the $(j,j)$ element of $(X'X)^{-1}$.

[30] For a proof that, conditional on $X$, $\hat{v} \sim \chi^2(n-k)$, see Wooldridge (2016) p. 727, Hayashi (2000) p. 36-37, or Lejeune (2011) p. 48-49 and p. 108 (under the assumption of $X$ nonstochastic). Recall that in Wooldridge (2016), $\hat{s}^2$ is noted $\hat{\sigma}^2$, and the degrees of freedom appearing in the equations is $n-(k+1)$ rather than $(n-k)$ because he considers a model with $k$ explanatory variables + an intercept, while here the intercept is included in the set of the explanatory variables.

[31] The independence of $\hat{z}$ and $\hat{v}$ (conditional on X) follows from the independence of $\hat{\beta}$ and $\hat{s}^2$ (conditional on $X$). For a proof of the latter, see again Wooldridge (2016) p. 727 or Hayashi (2000) p. 36-37.

[32] If $z \sim N(0,1)$, $v \sim \chi^2(m)$, and $z$ and $v$ are independent, then $t = \frac{z}{\sqrt{\frac{v}{m}}} \sim t(m)$, where $t(m)$ denotes the Student distribution with $m$ degrees of freedom. See Wooldridge (2016), Appendix B-5e. This carries over conditional distributions.

[33] It follows from the facts that, conditional on $X$, $\hat{v} \sim \chi^2(n-k)$ and $\hat{z}_o$ and $\hat{v}$ are independent, and from the definition of the non-central Student distribution given below.

where $t(\delta^*, n-k)$ denotes the non-central Student distribution[34] with $(n-k)$ degrees of freedom and non-centrality parameter $\delta^* = \frac{\beta_j^* - \beta_j^o}{s.e.(\hat{\beta}_j|X)}$.

In words, if the unknown variance $\sigma^2$ appearing in the standard error $s.e.(\hat{\beta}_j|X)$ of statistics (16), (17) and (18) is replaced by its unbiased estimator $\hat{s}^2$, so that the standard error $s.e.(\hat{\beta}_j|X)$ is replaced by its estimator $s.\hat{e}.(\hat{\beta}_j)$, then the distribution of (16), (17) and (18) switches from normal to Student. Note that it only makes a real difference when the sample size $n$ is small, because for $m$ large, $t(m) \approx N(0,1)$ and $t(\delta, m) \approx N(\delta, 1)$.

- The distributional result (19) is the basis for deriving a confidence interval for $\beta_j$. On the other hand, the distributional results (20) and (21) are the basis for hypothesis testing about $\beta_j$.

### 4.1.1. Confidence interval for $\beta_j$

- The distributional result (19) holds conditional on $X$. But as the conditional distribution of $\hat{t} = \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)}$ actually does not depend on $X$, it also holds unconditionally[35], and we can write:

$$\mathbb{P}\left(-t_{n-k;1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j}{s.\hat{e}.(\hat{\beta}_j)} \leq t_{n-k;1-\frac{\alpha}{2}}\right) = 1-\alpha \tag{22}$$

where $t_{n-k;1-\frac{\alpha}{2}}$ is the quantile of order $1-\frac{\alpha}{2}$ of the $t(n-k)$ Student distribution, i.e., the value such that $\mathbb{P}(t \leq t_{n-k;1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$, where $t \sim t(n-k)$. From (22), we have:

$$\mathbb{P}\left(\hat{\beta}_j - t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\right) = 1-\alpha$$

so that a $(1-\alpha)\times 100\%$ confidence interval for $\beta_j$ is given by:

$$\left[\hat{\beta}_j - t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\,;\, \hat{\beta}_j + t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)\right] \tag{23}$$

### 4.1.2. *t*-tests about $\beta_j$

- The distributional result (20) states that if the value of $\beta_j$ in the population is equal to $\beta_j^o$, then the statistic $\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)}$ will yield values distributed around zero according to its $t(n-k)$ Student distribution. Note that this distribution result does not hold only conditional on $X$: as the conditional distribution of $\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)}$ actually does not depend on $X$, it also holds unconditionally. On the other hand, the distributional result (21) states that if the value of $\beta_j$ in

[34] If $z \sim N(\delta, 1)$, $v \sim \chi^2(m)$, and $z$ and $v$ are independent, then $t = \frac{z}{\sqrt{\frac{v}{m}}} \sim t(\delta, m)$. This carries over conditional distributions.

[35] This should not be confused with the fact that the value of $\hat{t}$ itself depends on $X$.

the population is equal to $\beta_j^* \neq \beta_j^o$, then the same statistic $\hat{t}_o = \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)}$ will yield values distributed with a central tendency away from zero – either positive if $\beta_j^* > \beta_j^o$ or negative if $\beta_j^* < \beta_j^o$ – according to its $t(\delta^*, n-k)$ Student distribution.

This behavior makes $\hat{t}_o$ a natural test statistic for testing hypotheses such as $\mathrm{H}_0 : \beta_j = \beta_j^o$ against $\mathrm{H}_1 : \beta_j \neq \beta_j^o$ (two-sided test) or $\mathrm{H}_0 : \beta_j \leq \beta_j^o$ (resp. $\beta_j \geq \beta_j^o$) against $\mathrm{H}_1 : \beta_j > \beta_j^o$ (resp. $\beta_j < \beta_j^o$) (one-sided tests).

- A two-sided test at (significance) level $\alpha$ of $\mathrm{H}_0 : \beta_j = \beta_j^o$ against $\mathrm{H}_1 : \beta_j \neq \beta_j^o$ is given by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } |\hat{t}_o| = \left| \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \right| > t_{n-k;1-\frac{\alpha}{2}} \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

where the critical value $t_{n-k;1-\frac{\alpha}{2}}$ is the quantile of order $1 - \frac{\alpha}{2}$ of the $t(n-k)$ Student distribution. The $p$-value of this test[36], for a value $\hat{t}_o^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{t}_o^*} = I\!P(|t| > |\hat{t}_o^*|), \quad \text{where } t \sim t(n-k)$$

This two-sided test can equivalently be performed using the decision rule[37] :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } \beta_j^o \text{ does not fall in the } (1-\alpha) \times 100\% \\ \qquad\qquad\quad \text{confidence interval (23) for } \beta_j \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

- A right-sided test at (significance) level $\alpha$ of $\mathrm{H}_0 : \beta_j \leq \beta_j^o$ against $\mathrm{H}_1 : \beta_j > \beta_j^o$ is given by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} > t_{n-k;1-\alpha} \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

where the critical value $t_{n-k;1-\alpha}$ is the quantile of order $1-\alpha$ of the $t(n-k)$ Student distribution. The $p$-value of this test, for a value $\hat{t}_o^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{t}_o^*} = I\!P(t > \hat{t}_o^*), \quad \text{where } t \sim t(n-k)$$

- Symmetrically, a left-sided test at (significance) level $\alpha$ of $\mathrm{H}_0 : \beta_j \geq \beta_j^o$ against

[36] As a reminder, the $p$-value of a test is the smallest (significance) level $\alpha$ for which we can reject the null hypothesis $\mathrm{H}_0$ of the test based on the value $\hat{t}_o^*$ of the test statistic obtained in a particular sample.

[37] The two-sided $t$-test at level $\alpha$ does not reject $\mathrm{H}_0$ if $|\hat{t}_o| = \left| \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \right| \leq t_{n-k;1-\frac{\alpha}{2}}$, i.e., if $-t_{n-k;1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} \leq t_{n-k;1-\frac{\alpha}{2}}$, or equivalently if $\hat{\beta}_j - t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j) \leq \beta_j^o \leq \hat{\beta}_j + t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(\hat{\beta}_j)$. The endpoints of this latter interval are the endpoints of the $(1-\alpha) \times 100\%$ confidence interval for $\beta_j$.

$H_1: \beta_j < \beta_j^o$ is given by the decision rule:

$$\begin{cases} \text{- Reject } H_0 \text{ if } \hat{t}_o = \dfrac{\hat{\beta}_j - \beta_j^o}{s.\hat{e}.(\hat{\beta}_j)} < t_{n-k;\alpha} (= -t_{n-k;1-\alpha}) \\ \text{- Do not reject } H_0 \text{ otherwise} \end{cases}$$

where the critical value $t_{n-k;\alpha}$ $(= -t_{n-k;1-\alpha})$ is the quantile of order $\alpha$ of the $t(n-k)$ Student distribution. The $p$-value of this test, for a value $\hat{t}_o^*$ of the test statistic obtained in a particular sample, is given by:

$$p_{\hat{t}_o^*} = I\!P(t < \hat{t}_o^*), \quad \text{where } t \sim t(n-k)$$

## 4.2. Exact confidence interval and hypothesis testing about a single linear combination of parameters

- Exact in finite sample confidence interval and hypothesis tests about a single linear combination $R_0\beta$ of the vector of parameters $\beta$, where $R_0$ is a (row) $1 \times k$ vector of constants, can be derived in the same way as exact in finite sample confidence interval and hypothesis tests about a single parameter.

- From Property 4, under assumptions E.1 – E.5, and thus also under the seminal assumptions MLR.1 – MLR.6, we have:

$$\hat{\beta}|X \sim N\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

which implies that, for any single linear combination $R_0\beta$ of $\beta$, we have:

$$R_0\hat{\beta}|X \sim N(R_0\beta, \sigma^2 R_0 (X'X)^{-1} R_0')$$

so that, conditional on $X$, we have:

$$\hat{z} = \frac{R_0\hat{\beta} - R_0\beta}{s.e.(R_0\hat{\beta}|X)} \sim N(0,1) \tag{24}$$

where $s.e.(R_0\hat{\beta}|X) = \sqrt{Var(R_0\hat{\beta}|X)} = \sqrt{\sigma^2 R_0 (X'X)^{-1} R_0'} = \sqrt{R_0 V(\hat{\beta}|X) R_0'}$.

In particular, if $R_0\beta = r_0$, still conditional on $X$, we have:

$$\hat{z}_o = \frac{R_0\hat{\beta} - r_0}{s.e.(R_0\hat{\beta}|X)} \sim N(0,1) \tag{25}$$

while if $R_0\beta \neq r_0$, we have:

$$\hat{z}_o = \frac{R_0\hat{\beta} - r_0}{s.e.(R_0\hat{\beta}|X)} \sim N\left(\frac{R_0\beta - r_0}{s.e.(R_0\hat{\beta}|X)}, 1\right) \tag{26}$$

- As in the single parameter case, under the same assumptions, conditional on $X$,

we have that $\hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)$ and that $\hat{z}$ and $\hat{v}$ are independent[38], so that from the definition of the Student distribution, still conditional on $X$, we have:

$$\hat{t} = \frac{\hat{z}}{\sqrt{\frac{\hat{v}}{n-k}}} = \frac{\frac{R_0\hat{\beta}-R_0\beta}{\sqrt{\sigma^2 R_0(X'X)^{-1}R_0'}}}{\sqrt{\frac{\hat{s}^2}{\sigma^2}}} = \frac{R_0\hat{\beta}-R_0\beta}{\sqrt{\hat{s}^2 R_0(X'X)^{-1}R_0'}} \sim t(n-k)$$

i.e.:

$$\hat{t} = \frac{R_0\hat{\beta}-R_0\beta}{s.\hat{e}.(R_0\hat{\beta})} \sim t(n-k) \tag{27}$$

where $s.\hat{e}.(R_0\hat{\beta}) = \sqrt{\hat{Var}(R_0\hat{\beta})} = \sqrt{\hat{s}^2 R_0(X'X)^{-1}R_0'} = \sqrt{R_0\hat{V}(\hat{\beta})R_0'}$.

In particular, if $R_0\beta = r_0$, still conditional on $X$, we have:

$$\hat{t}_o = \frac{R_0\hat{\beta}-r_0}{s.\hat{e}.(R_0\hat{\beta})} \sim t(n-k) \tag{28}$$

while if $R_0\beta \neq r_0$, it may be checked[39] that we have:

$$\hat{t}_o = \frac{R_0\hat{\beta}-r_0}{s.\hat{e}.(R_0\hat{\beta})} \sim t(\delta^*, n-k)\,, \quad \text{where } \delta^* = \frac{R_0\beta-r_0}{s.e.(R_0\hat{\beta}|X)} \tag{29}$$

In words, if the unknown variance $\sigma^2$ appearing in the standard error $s.e.(R_0\hat{\beta}|X)$ of statistics (24), (25) and (26) is replaced by its unbiased estimator $\hat{s}^2$, so that the standard error $s.e.(R_0\hat{\beta}|X)$ is replaced by its estimator $s.\hat{e}.(R_0\hat{\beta})$, then the distribution of (24), (25) and (26) switches from normal to Student. But it only makes a real difference when the sample size $n$ is small, because for $m$ large, $t(m) \approx N(0,1)$ and $t(\delta, m) \approx N(\delta, 1)$.

- The distributional results (27), (28) and (29) are basically the same as the distributional results (19) (20) and (21) derived in the single parameter case: the single parameter $\hat{\beta}_j$ is simply replaced by the linear combination $R_0\hat{\beta}$, and likewise $\beta_j$ by $R_0\beta$, $\beta_j^o$ by $r_0$ and $s.\hat{e}.(\hat{\beta}_j)$ by $s.\hat{e}.(R_0\hat{\beta})$. Similarly, the distributional result (27) is the basis for deriving a confidence interval for $R_0\beta$, and the distributional results (28) and (29) are the basis for hypothesis testing about $R_0\beta$.

### 4.2.1. Confidence interval for $R_0\beta$

- As in the single parameter case, the distributional result (27) holds conditional

[38] As in the single parameter case, the independence of $\hat{z}$ and $\hat{v}$ (conditional on X) follows from the independence of $\hat{\beta}$ and $\hat{s}^2$ (conditional on $X$).

[39] It follows from the facts that, conditional on $X$, $\hat{v} \sim \chi^2(n-k)$ and $\hat{z}_o$ and $\hat{v}$ are independent, and from the definition of the non-central Student distribution previously given.

on $X$, but also unconditionally, and we can write :

$$\mathbb{P}\left(-t_{n-k;1-\frac{\alpha}{2}} \leq \frac{R_0\hat{\beta} - R_0\beta}{s.\hat{e}.(R_0\hat{\beta})} \leq t_{n-k;1-\frac{\alpha}{2}}\right) = 1-\alpha$$

or equivalently :

$$\mathbb{P}\left(R_0\hat{\beta} - t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(R_0\hat{\beta}) \leq R_0\beta \leq R_0\hat{\beta} + t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(R_0\hat{\beta})\right) = 1-\alpha$$

so that a $(1-\alpha)\times 100\%$ confidence interval for $R_0\beta$ is given by :

$$\left[R_0\hat{\beta} - t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(R_0\hat{\beta})\,;\; R_0\hat{\beta} + t_{n-k;1-\frac{\alpha}{2}} s.\hat{e}.(R_0\hat{\beta})\right] \tag{30}$$

- Remark : by smartly reparametrizing the original model, it is in practice possible to calculate the above confidence interval for a linear combination $R_0\beta$ as a confidence interval for a single parameter. See Wooldridge (2016) Section 4-4 for details.

### 4.2.2. *t*-tests about $R_0\beta$

- Again as in the single parameter case, the distributional result (28) states that if the value of $\beta$ in the population is such that $R_0\beta = r_0$, then the statistic $\hat{t}_o = \frac{R_0\hat{\beta}-r_0}{s.\hat{e}.(R_0\hat{\beta})}$ will yield values distributed around zero according to its $t(n-k)$ Student distribution. Note likewise that this distribution result does not hold only conditional on $X$ : it also holds unconditionally. On the other hand, the distributional result (29) states that if the value of $\beta$ in the population is such that $R_0\beta \neq r_0$, then the same statistic $\hat{t}_o = \frac{R_0\hat{\beta}-r_0}{s.\hat{e}.(R_0\hat{\beta})}$ will yield values distributed with a central tendency away from zero – either positive if $R_0\beta > r_0$ or negative if $R_0\beta < r_0$ – according to its $t(\delta^*, n-k)$ Student distribution.

  This behavior likewise makes $\hat{t}_o$ a natural test statistic for testing hypotheses such as $\mathrm{H}_0 : \; R_0\beta = r_0$ against $\mathrm{H}_1 : \; R_0\beta \neq r_0$ (two-sided test) or $\mathrm{H}_0 : \; R_0\beta \leq r_0$ (resp. $R_0\beta \geq r_0$) against $\mathrm{H}_1 : \; R_0\beta > r_0$ (resp. $R_0\beta < r_0$) (one-sided tests).

- A two-sided test at (significance) level $\alpha$ of $\mathrm{H}_0 : \; R_0\beta = r_0$ against $\mathrm{H}_1 : \; R_0\beta \neq r_0$ is given by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } |\hat{t}_o| = \left|\dfrac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})}\right| > t_{n-k;1-\frac{\alpha}{2}} \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

  The $p$-value of this test, for a value $\hat{t}_o^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{t}_o^*} = \mathbb{P}(|t| > |\hat{t}_o^*|)\,, \quad \text{where } t \sim t(n-k)$$

This two-sided test can also equivalently be performed using the decision rule[40] :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } & r_0 \text{ does not fall in the } (1-\alpha)\times 100\% \\ & \text{confidence interval (30) for } R_0\beta \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

- A right-sided test at (significance) level $\alpha$ of $\mathrm{H}_0 : \; R_0\beta \leq r_0$ against $\mathrm{H}_1 : \; R_0\beta > r_0$ is given by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } \hat{t}_o = \dfrac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})} > t_{n-k;1-\alpha} \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

The $p$-value of this test, for a value $\hat{t}_o^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{t}_o^*} = I\!P(t > \hat{t}_o^*) \,, \quad \text{where } t \sim t(n-k)$$

- Symmetrically, a left-sided test at (significance) level $\alpha$ of $\mathrm{H}_0 : \; R_0\beta \geq r_0$ against $\mathrm{H}_1 : \; R_0\beta < r_0$ is given by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } \hat{t}_o = \dfrac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})} < t_{n-k;\alpha} (= -t_{n-k;1-\alpha}) \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases}$$

The $p$-value of this test, for a value $\hat{t}_o^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{t}_o^*} = I\!P(t < \hat{t}_o^*) \,, \quad \text{where } t \sim t(n-k)$$

- Remark : As for the confidence interval for $R_0\beta$, by smartly reparametrizing the original model, it is in practice possible to perform the above two-sided and one-sided $t$-tests about a linear combination $R_0\beta$ as two-sided and one-sided $t$-tests about a single parameter. Again, see Wooldridge (2016) Section 4-4 for details.

## 4.3. Confidence interval and hypothesis testing without the normality assumption

- Hereafter, we show that the exact in finite sample confidence intervals and hypothesis tests for a single parameter and a single linear combination of parameters derived under assumptions E.1 – E.5, which are thus likewise exact in finite sample under the seminal assumptions MLR.1 – MLR.6, remain asymptot-

[40] The two-sided $t$-test at level $\alpha$ does not reject $\mathrm{H}_0$ if $|\hat{t}_o| = \left|\frac{R_0\hat{\beta}-r_0}{s.\hat{e}.(R_0\hat{\beta})}\right| \leq t_{n-k;1-\frac{\alpha}{2}}$, i.e., if $-t_{n-k;1-\frac{\alpha}{2}} \leq \frac{R_0\hat{\beta}-r_0}{s.\hat{e}.(R_0\hat{\beta})} \leq t_{n-k;1-\frac{\alpha}{2}}$, or equivalently if $R_0\hat{\beta} - t_{n-k;1-\frac{\alpha}{2}}\, s.\hat{e}.(R_0\hat{\beta}) \leq r_0 \leq R_0\hat{\beta} + t_{n-k;1-\frac{\alpha}{2}}\, s.\hat{e}.(R_0\hat{\beta})$. The endpoints of this latter interval are the endpoints of the $(1-\alpha)\times 100\%$ confidence interval for $R_0\beta$.

ically valid – i.e., approximately valid for $n$ sufficiently large – under assumptions MLR.1 – MLR.5, i.e., without having to rely on the normality assumption MLR.6.

- Because confidence interval and hypothesis tests for a single parameter are just special cases of confidence interval and hypothesis tests for a single linear combination of parameters[41], we focus on the latter.

- From Property 7, under assumptions MLR.1 – MLR.5, we have the limiting distributional result :

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 A^{-1}), \quad \text{where } A = E(X_i' X_i)$$

or, in terms of approximation for $n$ sufficiently large :

$$\hat{\beta} \approx N(\beta, \sigma^2 A^{-1}/n)$$

so that, for any single linear combination $R_0\beta$ of $\beta$, we have :

$$R_0\hat{\beta} \approx N(R_0\beta, \sigma^2 R_0(A^{-1}/n)R_0')$$

and

$$\hat{z}_{as} = \frac{R_0\hat{\beta} - R_0\beta}{As.e.(R_0\hat{\beta})} \approx N(0,1) \tag{31}$$

where $As.e.(R_0\hat{\beta}) = \sqrt{Avar(R_0\hat{\beta})} = \sqrt{\sigma^2 R_0(A^{-1}/n)R_0'} = \sqrt{R_0 Avar(\hat{\beta})R_0'}$.

In particular, if $R_0\beta = r_0$, we have :

$$\hat{z}_{aso} = \frac{R_0\hat{\beta} - r_0}{As.e.(R_0\hat{\beta})} \approx N(0,1) \tag{32}$$

while if $R_0\beta \neq r_0$, we have :

$$\hat{z}_{aso} = \frac{R_0\hat{\beta} - r_0}{As.e.(R_0\hat{\beta})} \approx N\left(\frac{R_0\beta - r_0}{As.e.(R_0\hat{\beta})}, 1\right) \tag{33}$$

- It may be shown that, from an asymptotic point of view, we can replace the unknown parameters $\sigma^2$ and $A$ appearing in the asymptotic standard error $As.e.(R_0\hat{\beta})$ of statistics (31), (32) and (33) by their consistent estimator $\hat{s}^2$ and $X'X/n$ – which means that $As.e.(R_0\hat{\beta})$ is replaced by its estimator $s.\hat{e}.(R_0\hat{\beta})$ – without affecting their approximate distribution[42], so that we also have :

$$\hat{t} = \frac{R_0\hat{\beta} - R_0\beta}{s.\hat{e}.(R_0\hat{\beta})} \approx N(0,1) \tag{34}$$

[41] A confidence interval or hypothesis $t$-test for any $\beta_j$ is obtained by defining $R_0$ as a $1 \times k$ vector of zeros with a one in the $j$-th position.

[42] Properly showing this requires advanced asymptotic theory as developped in Wooldridge (2010), Chapter 3. See also Hayashi (2010), Chapter 2.

where $s.\hat{e}.(R_0\hat{\beta}) = \sqrt{\hat{Var}(R_0\hat{\beta})} = \sqrt{\hat{s}^2 R_0 (X'X)^{-1} R_0'} = \sqrt{R_0 \hat{V}(\hat{\beta}) R_0'}$.

In particular, if $R_0\beta = r_0$, we have:

$$\hat{t}_o = \frac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})} \approx N(0,1) \tag{35}$$

while if $R_0\beta \neq r_0$, we have:

$$\hat{t}_o = \frac{R_0\hat{\beta} - r_0}{s.\hat{e}.(R_0\hat{\beta})} \approx N\left(\frac{R_0\beta - r_0}{As.e.(R_0\hat{\beta})}, 1\right) \tag{36}$$

- The distributional results (34), (35) and (36) are the same as the distributional results (27), (28) and (29) on which rely the exact in finite sample confidence interval and hypothesis tests about $R_0\beta$, except that they only hold asymptotically – i.e., approximately for $n$ sufficiently large –, and that they feature normal rather than Student distributions. Accordingly, deriving approximately valid, for $n$ sufficiently large, confidence interval and hypothesis tests about $R_0\beta$ on the basis of the distributional results (34), (35) and (36) will yield the same confidence interval and hypothesis tests as the exact in finite sample ones, except that they will feature quantiles of the standard normal $N(0,1)$ rather than quantiles of the $t(n-k)$ Student distribution.

- As $n \to \infty$, the quantiles of the $t(n-k)$ Student distribution become the same as the quantiles of the standard normal $N(0,1)$[43]. As a result, the exact in finite sample confidence interval and hypothesis tests about a single linear combination $R_0\beta$ of parameters derived under assumptions E.1 – E.5, which are thus likewise exact in finite sample under the seminal assumptions MLR.1 – MLR.6, remain asymptotically valid – i.e., approximately valid for $n$ sufficiently large – under assumptions MLR.1 – MLR.5, i.e., without having to rely on the normality assumption MLR.6. As confidence interval and hypothesis tests for a single parameter are just special cases of confidence interval and hypothesis tests for a single linear combination of parameters, the same hold for the exact in finite sample confidence interval and hypothesis tests about a single of parameter $\beta_j$.

## 4.4. Testing multiple linear restrictions

- It is usual that we want to test not just one but multiple hypotheses about parameters, or further multiple hypotheses about linear combination of parameters. Such tests are special cases of the general test:

$$\mathrm{H}_0: R_0\beta = r_0 \text{ against } \mathrm{H}_1: R_0\beta \neq r_0$$

where $R_0$ is now a $q \times k$ matrix of constants ($q \leq k$) and $r_0$ is a $q \times 1$ of constants. The number $q$ of rows of $R_0$ is the number of (linear) restrictions

[43] Formally, this comes from the fact that if $t \sim t(m)$, then as $m \to \infty$, $t \xrightarrow{d} N(0,1)$. Informally, for $m$ large, $t(m) \approx N(0,1)$.

which are jointly tested.

### 4.4.1. Exact hypothesis testing: the $F$-test

- We first consider an exact in finite sample test, i.e., a test which is valid when the Gauss-Markov assumptions E.1 – E.4 as well as the normality assumption E.5 hold, which is the case if the seminal assumptions MLR.1 – MLR.5 as well as the normality assumption MLR.6 hold.

- From Property 4, under assumptions E.1 – E.5, and thus also under the seminal assumptions MLR.1 – MLR.6, we have:

$$\hat{\beta}|X \sim N\left(\beta, \sigma^2(X'X)^{-1}\right)$$

which implies that, for any multiple linear combination $R_0\beta - r_0$ of $\beta$, we have:

$$(R_0\hat{\beta} - r_0)|X \sim N(R_0\beta - r_0, R_0V(\hat{\beta}|X)R_0')$$

where $V(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$. If the value of $\beta$ is such that $R_0\beta = r_0$, i.e., if $H_0$ is true, by standard property of the multivariate normal distribution[44], conditional on $X$, we have:

$$\begin{aligned}\hat{\chi}_0^2 &= (R_0\hat{\beta} - r_0)' \left[R_0V(\hat{\beta}|X)R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \sim \chi^2(q) \end{aligned} \tag{37}$$

where $\chi^2(q)$ denotes the chi-squared distribution[45] with $q$ degrees of freedom. On the other hand, if the value of $\beta$ is such that $R_0\beta \neq r_0$, i.e., if $H_0$ is false, again by standard property of the multivariate normal distribution[46], conditional on $X$, we have:

$$\begin{aligned}\hat{\chi}_0^2 &= (R_0\hat{\beta} - r_0)' \left[R_0V(\hat{\beta}|X)R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \sim \chi^2(\delta^*, q) \end{aligned} \tag{38}$$

where $\chi^2(\delta^*, q)$ denotes the non-central chi-squared distribution with $q$ degrees of freedom, and non-centrality parameter equal to $\delta^* = (R_0\beta - r_0)' \left[R_0V(\hat{\beta}|X)R_0'\right]^{-1} (R_0\beta - r_0)$.

- As previously, under the same assumptions, conditional on $X$, we have that $\hat{v} = \frac{(n-k)\hat{s}^2}{\sigma^2} \sim \chi^2(n-k)$ and that $\hat{\chi}_0^2$ and $\hat{v}$ are independent[47], so that if the value of $\beta$ is such that $R_0\beta = r_0$, i.e., if $H_0$ is true, from the definition of the

[44] Let $X$ be a $q \times 1$ random vector. If $X \sim N(0, \Sigma)$, then $X'\Sigma^{-1}X \sim \chi^2(q)$. This carries over conditional distributions.

[45] See Wooldridge (2016), Appendix B-5d.

[46] Let again $X$ be a $q \times 1$ random vector. If $X \sim N(m, \Sigma)$, then $X'\Sigma^{-1}X \sim \chi^2(\delta, q)$, where $\delta = m'\Sigma^{-1}m$. This carries over conditional distributions.

[47] As previously, the independence of $\hat{\chi}_0^2$ and $\hat{v}$ (conditional on X) follows from the independence of $\hat{\beta}$ and $\hat{s}^2$ (conditional on $X$).

Fisher[48] distribution, still conditional on $X$, we have:

$$\begin{aligned}\hat{F}_0 &= \frac{\frac{\hat{\chi}_0^2}{q}}{\frac{\hat{v}}{n-k}} = \frac{\frac{1}{q\sigma^2}(R_0\hat{\beta}-r_0)'\left[R_0(X'X)^{-1}R_0'\right]^{-1}(R_0\hat{\beta}-r_0)}{\frac{\hat{s}^2}{\sigma^2}} \\ &= \frac{1}{q\hat{s}^2}(R_0\hat{\beta}-r_0)'\left[R_0(X'X)^{-1}R_0'\right]^{-1}(R_0\hat{\beta}-r_0) \\ &= \frac{1}{q}(R_0\hat{\beta}-r_0)'\left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1}(R_0\hat{\beta}-r_0) \sim F(q,n-k) \qquad (39)\end{aligned}$$

where $F(q,n-k)$ denotes the Fisher distribution with $q$ and $n-k$ degrees of freedom[49] and $\hat{V}(\hat{\beta}) = \hat{s}^2(X'X)^{-1}$. On the other hand, if the value of $\beta$ is such that $R_0\beta \neq r_0$, i.e., if $\mathrm{H}_0$ is false, from the definition of the non-central Fisher[50] distribution, still conditional on $X$, we similarly have:

$$\begin{aligned}\hat{F}_0 &= \frac{1}{q\hat{s}^2}(R_0\hat{\beta}-r_0)'\left[R_0(X'X)^{-1}R_0'\right]^{-1}(R_0\hat{\beta}-r_0) \\ &= \frac{1}{q}(R_0\hat{\beta}-r_0)'\left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1}(R_0\hat{\beta}-r_0) \sim F(\delta^*,q,n-k) \qquad (40)\end{aligned}$$

where $F(\delta^*,q,n-k)$ denotes the non-central Fisher distribution with $q$ and $n-k$ degrees of freedom, and non-centrality parameter equal to $\delta^* = (R_0\beta - r_0)'\left[R_0V(\hat{\beta}|X)R_0'\right]^{-1}(R_0\beta - r_0)$.

In words, if the unknown variance $\sigma^2$ appearing in the variance-covariance $V(\hat{\beta}|X)$ of statistics (37) and (38) is replaced by its unbiased estimator $\hat{s}^2$, so that the variance-covariance $V(\hat{\beta}|X)$ is replaced by its estimator $\hat{V}(\hat{\beta})$, then the distribution of (37) and (38), after division by $q$, switches from chi-squared to Fisher.

- The distributional results (39) and (40) imply that the statistic $\hat{F}_0$ (which can not be negative) will yield systematically higher values when the value of $\beta$ is such that $R_0\beta \neq r_0$, i.e., when $\mathrm{H}_0$ is false, than when the value of $\beta$ is such that $R_0\beta = r_0$, i.e., when $\mathrm{H}_0$ is true[51]. Also note that the distributional result (39) does not hold only conditional on $X$: as the conditional distribution of $\hat{F}_0$ actually does not depend on $X$, it also holds unconditionally. This behavior makes $\hat{F}_0$ a natural test statistic for testing $\mathrm{H}_0: R_0\beta = r_0$ against $\mathrm{H}_1: R_0\beta \neq r_0$.

- A test at (significance) level $\alpha$ of $\mathrm{H}_0: R_0\beta = r_0$ against $\mathrm{H}_1: R_0\beta \neq r_0$ is given

[48] If $v_1 \sim \chi^2(m_1)$, $v_2 \sim \chi^2(m_2)$, and $v_1$ and $v_2$ are independent, then $F = \frac{\frac{v_1}{m_1}}{\frac{v_2}{m_2}} \sim F(m_1,m_2)$. See Wooldridge (2016), Appendix B-5f. This carries over conditional distributions.

[49] See Wooldridge (2016), Appendix B-5f.

[50] If $v_1 \sim \chi^2(\delta,m_1)$, $v_2 \sim \chi^2(m_2)$, and $v_1$ and $v_2$ are independent, then $F = \frac{\frac{v_1}{m_1}}{\frac{v_2}{m_2}} \sim F(\delta,m_1,m_2)$. This carries over conditional distributions.

[51] For a graphical evidence, see Lejeune (2011), p. 130.

by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } \hat{F}_0 > F_{q,n-k;1-\alpha} \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases} \tag{41}$$

where the critical value $F_{q,n-k;1-\alpha}$ is the quantile of order $1-\alpha$ of the $F(q, n-k)$ Fisher distribution, i.e., the value such that $\mathbb{P}(F \leq F_{q,n-k;1-\alpha}) = 1-\alpha$, where $F \sim F(q, n-k)$. The $p$-value of this test, for a value $\hat{F}_0^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{F}_0^*} = \mathbb{P}(F > \hat{F}_0^*)\,, \quad \text{where } F \sim F(q, n-k)$$

- Remarks :
  - The above $F$-test contains as special cases – and is fully equivalent to – the two-sided $t$-tests about a single parameter and about a single linear combination of parameters.
  - It may be shown that the $\hat{F}_0$ test statistic can also be written as[52] :

$$\hat{F}_0 = \frac{(\mathrm{SSR}_r - \mathrm{SSR}_{ur})/q}{\mathrm{SSR}_{ur}/(n-k)}$$

    where $\mathrm{SSR}_{ur}$ is the residual sum of squares[53] of the unrestricted model, i.e., the SSR of the model estimated without any restriction imposed, and $\mathrm{SSR}_r$ is the residual sum of squares of the restricted model, i.e., the SSR of the model estimated by restricted least-squares with the restriction $R_0\beta = r_0$ imposed. For examples of how $\mathrm{SSR}_r$ may easily be computed in some important cases (in particular when considering exclusion restrictions), see Wooldridge (2016), Section 4-5.

### 4.4.2. Hypothesis testing without the normality assumption

- Hereafter, we show that the exact in finite sample $F$-test of $\mathrm{H}_0$ : $R_0\beta = r_0$ against $\mathrm{H}_1$ : $R_0\beta \neq r_0$ derived under assumptions E.1 – E.5, which is thus likewise exact in finite sample under the seminal assumptions MLR.1 – MLR.6, remain asymptotically valid – i.e., approximately valid for $n$ sufficiently large – under assumptions MLR.1 – MLR.5, i.e., without having to rely on the normality assumption MLR.6.

- From Property 7, under assumptions MLR.1 – MLR.5, we have the limiting distributional result :

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 A^{-1}), \quad \text{where } A = E(X_i' X_i)$$

[52] For a hint of the proof, see Hayashi (2000) p. 42-43 and p. 74-75. Wooldridge (2016) exclusively uses this simpler form of the $F$-test (see Section 4-5). Note that the $F$-test can further be expressed in terms of $R$-squared. Be aware : in Wooldridge (2016), the degrees of freedom appearing in the denominator of $F$ is $n-(k+1)$ rather than $(n-k)$ because he considers a model with $k$ explanatory variables + an intercept, while here the intercept is included in the set of the explanatory variables.

[53] also called the sum of squared residuals.

or, in terms of approximation for $n$ sufficiently large :

$$\hat{\beta} \approx N(\beta, \sigma^2 A^{-1}/n)$$

which implies that, for any multiple linear combination $R_0\beta - r_0$ of $\beta$, we have :

$$R_0\hat{\beta} - r_0 \approx N(R_0\beta - r_0, R_0 Avar(\hat{\beta})R_0')$$

where $Avar(\hat{\beta}) = \sigma^2 A^{-1}/n$. If the value of $\beta$ is such that $R_0\beta = r_0$, i.e., if $H_0$ is true, by standard property of the multivariate normal distribution[54], we have :

$$\begin{aligned} \hat{\chi}^2_{as0} &= (R_0\hat{\beta} - r_0)' \left[R_0 Avar(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0(A^{-1}/n)R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(q) \end{aligned} \tag{42}$$

On the other hand, if the value of $\beta$ is such that $R_0\beta \neq r_0$, i.e., if $H_0$ is false, again by standard property of the multivariate normal distribution[55], we have :

$$\begin{aligned} \hat{\chi}^2_{as0} &= (R_0\hat{\beta} - r_0)' \left[R_0 Avar(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= \frac{1}{\sigma^2}(R_0\hat{\beta} - r_0)' \left[R_0^{-1}(A^{-1}/n)R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(\delta^*, q) \end{aligned} \tag{43}$$

where $\delta^* = (R_0\beta - r_0)' \left[R_0 Avar(\hat{\beta})R_0'\right]^{-1} (R_0\beta - r_0)$.

- It may be shown that, from an asymptotic point of view, we can replace the unknown parameters $\sigma^2$ and $A$ appearing in the asymptotic variance $Avar(\hat{\beta})$ of statistics (42) and (43) by their consistent estimator $\hat{s}^2$ and $X'X/n$ – which means that $Avar(\hat{\beta})$ is replaced by its estimator $\hat{V}(\hat{\beta})$ – without affecting their approximate distribution[56], so that if the value of $\beta$ is such that $R_0\beta = r_0$, i.e., if $H_0$ is true, we also have :

$$\begin{aligned} \hat{W}_0 &= \frac{1}{\hat{s}^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= (R_0\hat{\beta} - r_0)' \left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(q) \end{aligned} \tag{44}$$

where $\hat{V}(\hat{\beta}) = \hat{s}^2(X'X)^{-1}$, while if the value of $\beta$ is such that $R_0\beta \neq r_0$, i.e., if $H_0$ is false, we similarly have :

$$\begin{aligned} \hat{W}_0 &= \frac{1}{\hat{s}^2}(R_0\hat{\beta} - r_0)' \left[R_0(X'X)^{-1}R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \\ &= (R_0\hat{\beta} - r_0)' \left[R_0\hat{V}(\hat{\beta})R_0'\right]^{-1} (R_0\hat{\beta} - r_0) \approx \chi^2(\delta^*, q) \end{aligned} \tag{45}$$

where $\delta^* = (R_0\beta - r_0)' \left[R_0 Avar(\hat{\beta})R_0'\right]^{-1} (R_0\beta - r_0)$.

---

[54] As a reminder, if $X \sim N(0, \Sigma)$, then $X'\Sigma^{-1}X \sim \chi^2(q)$.

[55] As a reminder, if $X \sim N(m, \Sigma)$, then $X'\Sigma^{-1}X \sim \chi^2(\delta, q)$, where $\delta = m'\Sigma^{-1}m$.

[56] As previously outlined, properly showing this requires advanced asymptotic theory as developped in Wooldridge (2010), Chapter 3. See also Hayashi (2010), Chapter 2.

- The distributional results (44) and (45) imply that the statistic $\hat{W}_0$ (which can not be negative) will yield systematically higher values when the value of $\beta$ is such that $R_0\beta \neq r_0$, i.e., when $\mathrm{H}_0$ is false, than when the value of $\beta$ is such that $R_0\beta = r_0$, i.e., when $\mathrm{H}_0$ is true. This behavior makes $\hat{W}_0$ a natural test statistic for testing $\mathrm{H}_0 : \ R_0\beta = r_0$ against $\mathrm{H}_1 : \ R_0\beta \neq r_0$.

- A test at (significance) level $\alpha$ of $\mathrm{H}_0 : \ R_0\beta = r_0$ against $\mathrm{H}_1 : \ R_0\beta \neq r_0$ is given by the decision rule :

$$\begin{cases} \text{- Reject } \mathrm{H}_0 \text{ if } \hat{W}_0 > \chi^2_{q;1-\alpha} \\ \text{- Do not reject } \mathrm{H}_0 \text{ otherwise} \end{cases} \tag{46}$$

  where the critical value $\chi^2_{q;1-\alpha}$ is the quantile of order $1-\alpha$ of the $\chi^2(q)$ chi-squared distribution, i.e., the value such that $\mathbb{P}(v \leq \chi^2_{q;1-\alpha}) = 1-\alpha$, where $v \sim \chi^2(q)$. The $p$-value of this test, for a value $\hat{W}_0^*$ of the test statistic obtained in a particular sample, is given by :

$$p_{\hat{W}_0^*} = \mathbb{P}(v > \hat{W}_0^*) \,, \quad \text{where } v \sim \chi^2(q)$$

  This test is known as a Wald test[57].

- As $n \to \infty$, the decision rule (41) of the exact in finite sample $F$-test of $\mathrm{H}_0 : \ R_0\beta = r_0$ against $\mathrm{H}_1 : \ R_0\beta \neq r_0$ obtained previously becomes the same as the decision rule (46) of this Wald test : the two tests are asymptotically equivalent. As a matter of fact, on the one hand, the $\hat{F}_0$ test statistic is nothing but the Wald test statistic $\hat{W}_0$ divided by $q$ :

$$\hat{F}_0 = \frac{\hat{W}_0}{q}$$

  On the other hand, as $n \to \infty$, we have[58] :

$$F_{q,n-k;1-\alpha} \simeq \frac{\chi^2_{q;1-\alpha}}{q} \quad \Leftrightarrow \quad qF_{q,n-k;1-\alpha} \simeq \chi^2_{q;1-\alpha}$$

  where $F_{q,n-k;1-\alpha}$ and $\chi^2_{q;1-\alpha}$ are the quantiles of order $1-\alpha$ of respectively the $F(q, n-k)$ and the $\chi^2(q)$ distribution. As a result, the exact in finite sample $F$-test of $\mathrm{H}_0 : \ R_0\beta = r_0$ against $\mathrm{H}_1 : \ R_0\beta \neq r_0$ derived under assumptions E.1 – E.5, which is thus likewise exact in finite sample under the seminal assumptions MLR.1 – MLR.6, remain asymptotically valid – i.e., approximately valid for $n$ sufficiently large – under assumptions MLR.1 – MLR.5, i.e., without having to rely on the normality assumption MLR.6.

[57] See Wooldridge (2016), Appendix E-4.

[58] Formally, this comes from the fact that if $F \sim F(m_1, m_2)$, then as $m_2 \to \infty$, $m_1 F \xrightarrow{d} \chi^2(m_1)$. Informally, for $m_2$ large, $m_1 F(m_1, m_2) \approx \chi^2(m_1)$, or equivalently $F(m_1, m_2) \approx \frac{\chi^2(m_1)}{m_1}$.

# Appendix : properties of conditional expectation and conditional variance

- This appendix contains a brief summary of the properties of conditional expectation, and further conditional variance.

- The conditional expectation, also called the conditional mean, of $y$ given $x$ is by definition given, for $y$ a discrete random variable, by :

$$E(y|x) = \sum_y yf(y|x)$$

and for $y$ a continuous random variable, by :

$$E(y|x) = \int_{-\infty}^{\infty} yf(y|x)dy$$

where $f(y|x)$ denotes the conditional density of $y$ given $x$. Further, the conditional variance of $y$ given $x$ is defined as :

$$Var(y|x) = E\left[(y - E(y|x))^2|x\right] = E(y^2|x) - [E(y|x)]^2$$

- The main properties of conditional mean and conditional variance are the following[59] :

  - P.1 For any function $c(x)$, we have :

$$E\left(c(x)|x\right) = c(x)$$

and further :

$$Var\left(c(x)|x\right) = 0$$

  - P.2 For any function $a(x)$ and $b(x)$, we have :

$$E\left(a(x) + b(x)y|x\right) = a(x) + b(x)E(y|x)$$

and further :

$$Var\left(a(x) + b(x)y|x\right) = b(x)^2 Var(y|x)$$

  - P.3 If $x$ and $y$ are independent, then :

$$E(y|x) = E(y)$$

and also :

$$Var(y|x) = Var(y)$$

  - P.4 we have :

$$E(y) = E\left[E(y|x)\right]$$

[59] For more details, see Wooldridge (2016), Appendix B4-f and B4-g.

and more generally :

$$E(y|x) = E\left[E(y|x,z)|x\right]$$

This is the law of iterated expectations.

– P.5 If $E(y|x) = E(y)$, then :

$$Cov(x,y) = 0$$

• For $Y$ a $n \times 1$ random vector :

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

and $X$ a random vector or matrix, the conditional mean and the conditional variance of $Y$ given $X$ are by definition equal to :

$$E(Y|X) = E\left[\begin{array}{c|c} y_1 & \\ y_2 & \\ \vdots & X \\ y_n & \end{array}\right] = \begin{bmatrix} E(y_1|X) \\ E(y_2|X) \\ \vdots \\ E(y_n|X) \end{bmatrix}$$

and

$$\begin{aligned} V(Y|X) &= E\left[(Y - E(Y|X))(Y - E(Y|X))'|X\right] \\ &= E(YY'|X) - E(Y|X)E(Y|X)' \\ &= \begin{bmatrix} Var(y_1|X) & Cov(y_1,y_2|X) & \cdots & Cov(y_1,y_n|X) \\ Cov(y_2,y_1|X) & Var(y_2|X) & \cdots & Cov(y_2,y_n|X) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(y_n,y_1|X) & Cov(y_n,y_2|X) & \cdots & Var(y_n|X) \end{bmatrix} \end{aligned}$$

and the following similar properties hold[60] :

– P.1' For any $k \times 1$ vector function $C(X)$, we have :

$$E\left(C(X))|X\right) = C(X)$$

and further :

$$V\left(C(X))|X\right) = 0$$

– P.2' For any $k \times 1$ vector function $A(X)$ and $k \times n$ matrix function $B(X)$, we have :

$$E\left(A(X) + B(X)Y|X\right) = A(X) + B(X)E(Y|X)$$

[60] Note that in the expression of $E(Y|X)$ and its properties, $Y$ may also be a matrix. In the expression of $V(Y|X)$ and its properties, $Y$ may however only be a (column) vector. Note further that, by definition, in detailed form, $Cov(y_1,y_2|X) = E\left[(y_1 - E(y_1|X))(y_2 - E(y_2|X))|X\right] = E(y_1y_2|X) - E(y_1|X)E(y_2|X)$.

and further :

$$V\left(A(X)+B(X)Y|X\right)=B(X)V(Y|X)B(X)'$$

– P.3' If $X$ and $Y$ are independent, then :

$$E(Y|X)=E(Y)$$

and also :

$$V(Y|X)=V(Y)$$

– P.4' we have :

$$E(Y)=E\left[E(Y|X)\right]$$

and more generally :

$$E(Y|X)=E\left[E(Y|X,Z)|X\right]$$

where $Z$ is any random vector or matrix.

– P.5' If $E(Y|X)=E(Y)$, then :

$$Cov(x_{jk},y_i)=0, \text{ for all } i,\ j \text{ and } k$$

# Reference


Goldberger A.S. (1991), *A Course in Econometrics*, Harvard University Press.

Hayashi F. (2000), *Econometrics*, Princeton University Press.

Lejeune B. (2011), *Econométrie I*, Lecture notes (in french, latest version downloadable at www.microeco.ulg.ac.be/lejeune), HEC-Université de Liège.

Wooldridge J.M. (2010), *Econometric Analysis of Cross-Section and Panel Data*, Second Edition, MIT Press.

Wooldridge J.M. (2016), *Introductory Econometrics : A Modern Approach*, 6th Edition, Cengage Learning.