Econometrics and Statistics

Jean-Paul Renne

2023-02-17

Contents

Econometrics and statistics

This course covers various econometric topics, including linear regression models, discrete-choice models, and an introduction to time series analysis. It provides examples or simulations based on R codes. This course has been developed by Jean-Paul Renne.

The R codes use various packages that can be obtained from CRAN. Several pieces of code also involve procedures and data from a companion package (`AEC`). This `AEC` package is available on GitHub. To install it, one need to employ the `devtools` library:

```
install.packages("devtools") # This library allows to use "install_github".
library(devtools)
install_github("jrenne/AEC")
library(AEC)
```

Useful (R) links:

- Download R:
  - R software: https://cran.r-project.org (the basic R software)
  - RStudio: https://www.rstudio.com (a convenient R editor)
- Tutorials:
  - Rstudio: https://dss.princeton.edu/training/RStudio101.pdf (by Oscar Torres-Reyna)
  - R: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf (by Emmanuel Paradis)
  - My own tutorial: https://jrenne.shinyapps.io/Rtuto_publiShiny/

Chapter 1

Basic statistical results

1.1 Cumulative and probability density funtions (c.d.f. and p.d.f.)

Definition 1.1 (Cumulative distribution function (c.d.f.))**.** The random variable (r.v.) X admits the cumulative distribution function F if, for all a:

$$F(a) = \mathbb{P}(X \leq a).$$

Definition 1.2 (Probability distribution function (p.d.f.))**.** A continuous random variable X admits the probability density function f if, for all a and b such that $a < b$:

$$\mathbb{P}(a < X \leq b) = \int_a^b f(x)dx,$$

where $f(x) \geq 0$ for all x.

In particular, we have:

$$f(x) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}(x < X \leq x + \varepsilon)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{F(x + \varepsilon) - F(x)}{\varepsilon}. \tag{1.1}$$

and

$$F(a) = \int_{-\infty}^a f(x)dx.$$

This web interface illustrates the link between p.d.f. and c.d.f. Nonparametric estimates of p.d.f. are obtained by kernel-based methods (see this extra material).

Definition 1.3 (Joint cumulative distribution function (c.d.f.))**.** The random variables X and Y admit the joint cumulative distribution function F_{XY} if, for all a and b:

$$F_{XY}(a, b) = \mathbb{P}(X \leq a, Y \leq b).$$

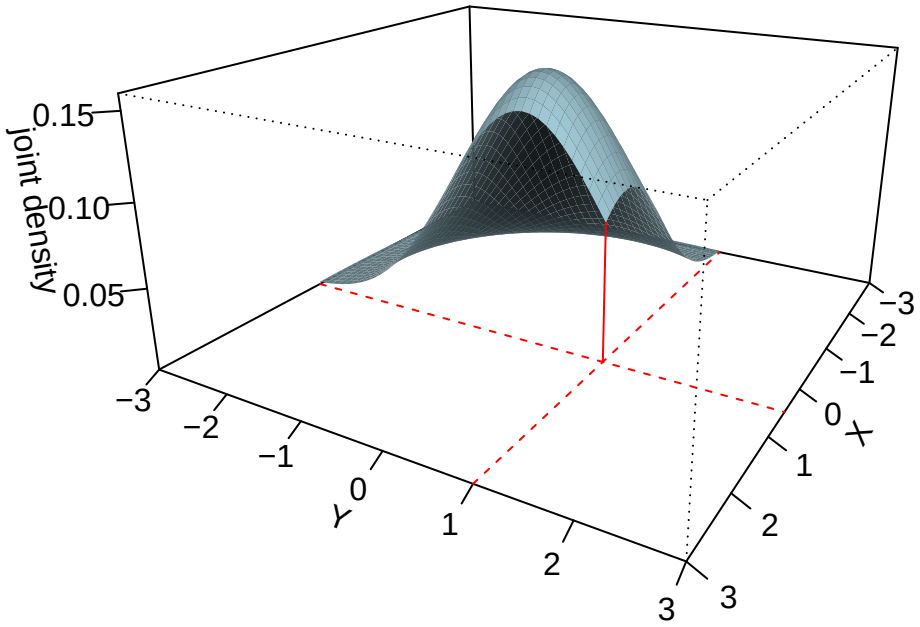


Figure 1.1: The volume between the horizontal plane ($z = 0$) and the surface is equal to $F_{XY}(0.5, 1) = \mathbb{P}(X < 0.5, Y < 1)$.

Definition 1.4 (Joint probability density function (p.d.f.))**.** The continuous random variables X and Y admit the joint p.d.f. f_{XY}, where $f_{XY}(x, y) \geq 0$ for all x and y, if:

$$\mathbb{P}(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{XY}(x, y)dxdy, \quad \forall a \leq b,\ c \leq d.$$

In particular, we have:

$$\begin{aligned} & f_{XY}(x, y) \\ = & \lim_{\varepsilon \to 0} \frac{\mathbb{P}(x < X \leq x + \varepsilon, y < Y \leq y + \varepsilon)}{\varepsilon^2} \\ = & \lim_{\varepsilon \to 0} \frac{F_{XY}(x + \varepsilon, y + \varepsilon) - F_{XY}(x, y + \varepsilon) - F_{XY}(x + \varepsilon, y) + F_{XY}(x, y)}{\varepsilon^2}. \end{aligned}$$

Definition 1.5 (Conditional probability distribution function)**.** If X and Y are continuous r.v., then the distribution of X conditional on $Y = y$, which we denote by $f_{X|Y}(x, y)$, satisfies:

$$f_{X|Y}(x, y) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}(x < X \leq x + \varepsilon | Y = y)}{\varepsilon}.$$

Proposition 1.1 (Conditional probability distribution function)**.** *We have*

$$f_{X|Y}(x, y) = \frac{f_{XY}(x, y)}{f_Y(y)}.$$

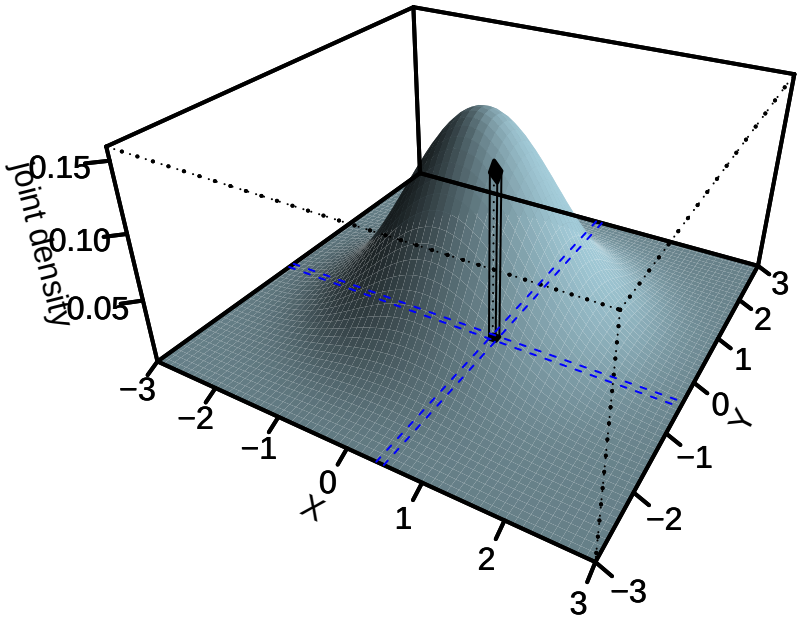


Figure 1.2: Assume that the basis of the black column is defined by those points whose x-coordinates are between x and $x+\varepsilon$ and y-coordinates are between y and $y+\varepsilon$. Then the volume of the black column is equal to $\mathbb{P}(x < X \leq x+\varepsilon, y < Y \leq y+\varepsilon)$, which is approximately equal to $f_{XY}(x,y)\varepsilon^2$ if ε is small.

Proof. We have:

$$\begin{aligned} f_{X|Y}(x,y) &= \lim_{\varepsilon\to 0} \frac{\mathbb{P}(x < X \leq x+\varepsilon | Y = y)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\mathbb{P}(x < X \leq x+\varepsilon | y < Y \leq y+\varepsilon) \\ &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\frac{\mathbb{P}(x < X \leq x+\varepsilon, y < Y \leq y+\varepsilon)}{\mathbb{P}(y < Y \leq y+\varepsilon)} \\ &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\frac{\varepsilon^2 f_{XY}(x,y)}{\varepsilon f_Y(y)}. \end{aligned}$$

□

Definition 1.6 (Independent random variables)**.** Consider two r.v., X and Y, with respective c.d.f. F_X and F_Y, and respective p.d.f. f_X and f_Y.

These random variables are independent if and only if (iff) the joint c.d.f. of X and Y (see Def. 1.3) is given by:

$$F_{XY}(x,y) = F_X(x) \times F_Y(y),$$

or, equivalently, iff the joint p.d.f. of (X,Y) (see Def. 1.4) is given by:

$$f_{XY}(x,y) = f_X(x) \times f_Y(y).$$

We have the following:

1. If X and Y are independent, $f_{X|Y}(x,y) = f_X(x)$. This implies, in particular, that $\mathbb{E}(g(X)|Y) = \mathbb{E}(g(X))$, where g is any function.
2. If X and Y are independent, then $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$ and $\mathbb{C}ov(g(X), h(Y)) = 0$, where g and h are any functions.

It is important to note that the absence of correlation between two variables is not a sufficient condition to have independence. Consider for instance the case where $X = Y^2$, with $Y \sim \mathcal{N}(0,1)$. In this case, we have $\mathbb{C}ov(X,Y) = 0$, but X and Y are not independent. Indeed, we have for instance $\mathbb{E}(Y^2 \times X) = 3$, which is not equal to $\mathbb{E}(Y^2) \times \mathbb{E}(X) = 1$. (If X and Y were independent, we should have $\mathbb{E}(Y^2 \times X) = \mathbb{E}(Y^2) \times \mathbb{E}(X)$ according to point 2 above.)

1.2 Law of iterated expectations

Proposition 1.2 (Law of iterated expectations)**.** *If X and Y are two random variables and if $\mathbb{E}(|X|) < \infty$, we have:*

$$\boxed{\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)).}$$

Proof. (in the case where the p.d.f. of (X,Y) exists) Let us denote by f_X, f_Y and f_{XY} the probability distribution functions (p.d.f.) of X, Y and (X,Y), respectively. We have:

$$f_X(x) = \int f_{XY}(x,y)dy.$$

Besides, we have (Bayes equality, Prop. 1.1):

$$f_{XY}(x,y) = f_{X|Y}(x,y)f_Y(y).$$

Therefore:

$$\begin{aligned}\mathbb{E}(X) &= \int x f_X(x)dx = \int x \underbrace{\int f_{XY}(x,y)dy}_{=f_X(x)} dx = \int\int x f_{X|Y}(x,y)f_Y(y)dydx \\ &= \int \underbrace{\left(\int x f_{X|Y}(x,y)dx\right)}_{\mathbb{E}[X|Y=y]} f_Y(y)dy = \mathbb{E}\left(\mathbb{E}[X|Y]\right).\end{aligned}$$

□

Example 1.1 (Mixture of Gaussian distributions)**.** By definition, X is drawn from a mixture of Gaussian distributions if:

$$X = B \times Y_1 + (1 - B) \times Y_2,$$

where B, Y_1 and Y_2 are three independent variables drawn as follows:

$$B \sim \text{Bernoulli}(p), \quad Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad \text{and} \quad Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

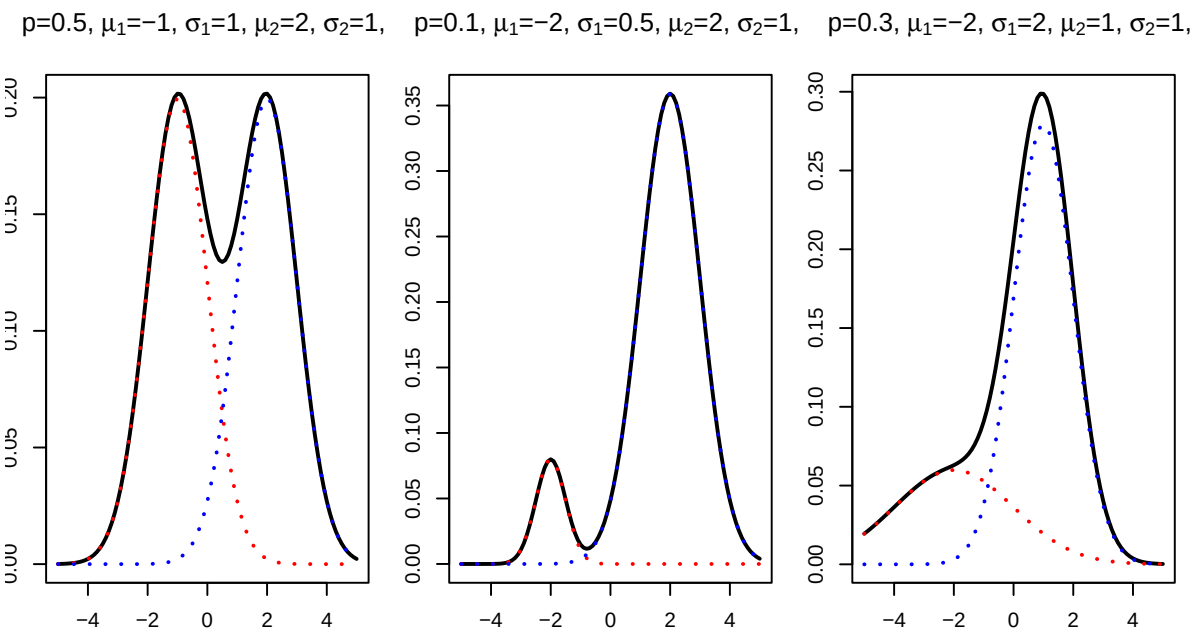


Figure 1.3: Example of pdfs of mixtures of Gaussian distribututions.

Figure 1.3 displays the pdfs associated with three different mixtures of Gaussian distributions. (This web-interface allows to produce the pdf associated for any other parameterization.)

The law of iterated expectations gives:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|B)) = \mathbb{E}(B\mu_1 + (1-B)\mu_2) = p\mu_1 + (1-p)\mu_2.$$

Example 1.2 (Buffon (1733)'s needles)**.** Suppose we have a floor made of parallel strips of wood, each the same width $[w = 1]$. We drop a needle, of length $1/2$, onto the floor. What is the probability that the needle crosses the grooves of the floor?

Let's define the random variable X by

$$X = \begin{cases} 1 & \text{if the needle crosses a line} \\ 0 & \text{otherwise.} \end{cases}$$

Conditionally on θ, it can be seen that we have $\mathbb{E}(X|\theta) = \cos(\theta)/2$ (see Figure 1.4).

It is reasonable to assume that θ is uniformly distributed on $[-\pi/2, \pi/2]$, therefore:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|\theta)) = \mathbb{E}(\cos(\theta)/2) = \int_{-\pi/2}^{\pi/2} \frac{1}{2}\cos(\theta)\left(\frac{d\theta}{\pi}\right) = \frac{1}{\pi}.$$

[This web-interface allows to simulate the present experiment (select Worksheet "Buffon's needles").]

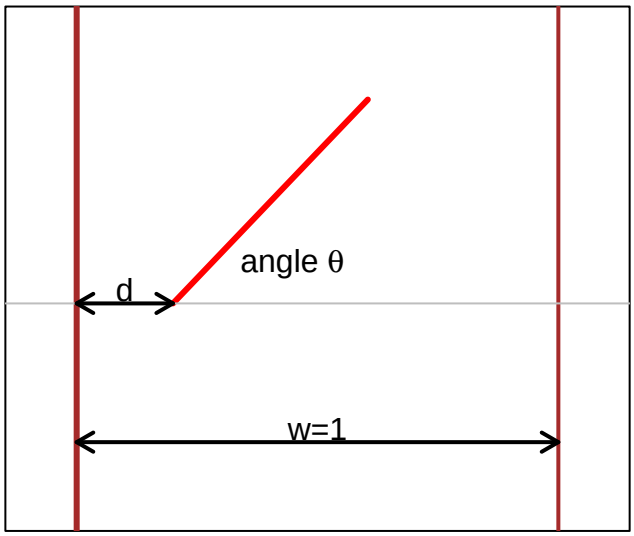


Figure 1.4: Schematic representation of the problem.

1.3 Law of total variance

Proposition 1.3 (Law of total variance)**.** *If X and Y are two random variables and if the variance of X is finite, we have:*

$$\boxed{\mathbb{V}ar(X) = \mathbb{E}(\mathbb{V}ar(X|Y)) + \mathbb{V}ar(\mathbb{E}(X|Y)).}$$

Proof. We have:

$$\begin{aligned}
\mathbb{V}ar(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \mathbb{E}(\mathbb{E}(X^2|Y)) - \mathbb{E}(X)^2 \\
&= \mathbb{E}(\mathbb{E}(X^2|Y) - \mathbb{E}(X|Y)^2) + \mathbb{E}(\mathbb{E}(X|Y)^2) - \mathbb{E}(X)^2 \\
&= \mathbb{E}(\underbrace{\mathbb{E}(X^2|Y) - \mathbb{E}(X|Y)^2}_{\mathbb{V}ar(X|Y)}) + \underbrace{\mathbb{E}(\mathbb{E}(X|Y)^2) - \mathbb{E}(\mathbb{E}(X|Y))^2}_{\mathbb{V}ar(\mathbb{E}(X|Y))}.
\end{aligned}$$

□

Example 1.3 (Mixture of Gaussian distributions (cont'd))**.** Consider the case of a mixture of Gaussian distributions (Example 1.1). We have:

$$\begin{aligned}
\mathbb{V}ar(X) &= \mathbb{E}(\mathbb{V}ar(X|B)) + \mathbb{V}ar(\mathbb{E}(X|B)) \\
&= p\sigma_1^2 + (1-p)\sigma_2^2 + p(1-p)(\mu_1 - \mu_2)^2.
\end{aligned}$$

1.4 About consistent estimators

The objective of econometrics is to estimate parameters out of data observations (samples). Examples of parameters of interest include, among many others: causal effect of a variable on

another, elasticities, parameters defining some distribution of interest, preference parameters (risk aversion)...

Except in degenerate cases, the estimates are different from the "true" (or *population*) value. A good estimator is expected to converge to the true value when the sample size increases. That is, we are interested in the *consistency* of the estimator.

Denote by $\widehat{\theta}_n$ the estimate of θ based on a sample of length n. We say that $\widehat{\theta}$ is a consistent estimator of θ if, for any $\varepsilon > 0$ (even if very small), the probability that $\widehat{\theta}_n$ is in $[\theta - \varepsilon, \theta + \varepsilon]$ goes to 1 when n goes to ∞. Formally:

$$\lim_{n \to +\infty} \mathbb{P}\left(\widehat{\theta}_n \in [\theta - \varepsilon, \theta + \varepsilon]\right) = 1.$$

That is, $\widehat{\theta}$ is a consistent estimator if $\widehat{\theta}_n$ *converges in probability* (Def. 9.16) to θ. Note that there exist different types of stochastic convergence.[1]

Example 1.4 (Example of non-convergent estimator)**.** Assume that $X_i \sim i.i.d.\text{Cauchy}$ with a location parameter of 1 and a scale parameter of 1 (Def. 9.14). The sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ does not converge in probability. This is because a Cauchy distribution has no mean; hence the law of large numbers (Theorem 2.1) does not apply.

```
N <- 5000
X <- rcauchy(N)
X.bar <- cumsum(X)/(1:N)
par(plt=c(.1,.95,.3,.85),mfrow=c(1,2))
plot(X,type="l",xlab="sample size (n)",
     ylab="",main=expression(X[n]),lwd=2)
plot(X.bar,type="l",xlab="sample size (n)",
     ylab="",main=expression(bar(X)[n]),lwd=2)
abline(h=0,lty=2,col="grey")
```

[1] Appendix 9.3.3 notably provides the definitions of the convergence in distribution (Def. 9.19) and the mean-square convergence (Def. 9.17).

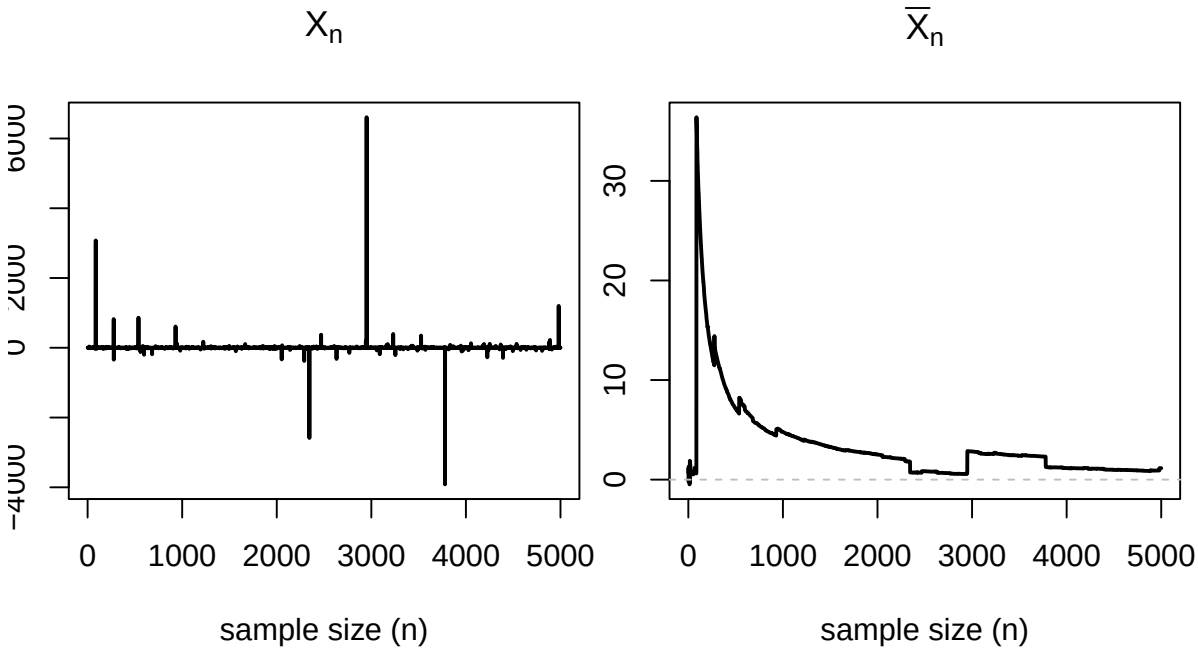


Figure 1.5: Simulation of $\bar{X}_n$ when $X_i \sim i.i.d.$Cauchy.

Chapter 2

Central Limit Theorem

The law of large numbers (LLN) and the central limit theorem (CLT) are two fundamental theorems of probability. The former states that the sample mean converges in probability to the population mean (when it exists). The latter states that the distribution of the sum (or average) of a large number of independent, identically distributed (i.i.d.) variables with finite variance is approximately normal, regardless of the underlying distribution (as long as it features a finite variance).

2.1 Law of large numbers

Definition 2.1 (Sample and Population)**.** In statistics, a *sample* refers to a finite set of observations drawn from a *population*.

Most of the time, it is necessary to use samples for research, because it is impractical to study the whole population. Typically, it is rare to observe the population mean, and we often use instead sample means (or averages). The sample average of $\{x_1, \dots, x_n\}$ is given by:

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This sample mean is an estimator of the true (population) mean $\mathbb{E}(x_i)$ (assuming the latter exists). According to the law of large numbers (Theorem 2.1), this estimator is *consistent*. (see Def. 9.16 for the definition of convergence in probability.)

Theorem 2.1 (Law of large numbers)**.** *The sample mean is a consistent estimator of the population mean. In other words, the sample mean converges in probability to the population mean.*

Proof. See Appendix 9.3.4. □

Example 2.1. Suppose one is interested in the average number of doctor consultations over 12 months, for Swiss people. One can get an estimate by computing the sample average of a large dataset, for instance using the Swiss Household Panel (SHP):

Figure 2.1 shows the empirical distribution of the number of doctor visits.

```
library(AEC) # to load the shp data-frame.
Nb.doct.visits <- shp$p19c15
Nb.doct.visits <- Nb.doct.visits[Nb.doct.visits>=0] # remove irrelevant observations
par(plt=c(.15,.95,.2,.95))
barplot(table(Nb.doct.visits),xlim=c(0,25))
```

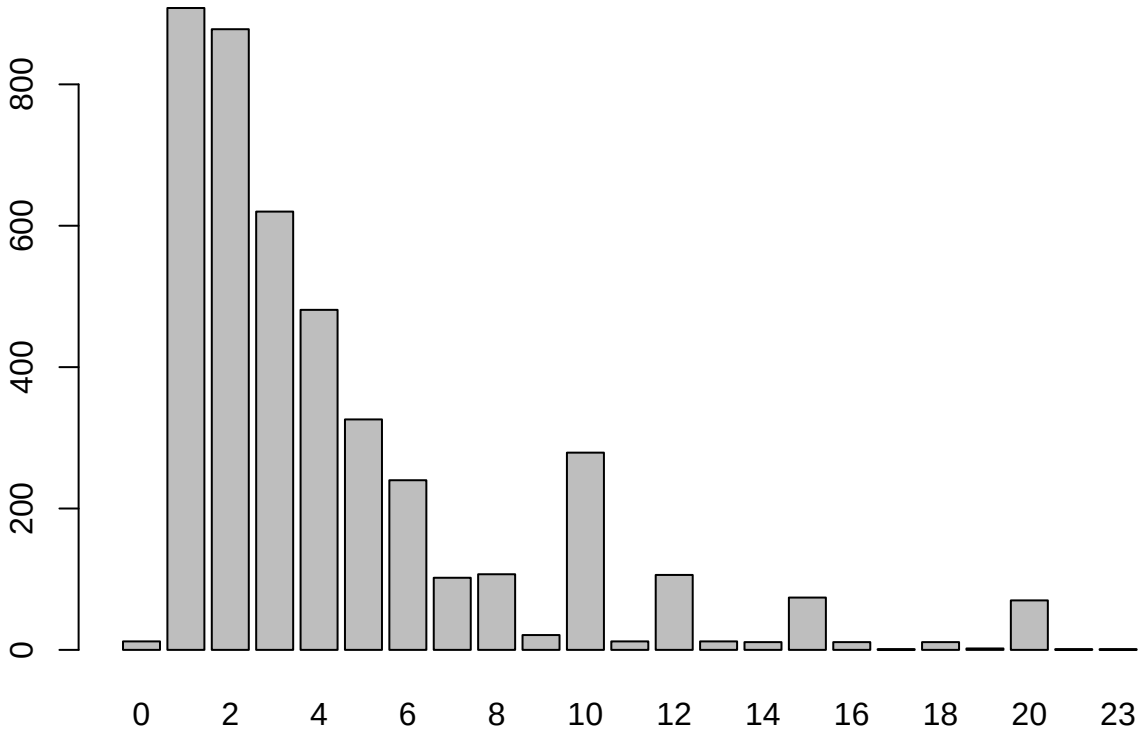


Figure 2.1: Distribution of the number of doctor visits. Source: Swiss Household Panel.

Let us compute the sample mean, variance, and standard deviation:

```
cbind(mean(Nb.doct.visits),var(Nb.doct.visits),sd(Nb.doct.visits))
```

```
##          [,1]     [,2]   [,3]
## [1,] 5.253816 79.50575 8.9166
```

How to know whether the sample mean is a good estimate of the true (population) average? In other words, how to measure the accuracy of the estimator?

A first answer is provided by the Chebychev inequality. Assume the x_i's are independently drawn from a distribution of mean μ and variance σ^2. We have:

$$\mathbb{E}(\bar{x}_n) = \mu \quad \text{and} \quad \mathbb{V}ar(\mu - \bar{x}_n) = \frac{1}{n}\sigma^2 \underset{n\to\infty}{\to} 0.$$

The Chebychev inequality (Proposition 9.11) states that, for any $\varepsilon > 0$ (even very small), we have:

$$\mathbb{P}(|\bar{x}_n - \mu| > \varepsilon) \leq \frac{\mathbb{V}ar(\mu - \bar{x}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \underset{n\to\infty}{\to} 0.$$

Consider for instance the estimation of the average number of doctor visits (Example 2.1). In this example, the sample length is $n = 4389$, we have $\bar{x}_n = 5.25$, and the variance σ^2 was approximately equal to 79.51. Therefore, taking $\varepsilon = 0.5$, we have that $\mathbb{P}(\bar{x}_n - 0.5 < \mu < \bar{x}_n + 0.5)$ is lower than $\frac{\sigma^2}{n\varepsilon^2} \approx 0.072$.

However, this only gives bounds for such probabilities. The central limit theorem provides richer information, as it gives the approximate distribution of the estimation error.

2.2 Central Limit Theorem (CLT)

Theorem 2.2 (Lindberg-Levy Central limit theorem, CLT)**.** *If x_n is an i.i.d. sequence of random variables with mean μ and (finite) variance, then:*

$$\boxed{\sqrt{n}(\bar{x}_n - \mu) \overset{d}{\to} \mathcal{N}(0, \sigma^2), \quad \textit{where} \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.}$$

[“$\overset{d}{\to}$” denotes the convergence in distribution (see Def. 9.19)]

Proof. See Appendix 9.3.4. □

According to the CLT, when n is large and whatever the distribution of the x_i's (as long as it features an average μ and a variance σ^2), the error between $\bar{x}_n$ and μ can be seen as a random variable that is normally distributed; it is of mean zero and its standard deviation is $\sigma/\sqrt{n}$. (We then say that the convergence rate is *in* $\sqrt{n}$.) The knowledge of the quantiles of the normal distribution can then be used to compute approximate confidence intervals for μ (based on the sample $\{x_1, \ldots, x_n\}$).

This web interface illustrates the CLT with simulations (select the “CLT” block).

The CL theorem was first postulated by the mathematician Abraham de Moivre. In an article published in 1733, he used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This finding was nearly forgotten until Pierre-Simon Laplace recalled it in his Théorie analytique des probabilités, published in 1812. Laplace expanded De Moivre's finding by approximating the binomial distribution with the normal distribution.

Exercise 2.1. Consider Buffon's experiment (Example 1.2). We have seen that π can be estimated by counting the fraction of times the needles cross the grooves of the floor.[1] Using the CLT, determine the (approximate) minimal number n of needles that one has to throw to get an estimate $\hat{\pi}$ $(= 1/\bar{X}_n)$ of π that is such that there is a 95% chance that $\pi \in [\hat{\pi} - 0.01, \hat{\pi} + 0.01]$?

Exercise 2.2. Assume you have a lot of time and a coin. How would you approximate the 100 percentiles of the $\mathcal{N}(0,1)$ distribution?

The CLT can be extended to the case where each x_i is a vector. (In that case, and if the dimension of x_i is m, then $\mathbb{V}ar(x_i)$ is a $m \times m$ matrix.) Here is the multivariate CLT:

Theorem 2.3 (Multivariate Central limit theorem, CLT)**.** *If $\mathbf{x}_n$ is an i.i.d. sequence of m-dimensional random variables with mean μ and variance Σ, where Σ is a positive definite matrix, then:*

$$\boxed{\sqrt{n}(\bar{\mathbf{x}}_n - \mu) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \quad \textit{where} \quad \bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.}$$

2.3 Comparison of sample means

As said at the beginning of this chapter, the CLT is a key statistical result that has numerous applications. An important one is the statistical comparision of sample means:

Consider two samples $\{m_1, \dots, m_{n_m}\}$ and $\{w_1, \dots, w_{n_w}\}$. Assume that the m_i's (respectively the w_i's) are i.i.d., with mean μ_m and variance σ_m^2 and of w_i's (resp. with mean μ_w and variance σ_w^2).

If n is large, according to the CLT, we approximately have:

$$\sqrt{n_m}(\bar{m} - \mu_m) \sim \mathcal{N}(0, \sigma_m^2) \quad \textit{and} \quad \sqrt{n_w}(\bar{w} - \mu_w) \sim \mathcal{N}(0, \sigma_w^2),$$

or

$$\bar{m} \sim \mathcal{N}\left(\mu_m, \frac{\sigma_m^2}{n_m}\right) \quad \textit{and} \quad \bar{w} \sim \mathcal{N}\left(\mu_w, \frac{\sigma_w^2}{n_w}\right).$$

if the m_i's and the w_i's are independent, then $\bar{m}$ and $\bar{w}$ also are, and we get:

$$\bar{m} - \bar{w} \sim \mathcal{N}\left(\mu_m - \mu_w, \sigma^2\right) \quad \textit{with} \quad \sigma^2 = \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}.$$

This can be used to test the null hypothesis $H_0 : \mu_m - \mu_w = 0$ (see Section 3). Indeed under this assumption, the probability that $\bar{m} - \bar{w}$ is in $[-1.96\sigma, 1.96\sigma]$ (respectively in $[-2.58\sigma, 2.58\sigma]$) is approximately 0.95 (respectively 0.99) [see Table 9.1].

For instance, if we find, using our sample, that $\bar{m} - \bar{w}$ is not in $[-2.58\sigma, 2.58\sigma]$, then we reject the null hypothesis at the 1% significance level.

[1] More precisely, the expectation of this fraction is equal to $1/\pi$.

Example 2.2 (Comparison of sample means)**.** Let us use the Swiss Household Panel (SHP) dataset to test the equality between men and women incomes, for Swiss people under the age of 35.

The next lines of codes construct two samples of incomes; one for men (`income.m`), and one for women (`income.w`).

```
library(AEC);library(logKDE)
dataset.m <- subset(shp,(sex19==1)&(age19<=35))
dataset.w <- subset(shp,(sex19==2)&(age19<=35))
income.m <- dataset.m$i19ptotg
income.w <- dataset.w$i19ptotg
income.m <- income.m[income.m>=0] # remove irrelevant data
income.w <- income.w[income.w>=0] # remove irrelevant data
```

Figure 2.2 displays the empirical densities of incomes for men and women. Vertical dashed lines indicate sample means:

```
par(plt=c(.1,.95,.2,.95)) # adjust margins
plot(logdensity(income.w),lwd=2,xlim=c(0,200000),main="",
     xlab="Yearly gross income (in CHF)",ylab="") # x and y labels
lines(logdensity(income.m),col="red",lwd=2)
abline(v=mean(income.m),col="red",lty=3,lwd=2)
abline(v=mean(income.w),col="black",lty=3,lwd=2)
legend("topright",c("men yearly income","women"),lty=1,lwd=2,col=c("red","black"))
```

Let us compute $\bar{m}$, $\bar{w}$, and σ:

```
m.bar <- mean(income.m)
w.bar <- mean(income.w)
sigma <- sqrt(var(income.m)/length(income.m) + var(income.w)/length(income.w))
print(c(m.bar,w.bar,sigma))
```

```
## [1] 53035.023 42708.227  2427.425
```

Under the "equality assumption", their would a probability of only 5% (respectively 1%) for $\bar{m}-\bar{w}$ not to be in [-4758,4758] (respectively in [-6263,6263]). Since $\bar{m}-\bar{w}$ is equal to 10327, we reject the null hypothesis at the 5%, and even 1%, significance levels.

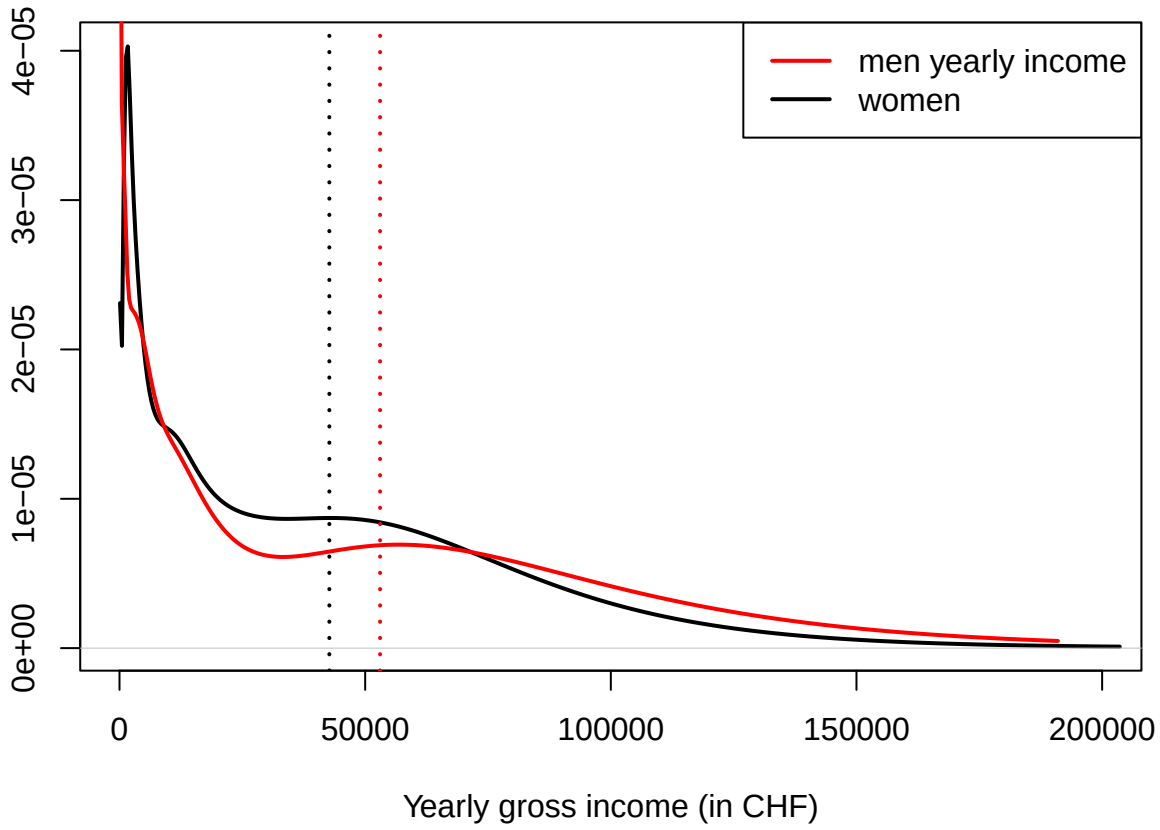


Figure 2.2: Distribution of yearly gross incomes for Swiss residents under the age of 35. Source: SHP. Vertical dashed lines indicates the sample mean.

Chapter 3

Statistical tests

We run a statistical test when we want to know whether some hypothesis about a vector of parameters θ —that is imperfectly observed— is consistent with data that are seen as random, and whose randomness depend on θ.

Typically, assume you observe a sample $\mathbf{x} = \{x_1, ..., x_n\}$ where the x_i's are i.i.d., of mean μ and variance σ^2 (these two parameters being unobseved). One may want to know whether $\mu = 0$ or, maybe, whether $\sigma = 1$.

The hypothesis the researcher wants to test is called the *null hypothesis*. It is often denoted by H_0. It is a conjecture about a given property of a population. Without loss of generality, it can be stated as:

$$H_0 : \{\theta \in \Theta\}.$$

It can also be defined through a function h (say):[1]

$$H_0 : h(\theta) = 0.$$

The *alternative hypothesis*, often denoted H_1 is then defined by $H_1 : \{\theta \in \Theta^c\}$.[2]

The ingredients of a statistical test are:

- a vector of (unknown) parameters (θ),
- a **test statistic**, that is a function of the sample components ($S(\mathbf{x})$, say), and
- a **critical region** (Ω, say), defined as a set of implausible values of S under H_0.

To implement a test statistic, we need to know the distribution of S under the null hypothesis H_0. Equipped, with such a distribution, we compute $S(\mathbf{x})$ and we look at its location; more specifically, we look whether it lies within the critical region Ω. The latter corresponds to "implausible" regions of this distribution (typically the tails). If $S \in \Omega$, then we reject the

[1] We can relate the previous and the next equation through $\Theta = \{\theta, \, s.t. \, h(\theta) = 0\}$.

[2] Note that researchers may want to find support for the null hypothesis H_0 as well as for the alternative hypothesis H_1.

null hypothesis. Looselly speaking; it amounts to saying: *if H_0 were true, it would have been unlikely to get such a large (or small) draw for S.*

Hence, there are two possible outcomes for a statisitcal test:

- H_0 is rejected if $S \in \Omega$;
- H_0 is not rejected if $S \notin \Omega$.

Except in extreme cases, there is always a non-zero probability to reject H_0 while it is true (**Type I error**, or **false positive**), or to fail to reject it while it is false (**Type II error**, or **false negative**).

This vocabulary is widely used. For instance, the notions of false positive and false negative are used in the context of Early Warning Signals (see Example 3.1).

Example 3.1 (Early Warning Signals)**.** These are approaches that aim to detect the occurrence of crises in advance. See, e.g., ECB (2014), for applications to financial crises.

To implement such approaches, researchers look for signals/indices forecasting crises. Suppose one index (W, say) tends to be large before financial crises; one may define an EWS by predicting a crisis when $W > a$, where a is a given threshold. It is easliy seen that the lower (respectively higeher) a, the larger the fraction of FP (resp. of FN).

3.1 Size and power of a test

Definition 3.1 (Size and Power of a test)**.** For a given test,

- the probability of type-I errors, denoted by α, is called the **size**, or **significance level**, of the test,
- the **power** of a test is equal to $1 - \beta$, where β is the probability of type-II errors.

Formally, the previous definitions can be written as follows:

$$\alpha = \mathbb{P}(S \in \Omega | H_0) \quad \text{(Proba. of a false positive)} \tag{3.1}$$
$$\beta = \mathbb{P}(S \notin \Omega | H_1) \quad \text{(Proba. of a false negative).} \tag{3.2}$$

The power is the probability that the test will lead to the rejection of the null hypothesis if the latter is false. Therefore, for a given size, we prefer tests with high power.

In most cases, there is a trade-off between size and power, whch is easily understood in the EWS context (Example 3.1): increasing the threshold a reduces the fraction of FP (thereby reducing the size of the test), but it increases the fraction of FN (thereby reducing the power of the test).

3.2 The different types of statistical tests

How to determine the critical region? Loosely speaking, we want the critical region to be a set of "implausible" values of the test statistic S under the null hypothesis H_0. The lower the size of the test (α), the more implausible these values. Recall that, by definition of the size of the test, $\alpha = \mathbb{P}(S \in \Omega | H_0)$. That is, if α is small, there is only a small probability that S lies in Ω under H_0.

Consider the case where, under H_0, the distribution of the test statistic is symmetrical (e.g., normal distribution or Student-t distribution). In this case, the critical region is usually defined by the union of the two tails of the distribution. The test is then said to be a **two-tailed test** or a **two-sided test**. This situation is illustrated by Figures 3.1 and 3.2. (Use this web interface to explore alternative situations.)

Figure 3.2 also illustrates the notion of *p-value* (in the case of a two-sided test). The p-value can be defined as the value of the size of the test α that is such that the computed test statistic, S, is at the "frontier" of the critical region. Given this definition, if the p-value is smaller (respectively larger) than the size of the test, we reject (resp. cannot rejet) the null hypothesis at the α significance level.

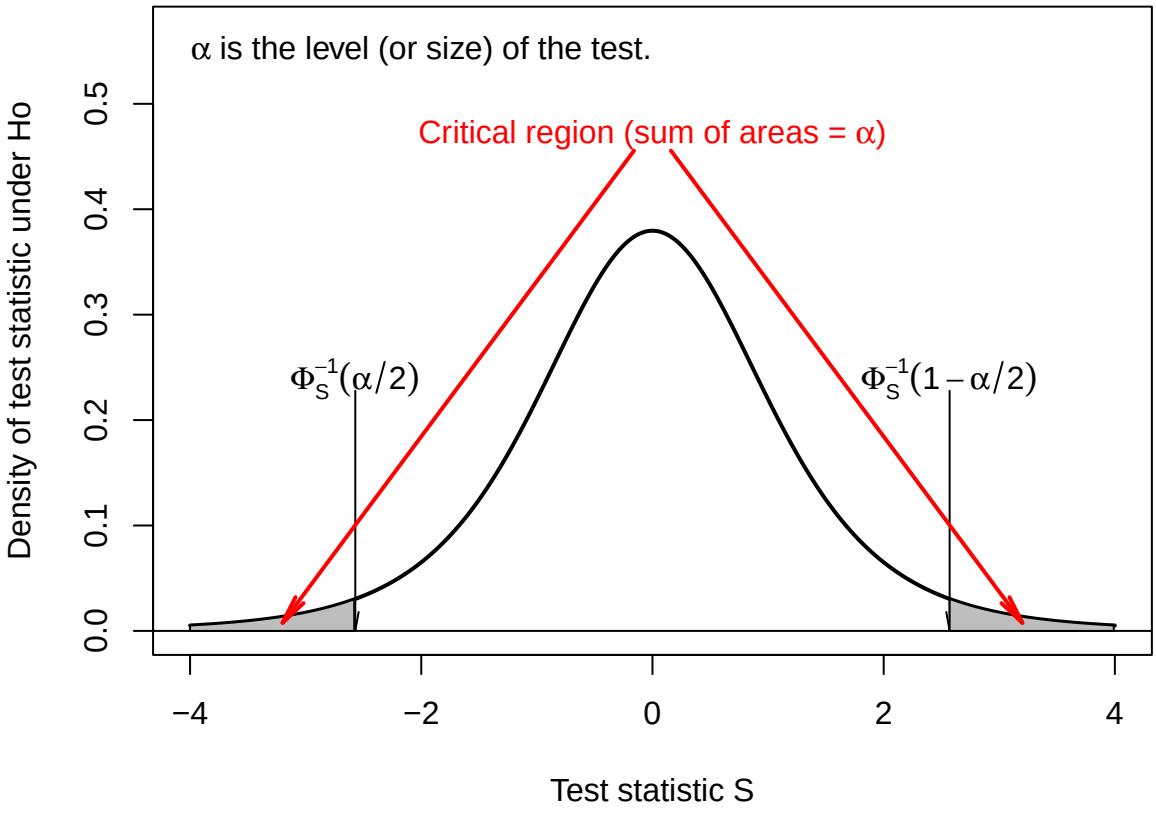


Figure 3.1: Two-sided test. Under H_0, $S \sim t(5)$. α is the size of the test.

Figures 3.3 and 3.4 illustrate the **one-tailed**, or **one-sided** situation. These tests are typically employed when the distribution of the test statistic under the null hypothesis has a

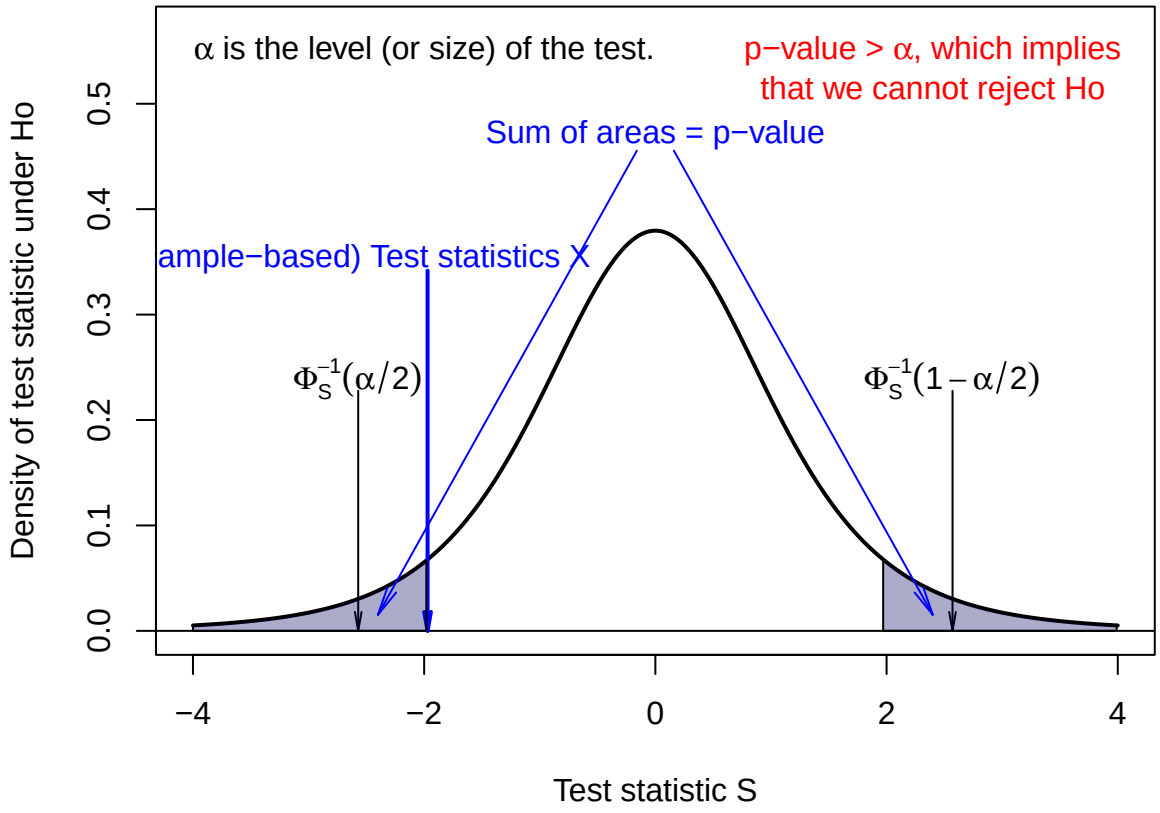


Figure 3.2: Two-sided test. Under H_0, $S \sim t(5)$. α is the size of the test.

support on $\mathbb{R}^+$ (e.g., the chi-square distribution χ^2, see Def. 9.13). Figure 3.4 also illustrates the notion of p-value associated with a one-sided statistical test.

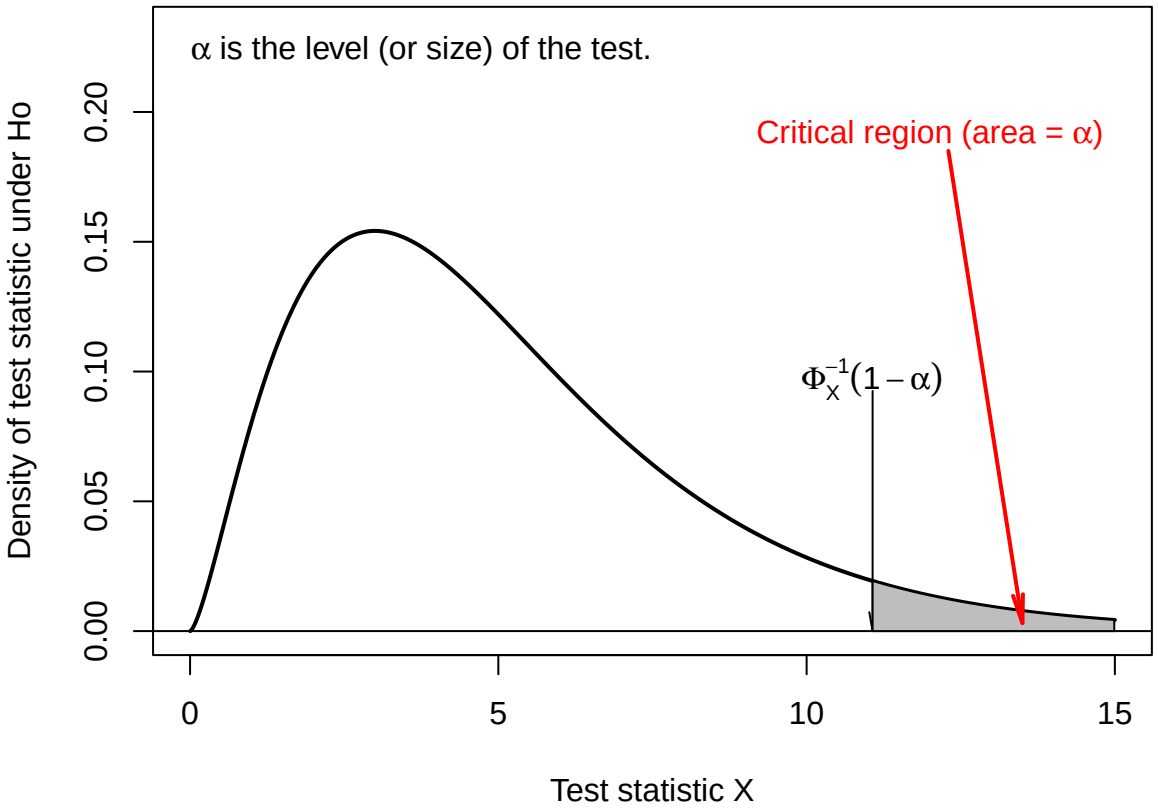


Figure 3.3: One-sided test. Under H_0, $S \sim \chi^2(5)$. α is the size of the test.

Example 3.2 (A practical illustration of size and power)**.** Consider a factory that produces metal cylinders whose diameter has to be equal to 1,cm. The tolerance is $a = 0.01$ cm. More precisely, more than 90% of the pieces have to satisfy the tolerance for the whole production (say 1.000.000 pieces) to be bought by the client.

The production technology is such that a proportion θ (imperfectly known) of the pieces does not satisfy the tolerance. The (population) parameter θ could be computed by measuring all the pieces but this would be costly. Instead, it is decided that $n \ll 1.000.000$ pieces will be measured. In this context, the null hypothesis H_0 is $\theta < 10\%$. The producing firm would like H_0 to be true.

Let us denote by d_i the binary indicator defined as:

$$d_i = \begin{cases} 0 & \text{if the size of the } i^{th} \text{ cylinder is in } [1-a, 1+a]; \\ 1 & \text{otherwise.} \end{cases}$$

We set $x_n = \sum_{i=1}^{n} d_i$. That is, x_n is the number of measured pieces that do not satisfy the tolerance (out of n).

The decision rule is: accept H_0 if $\dfrac{x_n}{n} \leq b$, reject otherwise. A natural choice for b is $b = 0.1$. However, this would not be a conservative choice, since it remains likely that $x_n < 0.1$ even

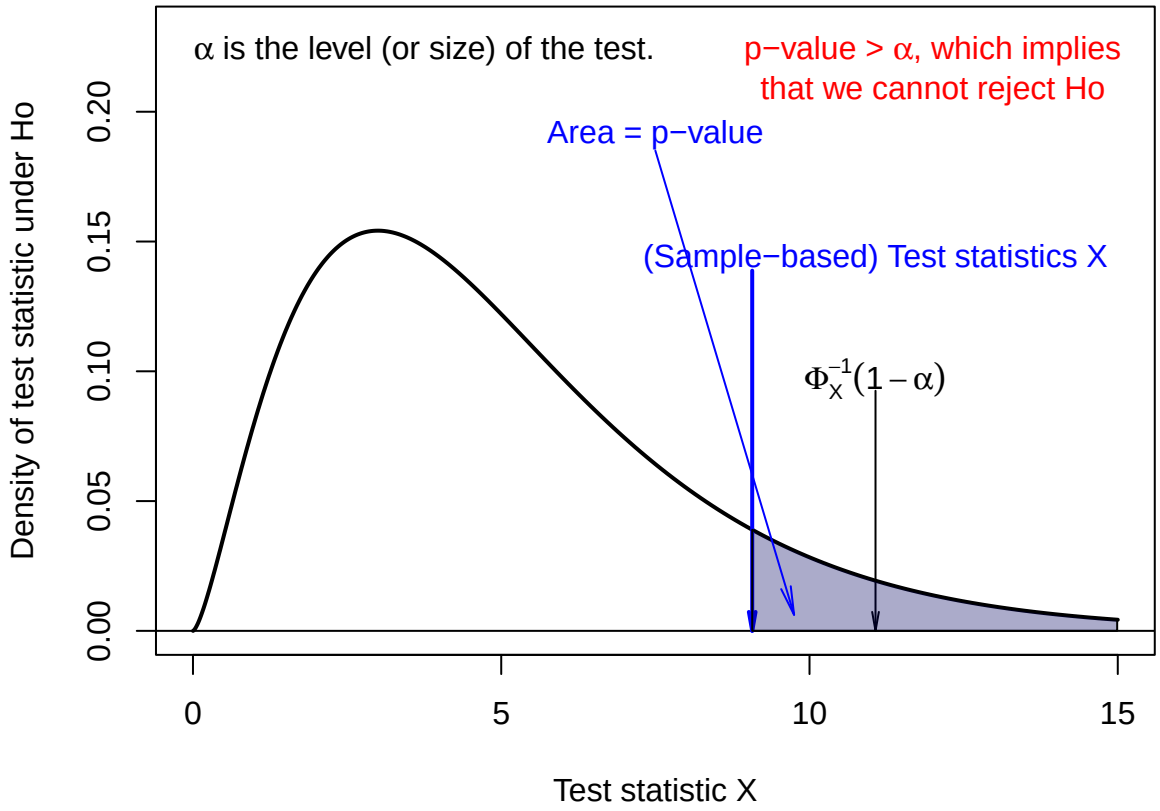


Figure 3.4: One-sided test. Under H_0, $S \sim \chi^2(5)$. α is the size of the test.

if $\mathbb{E}(x_n) = \theta > 0.1$ (especially if n is small). Hence, if one chooses $b = 0.1$, the probability of false negative may be high.

In this simple example, the size and the power of the test can be computed analytically. The test statistic is $S_n = \frac{x_n}{n}$ and the critical region is $\Omega = [b, 1]$. The probability to reject H_0 is:

$$\mathbb{P}_\theta(S_n \in \Omega) = \sum_{i=b\times n+1}^{n} C_n^i \theta^i (1-\theta)^{n-i}.$$

When $\theta < 0.1$, the previous expression gives the size of the test, while it gives the power of the test when $\theta > 0.1$. Figure 3.5 shows how this probability depends on θ for two sample sizes: $n = 100$ (upper plot) and $n = 500$ (lower plot).

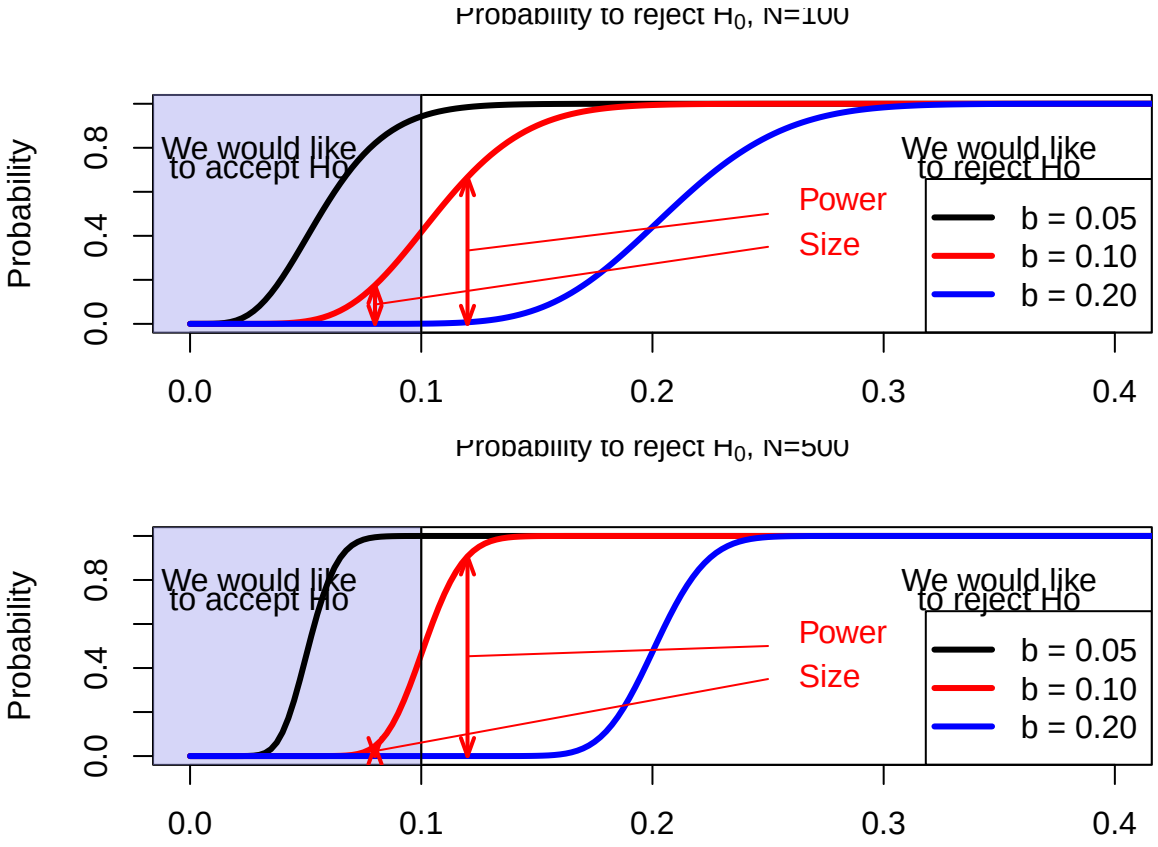


Figure 3.5: Factory example.

(Alternative situations can be explored by using this web interface.)

3.3 Asymptotic properties of statistical tests

In some cases, the distribution of the test statistic is known only asymptotically. For instance, it may be the case that the distribution of a given test statistic becomes normal only in large samples (e.g., related to the CLT, see Section 2). The level of the test is then not know

for small sample sizes, but only asymptotically, i.e., when n becomes infinitely large. That defines the asymptotic level of the test:

Definition 3.2 (Asymptotic level)**.** An asymptotic test with critical region Ω_n has an asymptotic level equal to α if:

$$\sup_{\theta\in\Theta} \lim_{n\to\infty} \mathbb{P}_\theta(S_n \in \Omega_n) = \alpha.$$

Example 3.3 (The factory example)**.** Let us come back to the factory example (Example 3.2). Because $S_n = \bar{d}_n$, and since $\mathbb{E}(d_i) = \theta$ and $\mathbb{V}ar(d_i) = \theta(1-\theta)$, the CLT (Theorem 9.5) leads to:

$$S_n \sim \mathcal{N}\left(\theta, \frac{1}{n}\theta(1-\theta)\right) \quad or \quad \frac{\sqrt{n}(S_n-\theta)}{\sqrt{\theta(1-\theta)}} \sim \mathcal{N}(0,1).$$

Hence, $\mathbb{P}_\theta(S_n \in \Omega_n) = \mathbb{P}_\theta(S_n > b) \approx 1 - \Phi\left(\frac{\sqrt{n}(b-\theta)}{\sqrt{\theta(1-\theta)}}\right)$. Since function $\theta \to 1 - \Phi\left(\frac{\sqrt{n}(b-\theta)}{\sqrt{\theta(1-\theta)}}\right)$ increases w.r.t. θ, we have:

$$\sup_{\theta\in\Theta=[0,0.1]} \mathbb{P}_\theta(S_n > b_n) = \mathbb{P}_{\theta=0.1}(S_n \in \Omega_n) \approx$$

$$1 - \Phi\left(\frac{\sqrt{n}(b_n - 0.1)}{0.3}\right).$$

Hence, if we set $b_n = 0.1 + 0.3\Phi^{-1}(1-\alpha)/\sqrt{n}$, we have $\sup_{\theta\in\Theta=[0,0.1]} \mathbb{P}_\theta(S_n > b_n) \approx \alpha$ for large values of n.

Let us now turn to the power of the test. Although it is often difficult to compute the power of the test, it is sometimes feasible to demonstrate that the power of the test converges to one when n goes to $+\infty$. In that case, if H_0 is false, the probability to reject it tends to be close to one in large samples.

Definition 3.3 (Asymptotically consistent test)**.** An asymptotic test with critical region Ω_n is consistent if:

$$\forall\theta \in \Theta^c, \quad \mathbb{P}_\theta(S_n \in \Omega_n) \to 1.$$

Example 3.4 (The factory example)**.** Let us come back to the factory example (Example 3.2). We proceed under the assumption that $\theta > 0.1$ and we consider $b_n = b = 0.1$. We still have:

$$\mathbb{P}_\theta(S_n \in \Omega_n) = \mathbb{P}_\theta(S_n > b) \approx 1 - \Phi\left(\frac{\sqrt{n}(b-\theta)}{\sqrt{\theta(1-\theta)}}\right).$$

Because $\frac{\sqrt{n}(b-\theta)}{\sqrt{\theta(1-\theta)}} \underset{n\to\infty}{\to} -\infty$, we have

$$\mathbb{P}_\theta(S_n > b) \approx 1 - \underbrace{\Phi\left(\frac{\sqrt{n}(b-\theta)}{\sqrt{\theta(1-\theta)}}\right)}_{\underset{n\to\infty}{\to} 0} \underset{n\to\infty}{\to} 1.$$

Therefore, with $b_n = b = 0.1$, the test is consistent.

3.4 Example: Normality tests

Because many statistical results are valid only if the underlying data is normally distributed, researchers often have tu conduct normality tests. Under the null hypothesis, the data at hand ($\mathbf{y} = \{y_1, \dots, y_n\}$, say) are drawn from a Gaussian distribution. A popular normality test is the Jarque-Bera test. It consists in verifying that the sample skewness and kurtosis of the y_i's are consistent with those of the normal distribution. Let us first define the skewness and kurtosis of a random variable.

Let f be the p.d.f. of Y. The k^{th} **standardized moment** of Y is defined as:

$$\psi_k = \frac{\mu_k}{\left(\sqrt{\mathbb{V}ar(Y)}\right)^k},$$

where $\mathbb{E}(Y) = \mu$ and

$$\mu_k = \mathbb{E}[(Y-\mu)^k] = \int_{-\infty}^{\infty} (y-\mu)^k f(y) dy$$

is the k^{th} **central moment** of Y. In particular, $\mu_2 = \mathbb{V}ar(Y)$. Therefore:

$$\psi_k = \frac{\mu_k}{\left(\mu_2^{1/2}\right)^k},$$

The **skewness** of Y corresponds to ψ_3 and the **kurtosis** to ψ_4 (Def. 9.6).

Proposition 3.1 (Skewness and kurtosis of the normal distribution)**.** *For a Gaussian var., the skewness (ψ_3) is 0 and the kurtosis (ψ_4) is 3.*

Proof. For a centered Gaussian distribution, $(-y)^3 f(-y) = -y^3 f(y)$. This implies that

$$\begin{aligned}\int_{-\infty}^{\infty} y^3 f(y) dy &= \int_{-\infty}^{0} y^3 f(y) dy + \int_{0}^{\infty} y^3 f(y) dy \\ &= -\int_{0}^{\infty} y^3 f(y) dy + \int_{0}^{\infty} y^3 f(y) dy = 0,\end{aligned}$$

which leads to the skewness result.

Moreover, for a Gaussian distribution, $df(y)/dy = -yf(y)$ and therefore $\frac{d}{dy}(y^3 f(y)) = 3y^2 f(y) - y^4 f(y)$. Partial integration leads to the kurtosis result. □

Let us now introduce the sample analog of standardized moments. The k^{th} **central sample moment** of Y is given by:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^k,$$

and the k^{th} **standardized sample moment** of Y is given by:

$$g_k = \frac{m_k}{m_2^{k/2}}.$$

Proposition 3.2 (Consistency of central sample moments)**.** *If the k^{th} central moment of Y, exists and if the y_i's are i.i.d., then the sample central moment m_k is a consistent estimate of the central moment μ_k.*

Proposition 3.3 (Asymptotic distribution of 3rd-order sample central moment of a normal distribution)**.** *If $y_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then $\sqrt{n} g_3 \xrightarrow{d} \mathcal{N}(0, 6)$.*

Proof. See, e.g. Lehmann (1999). □

Proposition 3.4 (Asymptotic distribution of 4th-order sample central moment of a normal distribution)**.** *If $y_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then $\sqrt{n}(g_4 - 3) \xrightarrow{d} \mathcal{N}(0, 24)$.*

Proposition 3.5 (Joint asymptotic distribution of 3rd and 4th-order sample central moments of a normal distribution)**.** *If $y_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then the vector $(\sqrt{n} g_3, \sqrt{n}(g_4 - 3))$ is asymptotically bivariate Gaussian. Further its elements are uncorrelated and therefore independent.*

The Jarque-Bera statistic is defined by:

$$JB = n \left(\frac{g_3^2}{6} + \frac{(g_4 - 3)^2}{24} \right) = \frac{n}{6} \left(g_3^2 + \frac{(g_4 - 3)^2}{4} \right).$$

Proposition 3.6 (Jarque-Bera asympt. distri.)**.** *If $y_i \sim$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $JB \xrightarrow{d} \chi^2(2)$.*

Proof. This directly derives from Proposition 3.5. □

Example 3.5 (Consistency of the Jarque-Bera normality test)**.** This example illustrates the consistency of the JB test (see Def. 9.8).

For each row of matrix `x`, the function `JB` (defined below) computes the Jarque-Bera test statistic. (That is, each row is considered as a given sample.)

```
JB <- function(x){
  N <- dim(x)[1] # number of samples
  n <- dim(x)[2] # sample size
  x.bar <- apply(x,1,mean)
  x.x.bar <- x - matrix(x.bar,N,n)
  m.2 <- apply(x.x.bar,1,function(x){mean(x^2)})
  m.3 <- apply(x.x.bar,1,function(x){mean(x^3)})
  m.4 <- apply(x.x.bar,1,function(x){mean(x^4)})
  g.3 <- m.3/m.2^(3/2)
  g.4 <- m.4/m.2^(4/2)
  return(n*(g.3^2/6 + (g.4-3)^2/24))
}
```

Let us first consider the case where H_0 is satisfied. Figure 3.6 displays, for different sample sizes n, the distribution of the JB statistics when the y_i's are normal, consistently with H_0. It appears that when n grows, the distribution indeed converges to the $\chi^2(2)$ distribution (as stated by Proposition 3.6).

```
all.n <- c(5,10,20,100)
nb.sim <- 10000
y <- matrix(rnorm(nb.sim*max(all.n)),nb.sim,max(all.n))

par(mfrow=c(2,2));par(plt=c(.1,.95,.15,.8))
for(i in 1:length(all.n)){
  n <- all.n[i]
  hist(JB(y[,1:n]),nclass = 200,freq = FALSE,
       main=paste("n = ",toString(n),sep=""),xlim=c(0,10))
  xx <- seq(0,10,by=.01)
  lines(xx,dchisq(xx,df = 2),col="red")
}
```

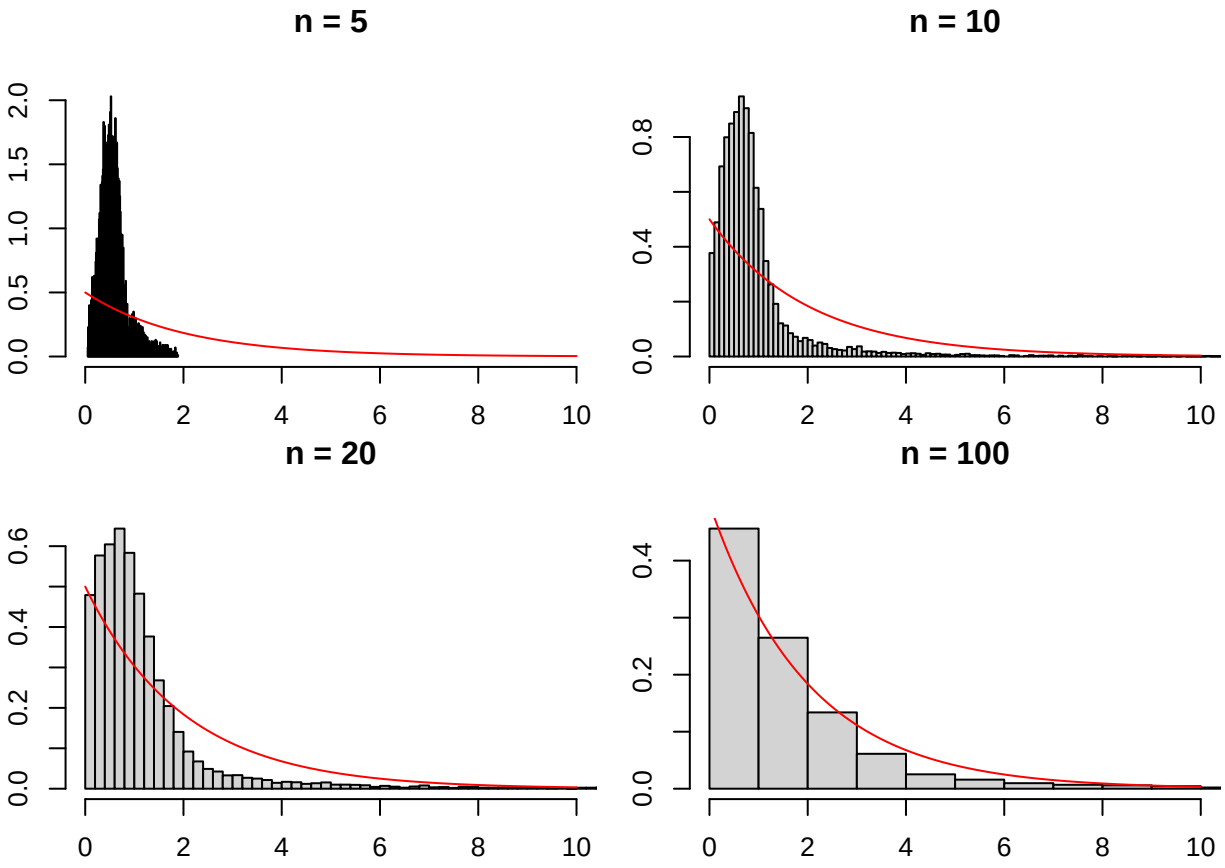


Figure 3.6: Distribution of the JB test statistic under H_0 (normality).

Now, replace `rnorm` with `runif`. The y_i's are then drawn from a uniform distribution. H_0 is

not satisfied. Figure 3.7 then shows that, when n grows, the distributions of the JB statistic shift to the right. This results in the consistency of the JB test (see Def. 9.8).

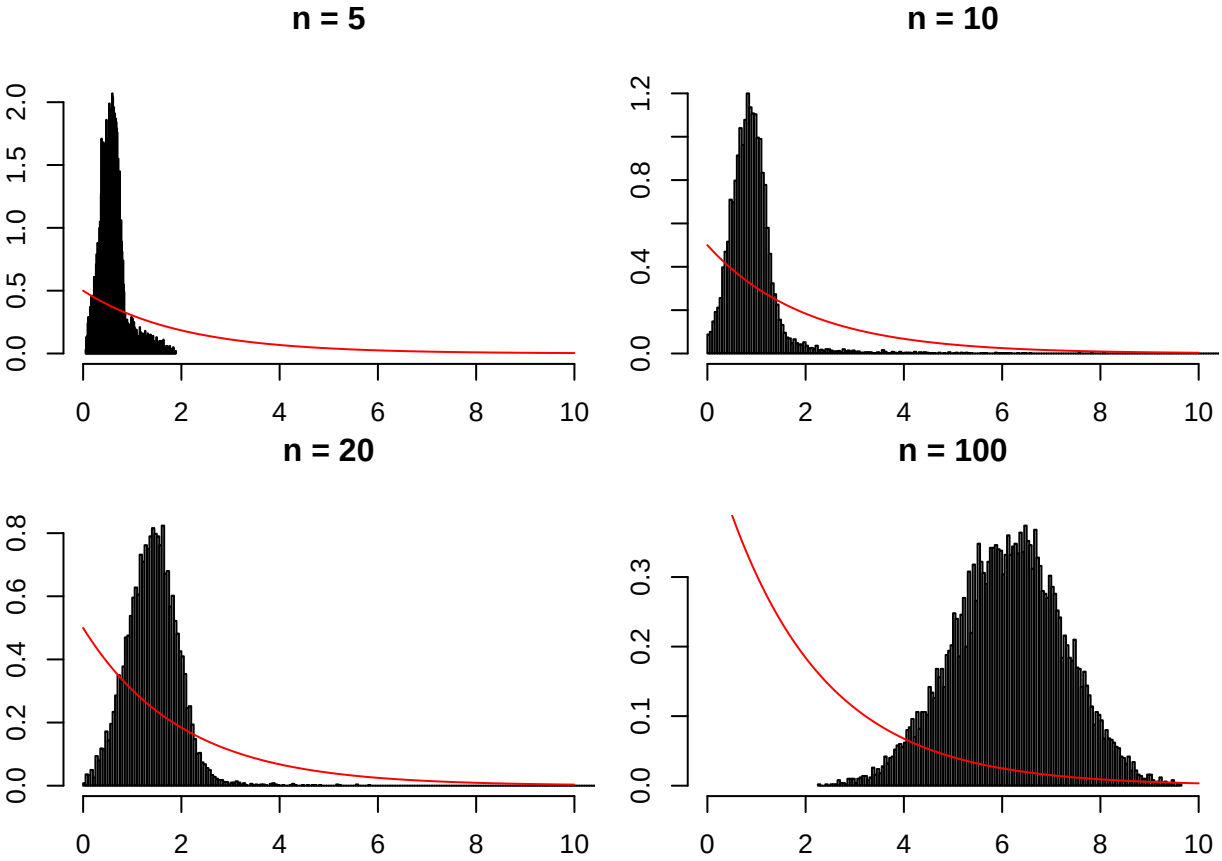


Figure 3.7: Distribution of the JB test statistic when the y_i's are drawn from a uniform distribution (hence H_0 is not satisfied).

Chapter 4

Linear Regressions

Definition 4.1. A linear regression model is of the form:

$$y_i = \beta' \mathbf{x}_i + \varepsilon_i, \tag{4.1}$$

where $\mathbf{x}_i = [x_{i,1}, \dots, x_{i,K}]'$ is a vector of dimension $K \times 1$.

For entity i, the $x_{i,k}$'s, for $k \in \{1, \dots, K\}$, are explanatory variables, regressors, or covariates. The variable of interest, y_i, is often called dependent variable, or regressand. The last term of the specification, namely ε_i, is called error, or disturbance.

The researcher is usually interested in the components of vector β, denoted by β_k, $k \in \{1, \dots, K\}$. She usually aims at estimating these coefficients based on observations of $\{y_i, \mathbf{x}_i\}$, $i \in \{1, \dots, n\}$, which constitutes a *sample*. In the following, we will denote the sample length by n.

To have an intercept in the specification (4.1), one has to set $x_{i,1} = 1$ for all i; β_1 then corresponds to the intercept.

4.1 Hypotheses

In this section, we introduce different assumptions regarding the covariates and/or the errors. The properties of the estimators used by the researcher depend on which of these assumptions are satisfied.

Hypothesis 4.1 (Full rank)**.** There is no exact linear relationship among the independent variables (the $x_{i,k}$'s, for a given $i \in \{1, \dots, n\}$).

Intuitively, when Hypothesis 4.1 is satisfied, then the estimation of the model parameters is unfeasible since, for any value of β, some changes in the explanatory variables will be exactly compensated by other changes in another set of explanatory variables, preventing the identification of these effects.

Let us denote by $\mathbf{X}$ the matrix containing all explanatory variables, of dimension $n \times K$. (That is, row i of $\mathbf{X}$ is $\mathbf{x}_i'$.) The following hypothesis concerns the relationship between the errors (gathered in ε, a n-dimensional vector) and the explanatory variables $\mathbf{X}$:

Hypothesis 4.2 (Conditional mean-zero assumption)**.**

$$\mathbb{E}(\varepsilon|\mathbf{X}) = 0. \tag{4.2}$$

Hypothesis 4.2 has important implications:

Proposition 4.1. *Under Hypothesis 4.2:*

i. $\mathbb{E}(\varepsilon_i) = 0$*;*
ii. The x_{ij}*'s and the* ε_i*'s are uncorrelated, i.e.* $\forall i, j \quad \mathbb{C}orr(x_{ij}, \varepsilon_i) = 0$.

Proof. Let us prove (i) and (ii):

i. By the law of iterated expectations:

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

ii. $\mathbb{E}(x_{ij}\varepsilon_i) = \mathbb{E}(\mathbb{E}(x_{ij}\varepsilon_i|\mathbf{X})) = \mathbb{E}(x_{ij}\underbrace{\mathbb{E}(\varepsilon_i|\mathbf{X})}_{=0}) = 0.$

□

The next two hypotheses (4.3 and 4.4) concern the stochastic properties of the errors ε_i:

Hypothesis 4.3 (Homoskedasticity)**.**

$$\forall i, \quad \mathbb{V}ar(\varepsilon_i|\mathbf{X}) = \sigma^2.$$

The following lines of code generate a figure comparing two situations: Panel (a) of Figure 4.1 corresponds to a situation of homoskedasticity, and Panel (b) corresponds to a situation of heteroskedasticity. Let us be more specific. In the two plots, we have $X_i \sim \mathcal{N}(0, 1)$ and $\varepsilon_i^* \sim \mathcal{N}(0, 1)$. In Panel (a) (homoskedasticity):

$$Y_i = 2 + 2X_i + \varepsilon_i^*.$$

In Panel (b) (heteroskedasticity):

$$Y_i = 2 + 2X_i + \left(2\mathbb{1}_{\{X_i<0\}} + 0.2\mathbb{1}_{\{X_i\geq 0\}}\right)\varepsilon_i^*$$

.

```
N <- 200
X <- rnorm(N);eps <- rnorm(N)
par(mfrow=c(1,2),plt=c(.2,.95,.2,.8))
Y <- 2 + 2*X + eps
plot(X,Y,pch=19,main="(a) Homoskedasticity",
     las=1,cex.lab=.8,cex.axis=.8,cex.main=.8,)
Y <- 2 + 2*X + eps*( (X<0)*2 + (X>=0)*.2 )
plot(X,Y,pch=19,main="(b) Heteroskedasticity",
     las=1,cex.lab=.8,cex.axis=.8,cex.main=.8,)
```

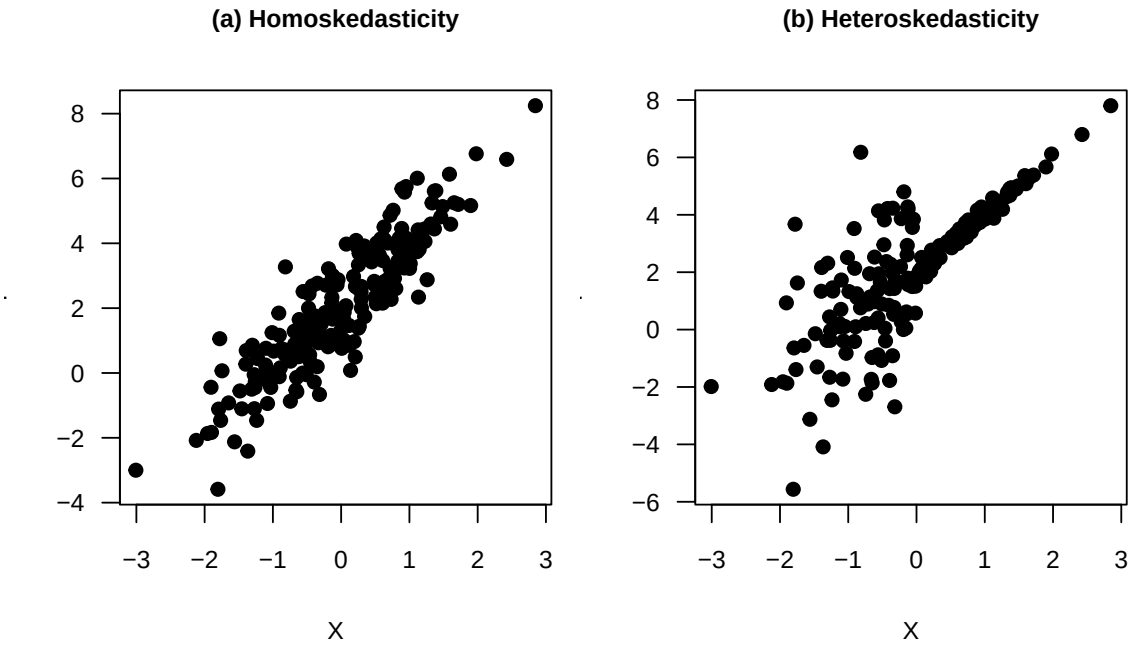


Figure 4.1: Homoskedasticity vs heteroskedasticity. See text for the exact specifications.

Figure 4.2 shows a real-data situation of heteroskedasticity, based on data taken from the Swiss Household Panel. The sample is restricted to persons (i) that are younger than 35 year in 2019, and (ii) that have completed at least 19 years of study. The figure shows that the dispersion of yearly income increases with age.

```
library(AEC)
table(shp$edyear19)
```

```
##
##    8    9   10   12   13   14   16   19   21
##   70  325  350 1985  454  117  990 1263  168
```

```
shp_higherEd <- subset(shp,(edyear19>18)&age19<35)
plot(i19wyg/1000~age19,data=shp_higherEd,pch=19,las=1,
     xlab="Age",ylab="Yearly work income")
abline(lm(i19wyg/1000~age19,data=shp_higherEd),col="red",lwd=2)
```

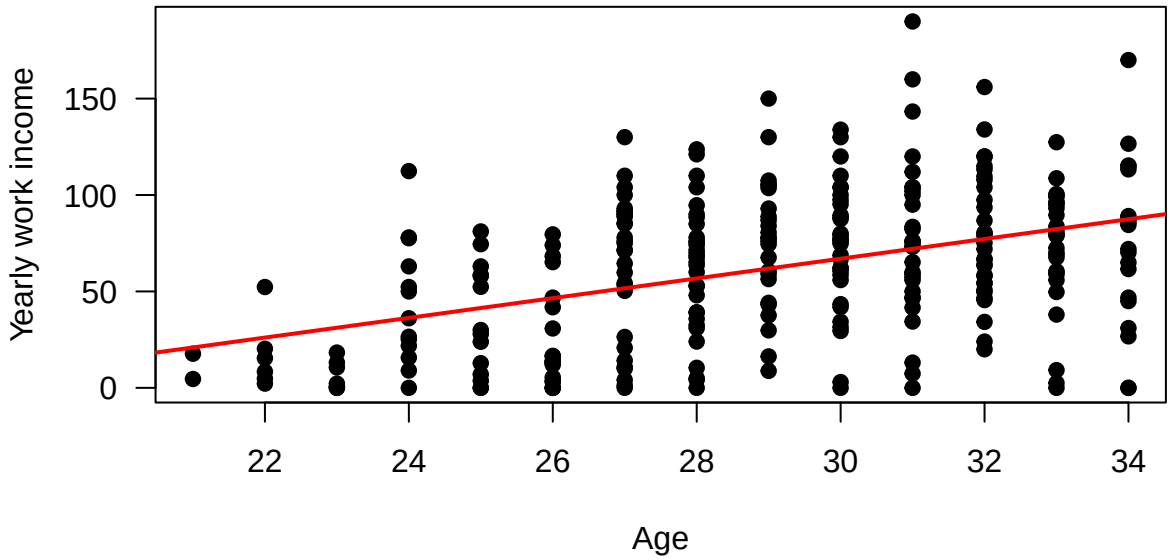


Figure 4.2: Income versus age. Data are from the Swiss Household Panel. The sample is restricted to persons that have completed at least 19 years of study. The figure shows that the dispersion of yearly income increases with age.

The next assumption concerns the correlation of the errors across entities.

Hypothesis 4.4 (Uncorrelated errors)**.**

$$\forall i \neq j, \quad \mathbb{C}ov(\varepsilon_i, \varepsilon_j | \mathbf{X}) = 0.$$

We will often need to work with the covariance matrix of the errors. Proposition 4.2 give the specific form of the covariance matrix of the errors —conditional on $\mathbf{X}$— when both Hypotheses 4.3 and 4.4 are satisfied:

Proposition 4.2. *If Hypotheses 4.3 and 4.4 hold, then:*

$$\mathbb{V}ar(\varepsilon | \mathbf{X}) = \sigma^2 Id,$$

where Id is the $n \times n$ identity matrix.

We will sometimes assume that errors are Gaussian—or normal. We will then invoke Hypothesis 4.5:

Hypothesis 4.5 (Normal distribution)**.**

$$\forall i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

4.2 Least square estimation

4.2.1 Derivation of the OLS formula

In this section, we will present and study the properties of the most popular estimation approach, namely the **Ordinary Least Squares (OLS)** approach. As suggested by its name, the OLS estimator of β is defined as the vector $\mathbf{b}$ that minimizes the sum of squared residuals. (The *residuals* are the estimates of the *errors* ε_i.)

For a given vector of coefficients $\mathbf{b} = [b_1, ..., b_K]'$, the sum of squared residuals is:

$$f(\mathbf{b}) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{K} x_{i,j} b_j \right)^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \mathbf{b})^2.$$

Minimizing this sum amounts to minimizing:

$$f(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}).$$

Since:[1]

$$\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b},$$

it comes that a necessary first-order condition (FOC) is:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{4.3}$$

Under Assumption 4.1, $\mathbf{X}'\mathbf{X}$ is invertible. Hence:

$$\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.}$$

Vector $\mathbf{b}$ minimizes the sum of squared residuals. (f is a non-negative quadratic function, it therefore admits a minimum.)

We have:

$$\mathbf{y} = \underbrace{\mathbf{X}\mathbf{b}}_{\text{fitted values } (\hat{\mathbf{y}})} + \underbrace{\mathbf{e}}_{\text{residuals}}$$

The estimated residuals are:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y}, \tag{4.4}$$

where $\mathbf{M} := \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the **residual maker** matrix.

Moreover, the fitted values $\hat{\mathbf{y}}$ are given by:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}, \tag{4.5}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a **projection matrix**.

These matrices $\mathbf{M}$ and $\mathbf{P}$ are such that:

[1] see Proposition 9.7.

- $\mathbf{MX} = \mathbf{0}$: if one regresses one of the explanatory variables on $\mathbf{X}$, the residuals are null.
- $\mathbf{My} = \mathbf{M}\varepsilon$ (because $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ and $\mathbf{MX} = \mathbf{0}$).

Here are some additional properties of $\mathbf{M}$ and $\mathbf{P}$:

- $\mathbf{M}$ is symmetric ($\mathbf{M} = \mathbf{M}'$) and idempotent ($\mathbf{M} = \mathbf{M}^2 = \mathbf{M}^k$ for $k > 0$).
- $\mathbf{P}$ is symmetric and idempotent.
- $\mathbf{PX} = \mathbf{X}$.
- $\mathbf{PM} = \mathbf{MP} = 0$.
- $\mathbf{y} = \mathbf{Py} + \mathbf{My}$ (decomposition of $\mathbf{y}$ into two orthogonal parts).

It is easily checked that $\mathbf{X}'\mathbf{e} = 0$. Each column of $\mathbf{X}$ is therefore orthogonal to $\mathbf{e}$. In particular, if an intercept is included in the regression ($x_{i,1} \equiv 1$ for all i's, i.e., the first column of $\mathbf{X}$ is filled with ones), the average of the residuals is null.

Example 4.1 (Bivariate case)**.** Consider a bivariate situation, where we regress y_i on a constant and an explanatory variable w_i. We have $K = 2$, and $\mathbf{X}$ is a $n \times 2$ matrix whose i^{th} row is $[x_{i,1}, x_{i,2}]$, with $x_{i,1} = 1$ (to account for the intercept) and with $w_i = x_{i,2}$ (say).

We have:

$$\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{bmatrix} n & \sum_i w_i \\ \sum_i w_i & \sum_i w_i^2 \end{bmatrix}, \\
(\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n\sum_i w_i^2 - (\sum_i w_i)^2} \begin{bmatrix} \sum_i w_i^2 & -\sum_i w_i \\ -\sum_i w_i & n \end{bmatrix}, \\
(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \frac{1}{n\sum_i w_i^2 - (\sum_i w_i)^2} \begin{bmatrix} \sum_i w_i^2 \sum_i y_i - \sum_i w_i \sum_i w_i y_i \\ -\sum_i w_i \sum_i y_i + n\sum_i w_i y_i \end{bmatrix} \\
&= \frac{1}{\frac{1}{n}\sum_i (w_i - \bar{w})^2} \begin{bmatrix} \frac{\bar{y}}{n}\sum_i w_i^2 - \frac{\bar{w}}{n}\sum_i w_i y_i \\ \frac{1}{n}\sum_i (w_i - \bar{w})(y_i - \bar{y}) \end{bmatrix}.
\end{aligned}$$

It can be seen that the second element of $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is:

$$b_2 = \frac{\overline{\mathbb{C}ov(W, Y)}}{\overline{\mathbb{V}ar(W)}},$$

where $\overline{\mathbb{C}ov(W, Y)}$ and $\overline{\mathbb{V}ar(W)}$ are sample estimates.

Since there is a constant in the regression, we have $b_1 = \bar{y} - b_2\bar{w}$.

4.2.2 Properties of the OLS estimate (small sample)

The OLS properties stated in Proposition 4.3 are valid for any sample size n:

Proposition 4.3 (Properties of the OLS estimator)**.** *We have:*

i. Under Assumptions 4.1 and 4.2, the OLS estimator is linear and unbiased.

ii. Under Hypotheses 4.1 to 4.4, the conditional covariance matrix of $\mathbf{b}$ *is:* $\mathbb{V}ar(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Proof. Under Hypothesis 4.1, $\mathbf{X}'\mathbf{X}$ can be inverted. We have:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \varepsilon.$$

i. Let us consider the expectation of the last term, i.e. $\mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \varepsilon)$. Using the law of iterated expectations, we obtain:

$$\mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \varepsilon) = \mathbb{E}(\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \varepsilon|\mathbf{X}]) = \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\varepsilon|\mathbf{X}]).$$

By Hypothesis 4.2, we have $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$. Hence $\mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \varepsilon) = 0$ and result (i) follows.

ii. $\mathbb{V}ar(\mathbf{b}|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. By Prop. 4.2, if 4.3 and 4.4 hold, then we have $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \mathbb{V}ar(\varepsilon|\mathbf{X}) = \sigma^2 Id$.

□

Together, Hypotheses 4.1 to 4.4 form the so-called Gauss-Markov set of assumptions. Under these assumptions, the OLS estimator feature the lowest possible variance within the family of linear unbiased estimates of β:

Theorem 4.1 (Gauss-Markov Theorem)**.** *Under Assumptions 4.1 to 4.4, for any vector* w*, the minimum-variance linear unbiased estimator of* $w'\beta$ *is* $w'\mathbf{b}$*, where* $\mathbf{b}$ *is the least squares estimator. (BLUE: Best Linear Unbiased Estimator.)*

Proof. Consider $\mathbf{b}^* = C\mathbf{y}$, another linear unbiased estimator of β. Since it is unbiased, we must have $\mathbb{E}(C\mathbf{y}|\mathbf{X}) = \mathbb{E}(C\mathbf{X}\beta + C\varepsilon|\mathbf{X}) = \beta$. We have $\mathbb{E}(C\varepsilon|\mathbf{X}) = C\mathbb{E}(\varepsilon|\mathbf{X}) = 0$ (by 4.2). Therefore $\mathbf{b}^*$ is unbiased if $\mathbb{E}(C\mathbf{X})\beta = \beta$. This has to be the case for any β, which implies that we must have $C\mathbf{X} = \mathbf{I}$. Let us compute $\mathbb{V}ar(\mathbf{b}^*|\mathbf{X})$. For this, we introduce $D = C - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which is such that $D\mathbf{y} = \mathbf{b}^* - \mathbf{b}$. The fact that $C\mathbf{X} = \mathbf{I}$ implies that $D\mathbf{X} = \mathbf{0}$. We have $\mathbb{V}ar(\mathbf{b}^*|\mathbf{X}) = \mathbb{V}ar(C\mathbf{y}|\mathbf{X}) = \mathbb{V}ar(C\varepsilon|\mathbf{X}) = \sigma^2 CC'$ (by Assumptions 4.3 and 4.4, see Prop. 4.2). Using $C = D + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and exploiting the fact that $D\mathbf{X} = \mathbf{0}$ leads to:

$$\mathbb{V}ar(\mathbf{b}^*|\mathbf{X}) = \sigma^2\left[(D + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(D + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'\right] = \mathbb{V}ar(\mathbf{b}|\mathbf{X}) + \sigma^2\mathbf{D}\mathbf{D}'.$$

Therefore, we have

$$\begin{aligned} & \mathbb{V}ar(w'\mathbf{b}^*|\mathbf{X}) = w'\mathbb{V}ar(\mathbf{b}|\mathbf{X})w + \sigma^2 w'\mathbf{D}\mathbf{D}'w \\ \geq\ & w'\mathbb{V}ar(\mathbf{b}|\mathbf{X})w = \mathbb{V}ar(w'\mathbf{b}|\mathbf{X}). \end{aligned}$$

□

The Frish-Waugh theorem (Theorem 4.2) reveals the relationship between the OLS estimator and the notion of partial correlation coefficient. Consider the linear least square regression of $\mathbf{y}$ on $\mathbf{X}$. We introduce the notations:

- $\mathbf{b}^{\mathbf{y}/\mathbf{X}}$: OLS estimates of β,
- $\mathbf{M}^{\mathbf{X}}$: residual-maker matrix of any regression on $\mathbf{X}$ (given by $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$),
- $\mathbf{P}^{\mathbf{X}}$: projection matrix of any regression on $\mathbf{X}$ (given by $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$).

Let us split the set of explanatory variables into two: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$. With obvious notations: $\mathbf{b}^{\mathbf{y}/\mathbf{X}} = [\mathbf{b}_1', \mathbf{b}_2']'$.

Theorem 4.2 (Frisch-Waugh Theorem)**.** *We have:*

$$\mathbf{b}_2 = \mathbf{b}^{\mathbf{M}^{\mathbf{X}_1}\mathbf{y}/\mathbf{M}^{\mathbf{X}_1}\mathbf{X}_2}.$$

Proof. The minimization of the least squares leads to (these are first-order conditions, see Eq. (4.3)):

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix}.$$

Use the first-row block of equations to solve for $\mathbf{b}_1$ first; it comes as a function of $\mathbf{b}_2$. Then use the second set of equations to solve for $\mathbf{b}_2$, which leads to:

$$\begin{aligned} \mathbf{b}_2 &= [\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)\mathbf{X}_1'\mathbf{X}_2]^{-1}\mathbf{X}_2'(Id - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)\mathbf{X}_1')\mathbf{y} \\ &= [\mathbf{X}_2'\mathbf{M}^{\mathbf{X}_1}\mathbf{X}_2]^{-1}\mathbf{X}_2'\mathbf{M}^{\mathbf{X}_1}\mathbf{y}. \end{aligned}$$

Using the fact that $\mathbf{M}^{\mathbf{X}_1}$ is idempotent and symmetric leads to the result. □

This suggests a second way of estimating $\mathbf{b}_2$:

1. Regress Y on X_1, regress X_2 on X_1.
2. Regress the residuals associated with the former regression on those associated with the latter regressions.

This is illustrated by the following code, where we run different regressions involving the number of Google searches for "parapluie" (*umbrella* in French). In the broad specification, we regress it on French precipitations and month dummies. Next, we deseasonalize both the dependent variable and the precipitations by regressing them on the month dummies. As stated by Theorem 4.2, regressing deseasonalized Google searches on deseasonalized precipitations give the same coefficient as in the baseline regression.

```
library(AEC)
dummies <- as.matrix(parapluie[,4:14])
eq_all <- lm(parapluie~dummies+precip,data=parapluie)
deseas_parapluie <- lm(parapluie~dummies,data=parapluie)$residuals
deseas_precip    <- lm(precip~dummies,data=parapluie)$residuals
eq_frac <- lm(deseas_parapluie~deseas_precip-1)
stargazer::stargazer(eq_all,eq_frac,omit=c(1:11,"Constant"),type="text",
                     omit.stat = c("f","ser"),digits=5,
                     add.lines=list(c('Monthly dummy','Yes','No')))
```

```
##
## ===========================================
##                    Dependent variable:
##                ---------------------------
##                 parapluie  deseas_parapluie
##                    (1)           (2)
## -------------------------------------------
## precip         0.13001***
##                 (0.03594)
##
## deseas_precip                 0.13001***
##                               (0.03277)
##
## -------------------------------------------
## Monthly dummy      Yes            No
## Observations       72             72
## R2                0.51793        0.18148
## Adjusted R2       0.41988        0.16995
## ===========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

When b_2 is scalar (and then $\mathbf{X}_2$ is of dimension $n \times 1$), Theorem 4.2 gives the expression of the **partial regression coefficient** b_2:

$$b_2 = \frac{\mathbf{X}_2' M^{\mathbf{X}_1} \mathbf{y}}{\mathbf{X}_2' M^{\mathbf{X}_1} \mathbf{X}_2}.$$

4.2.3 Goodness of fit

Define the total variation in y as the sum of squared deviations (from the sample mean):

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

We have:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

In the following, we assume that the regression includes a constant (i.e. for all i, $x_{i,1} = 1$). Denote by $\mathbf{M}^0$ the matrix that transforms observations into deviations from sample means. Using that $\mathbf{M}^0\mathbf{e} = \mathbf{e}$ and that $\mathbf{X}'\mathbf{e} = 0$, we have:

$$\begin{aligned} \underbrace{\mathbf{y}'\mathbf{M}^0\mathbf{y}}_{\text{Total sum of sq.}} &= (\mathbf{X}\mathbf{b} + \mathbf{e})'\mathbf{M}^0(\mathbf{X}\mathbf{b} + \mathbf{e}) \\ &= \underbrace{\mathbf{b}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\mathbf{b}}_{\text{"Explained" sum of sq.}} + \underbrace{\mathbf{e}'\mathbf{e}}_{\text{Sum of sq. residuals}} \\ TSS &= Expl.SS + SSR. \end{aligned}$$

We can now define the coefficient of determination:

$$\boxed{\text{Coefficient of determination} = \frac{Expl.SS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}.} \tag{4.6}$$

It can be shown (Greene (2003), Section 3.5) that:

$$\text{Coefficient of determination} = \frac{[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}.$$

That is, the R^2 is the sample squared correlation between y and the (regression-implied) y's predictions.

The hgher the R^2, the higher the goodness of fit of a model. One however has to be cautious with R^2. Indeed, it is easy to increase it: it suffices to add explanatory variables. As stated by Proposition 4.5, adding an explanatory variable (even if it does not truly relate to the dependent variable) mechanically results in an increase in the R^2. In the limit, taking any set of n non-linearly-dependent explanatory variables (i.e., variables satisfying Hypothesis 4.1) results in a R^2 equal to one.

Proposition 4.4 (Change in SSR when a variable is added)**.** *We have:*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}^{*\prime}\mathbf{z}^*) \qquad (\leq \mathbf{e}'\mathbf{e}) \tag{4.7}$$

where (i) $\mathbf{u}$ *and* $\mathbf{e}$ *are the residuals in the regressions of* $\mathbf{y}$ *on* $[\mathbf{X}, \mathbf{z}]$ *and of* $\mathbf{y}$ *on* $\mathbf{X}$*, respectively, (ii)* c *is the regression coefficient on* $\mathbf{z}$ *in the former regression and where* $\mathbf{z}^*$ *are the residuals in the regression of* $\mathbf{z}$ *on* $\mathbf{X}$*.*

Proof. The OLS estimates $[\mathbf{d}', \mathbf{c}]'$ in the regression of $\mathbf{y}$ on $[\mathbf{X}, \mathbf{z}]$ satisfies (first-order cond., Eq. (4.3)):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{z} \\ \mathbf{z}'\mathbf{X} & \mathbf{z}'\mathbf{z} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ c \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{z}'\mathbf{y} \end{bmatrix}.$$

Hence, in particular $\mathbf{d} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}c$, where $\mathbf{b}$ is the OLS of $\mathbf{y}$ on $\mathbf{X}$. Substituting in $\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{d} - \mathbf{z}c$, we get $\mathbf{u} = \mathbf{e} - \mathbf{z}^*c$. We therefore have:

$$\mathbf{u}'\mathbf{u} = (\mathbf{e} - \mathbf{z}^*c)(\mathbf{e} - \mathbf{z}^*c) = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}^{*'}\mathbf{z}^*) - 2c\mathbf{z}^{*'}\mathbf{e}. \tag{4.8}$$

Now $\mathbf{z}^{*'}\mathbf{e} = \mathbf{z}^{*'}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{z}^{*'}\mathbf{y}$ because $\mathbf{z}^*$ are the residuals in an OLS regression on $\mathbf{X}$. Since $c = (\mathbf{z}^{*'}\mathbf{z}^*)^{-1}\mathbf{z}^{*'}\mathbf{y}^*$ (by an application of Theorem 4.2), we have $(\mathbf{z}^{*'}\mathbf{z}^*)c = \mathbf{z}^{*'}\mathbf{y}^*$ and, therefore, $\mathbf{z}^{*'}\mathbf{e} = (\mathbf{z}^{*'}\mathbf{z}^*)c$. Inserting this in Eq. (4.8) leads to the results. □

Proposition 4.5 (Change in the coefficient of determination when a variable is added)**.** *Denoting by R^2_W the coefficient of determination in the regression of $\mathbf{y}$ on some variable $\mathbf{W}$, we have:*

$$R^2_{\mathbf{X},\mathbf{z}} = R^2_{\mathbf{X}} + (1 - R^2_{\mathbf{X}})(r^{\mathbf{X}}_{yz})^2,$$

where $r^{\mathbf{X}}_{yz}$ is the coefficient of partial correlation (see Definition 9.5).

Proof. Let's use the same notations as in Prop. 4.4. Theorem 4.2 implies that $c = (\mathbf{z}^{*'}\mathbf{z}^*)^{-1}\mathbf{z}^{*'}\mathbf{y}^*$. Using this in Eq. (4.7) gives $\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - (\mathbf{z}^{*'}\mathbf{y}^*)^2/(\mathbf{z}^{*'}\mathbf{z}^*)$. Using the definition of the partial correlation (Eq. (9.2)), we get $\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e}\left(1 - (r^{\mathbf{X}}_{yz})^2\right)$. The results is obtained by dividing both sides of the previous equation by $\mathbf{y}'\mathbf{M}_0\mathbf{y}$. □

Figure 4.3, below, illustrates the fact that one can obtain an R^2 of one by regressing a sample of length n on any set of n linearly-independent variables.

```
n <- 30;Y <- rnorm(n);X <- matrix(rnorm(n^2),n,n)
all_R2 <- NULL;all_adjR2 <- NULL
for(j in 0:(n-1)){
  if(j==0){eq <- lm(Y~1)}else{eq <- lm(Y~X[,1:j])}
  all_R2 <- c(all_R2,summary(eq)$r.squared)
  all_adjR2 <- c(all_adjR2,summary(eq)$adj.r.squared)
}
par(plt=c(.15,.95,.25,.95))
plot(all_R2,pch=19,ylim=c(min(all_adjR2,na.rm = TRUE),1),
     xlab="number of regressors",ylab="R2")
points(all_adjR2,pch=3);abline(h=0,col="light grey",lwd=2)
legend("topleft",c("R2","Adjusted R2"),
       lty=NaN,col=c("black"),pch=c(19,3),lwd=2)
```

In order to address the risk of adding irrelevant explanatory variables, measures of **adjusted** R^2 have been proposed. Compared to the standard R^2, these measures add penalties that depend on the number of covariates employed in the regression. A common adjusted R^2 measure, denoted by $\bar{R}^2$, is the following:

$$\boxed{\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(n-K)}{\mathbf{y}'\mathbf{M}^0\mathbf{y}/(n-1)} = 1 - \frac{n-1}{n-K}(1-R^2).}$$

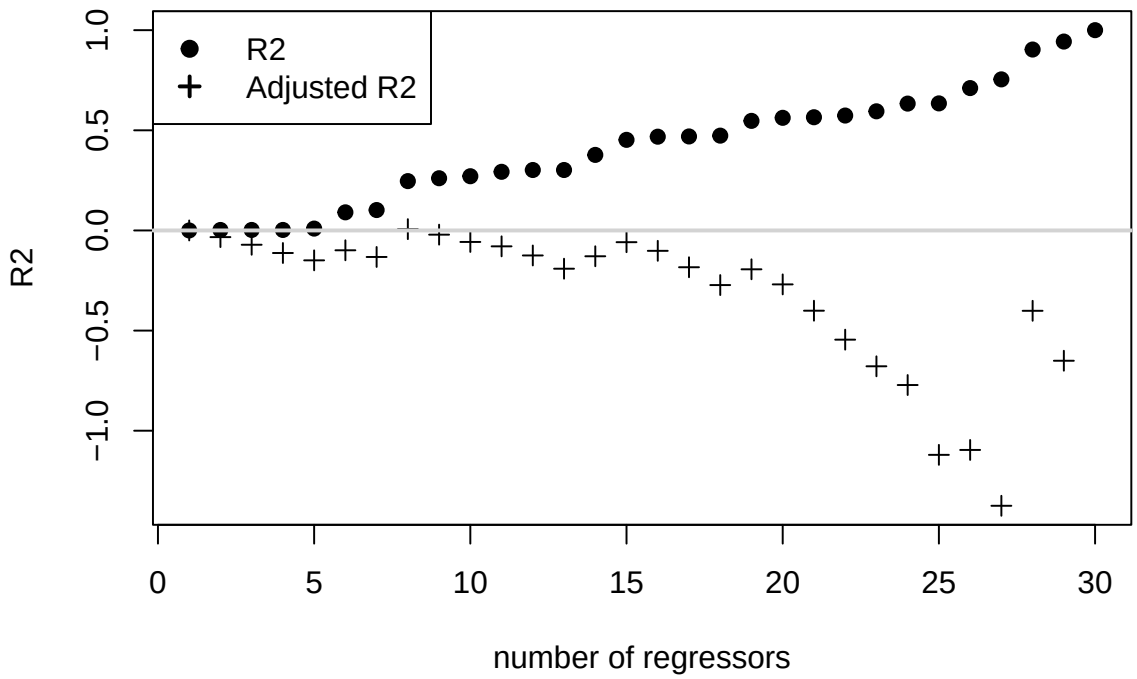


Figure 4.3: This figure illustrates the monotonous increase in the R^2 as a function of the number of explanatory variables. In the true model, there is no explanatory variables, i.e., $y_i = \varepsilon_i$. We then take (independent) regressors and regress y on the latter, progressively increasing the set of regressors.

4.2.4 Inference and confidence intervals (in small sample)

Under the normality assumption (Assumption 4.5), we know the distribution of $\mathbf{b}$ (conditional on $\mathbf{X}$). Indeed, $\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$. Therefore, conditional on $\mathbf{X}$, vector $\mathbf{b}$ is an affine combination of Gaussian variables—the components of ε. As a result, it is also Gaussian. Its distribution is therefore completely characterized by its mean and variance, and we have:

$$\mathbf{b}|\mathbf{X} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \tag{4.9}$$

Eq. (4.9) can be used to conduct inference and tests. However, in practice, we do not know σ^2 (which is a population parameter). The following proposition gives an unbiased estimate of σ^2.

Proposition 4.6. *Under 4.1 to 4.4, an unbiased estimate of σ^2 is given by:*

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}. \tag{4.10}$$

(It is sometimes denoted by σ^2_{OLS}.)

Proof. We have:

$$\begin{aligned} \mathbb{E}(\mathbf{e}'\mathbf{e}|\mathbf{X}) &= \mathbb{E}(\varepsilon'\mathbf{M}\varepsilon|\mathbf{X}) = \mathbb{E}(\mathrm{Tr}(\varepsilon'\mathbf{M}\varepsilon)|\mathbf{X})) \\ &= \mathrm{Tr}(\mathbf{M}\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})) = \sigma^2\mathrm{Tr}(\mathbf{M}). \end{aligned}$$

(Note that we have $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 Id$ by Assumptions 4.3 and 4.4, see Prop. 4.2.) Moreover:

$$\begin{aligned}\text{Tr}(\mathbf{M}) &= n - \text{Tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= n - \text{Tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = n - \text{Tr}(Id_{K\times K}),\end{aligned}$$

which leads to the result. □

Two results will prove important to produce inference:

i. We know the conditional distribution of s^2 (Prop. 4.7).
ii. s^2 and $\mathbf{b}$ are independent random variables (Prop. 4.8).

Proposition 4.7. *Under 4.1 to 4.5, we have:* $\frac{s^2}{\sigma^2}|\mathbf{X} \sim \frac{1}{n-K}\chi^2(n-K)$.

Proof. We have $\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}\varepsilon$. $\mathbf{M}$ is an idempotent symmetric matrix. Therefore it can be decomposed as PDP' where D is a diagonal matrix and P is an orthogonal matrix. As a result $\mathbf{e}'\mathbf{e} = (P'\varepsilon)'D(P'\varepsilon)$, i.e. $\mathbf{e}'\mathbf{e}$ is a weighted sum of independent squared Gaussian variables (the entries of $P'\varepsilon$ are independent because they are Gaussian —under 4.5— and uncorrelated). The variance of each of these i.i.d. Gaussian variable is σ^2. Because $\mathbf{M}$ is an idempotent symmetric matrix, its eigenvalues are either 0 or 1, and its rank equals its trace (see Propositions 9.3 and 9.4). Further, its trace is equal to $n-K$ (see proof of Eq. (4.10)). Therefore D has $n-K$ entries equal to 1 and K equal to 0. Hence, $\mathbf{e}'\mathbf{e} = (P'\varepsilon)'D(P'\varepsilon)$ is a sum of $n-K$ squared independent Gaussian variables of variance σ^2. Therefore $\frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = (n-K)\frac{s^2}{\sigma^2}$ is a sum of $n-k$ squared i.i.d. standard normal variables. The result follows by the definition of the chi-square distribution (see Def. 9.13). □

Proposition 4.8. *Under Hypotheses 4.1 to 4.5,* $\mathbf{b}$ *and* s^2 *are independent.*

Proof. We have $\mathbf{b} = \beta + [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}\varepsilon$ and $s^2 = \varepsilon'\mathbf{M}\varepsilon/(n-K)$. Hence $\mathbf{b}$ is an affine combination of ε and s^2 is a quadratic combination of the same Gaussian shocks. One can write s^2 as $s^2 = (\mathbf{M}\varepsilon)'\mathbf{M}\varepsilon/(n-K)$ and $\mathbf{b}$ as $\beta + \mathbf{T}\varepsilon$. Since $\mathbf{TM} = 0$, $\mathbf{T}\varepsilon$ and $\mathbf{M}\varepsilon$ are independent (because two uncorrelated Gaussian variables are independent), therefore $\mathbf{b}$ and s^2, which are functions of two sets of independent variables, are independent. □

Consistently with Eq. (4.9), under Hypotheses 4.1 to 4.5, the k^{th} entry of $\mathbf{b}$ satisfies:

$$b_k|\mathbf{X} \sim \mathcal{N}(\beta_k, \sigma^2 v_k),$$

where v_k is the kth component of the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$.

Moreover, we have (Prop. 4.7):

$$\frac{(n-K)s^2}{\sigma^2}|\mathbf{X} \sim \chi^2(n-K).$$

As a result (using Propositions 4.7 and 4.8), we have:

$$\boxed{t_k = \frac{\frac{b_k-\beta_k}{\sqrt{\sigma^2 v_k}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2(n-K)}}} = \frac{b_k-\beta_k}{\sqrt{s^2 v_k}} \sim t(n-K),} \tag{4.11}$$

where $t(n-K)$ denotes a Student t distribution with $n-K$ degrees of freedom (see Def. 9.12).[2]

Note that $s^2 v_k$ is not exactly the conditional variance of b_k: The variance of b_k conditional on $\mathbf{X}$ is $\sigma^2 v_k$. However $s^2 v_k$ is an unbiased estimate of $\sigma^2 v_k$ (by Prop. 4.6).

The previous result (Eq. (4.11)) can be extended to any linear combinations of elements of $\mathbf{b}$. (Eq. (4.11) is for its k^{th} component only.) Let us consider $\alpha'\mathbf{b}$, the OLS estimate of $\alpha'\beta$. From Eq. (4.9), we have:

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

Therefore:

$$\frac{\alpha'\mathbf{b}-\alpha'\beta}{\sqrt{\sigma^2\alpha'(\mathbf{X}'\mathbf{X})^{-1}\alpha}}|\mathbf{X} \sim \mathcal{N}(0,1).$$

Using the same approach as the one used to derive Eq. (4.11), one can show that Props. 4.7 and 4.8 also imply that:

$$\boxed{\frac{\alpha'\mathbf{b}-\alpha'\beta}{\sqrt{s^2\alpha'(\mathbf{X}'\mathbf{X})^{-1}\alpha}} \sim t(n-K).} \tag{4.12}$$

What precedes is widely exploited for statistical inference in the context of linear regressions. Indeed, Eq. (4.11) gives a sense of the distances between b_k and β_k that can be deemed as "likely" (or, conversely, "unlikely"). For instance, it implies that, if $\sqrt{v_k s^2}$ is equal to 1 (say), then the probability to obtain b_k smaller than $\beta_k - 4.587 \times \sqrt{v_k s^2}$ or larger than $\beta_k + 4.587 \times \sqrt{v_k s^2}$ is equal to 0.1% when $n-K=10$.

That means for instance that, under the assumption that $\beta_k = 0$, it would be extremely unlikely to have obtained $b_k/\sqrt{v_k s^2}$ smaller than -4.587 or larger than 4.587. More generally, this shows that the **t-statistic**, i.e., the ratio $b_k/\sqrt{v_k s^2}$, is the test statistic associated with the null hypothesis:

$$H_0 : \beta_k = 0.$$

Under the null hypothesis, the test statistic follows a Student-t distribution with $n-K$ degrees of freedom. The **t-statistic** is therefore of particular importance, and, as a result, it is routinely reported in regression outputs (see Example 4.2).

Example 4.2 (Education and income)**.** Consider regression that aims at determining covariates of households' income. This example makes use of data from the Swiss Household Panel (SHP); `edyear19` is the number of years of education and `age19` is the age of the respondent, as of 2019.

[2] We have $\frac{b_k-\beta_k}{\sqrt{\sigma^2 v_k}}|\mathbf{X} \sim \mathcal{N}(0,1)$ and $\frac{(n-K)s^2}{\sigma^2}|\mathbf{X} \sim \chi^2(n-K)$. These two distributions do not depend on $\mathbf{X} \Rightarrow$ the *marginal distribution* of t_k is also t.

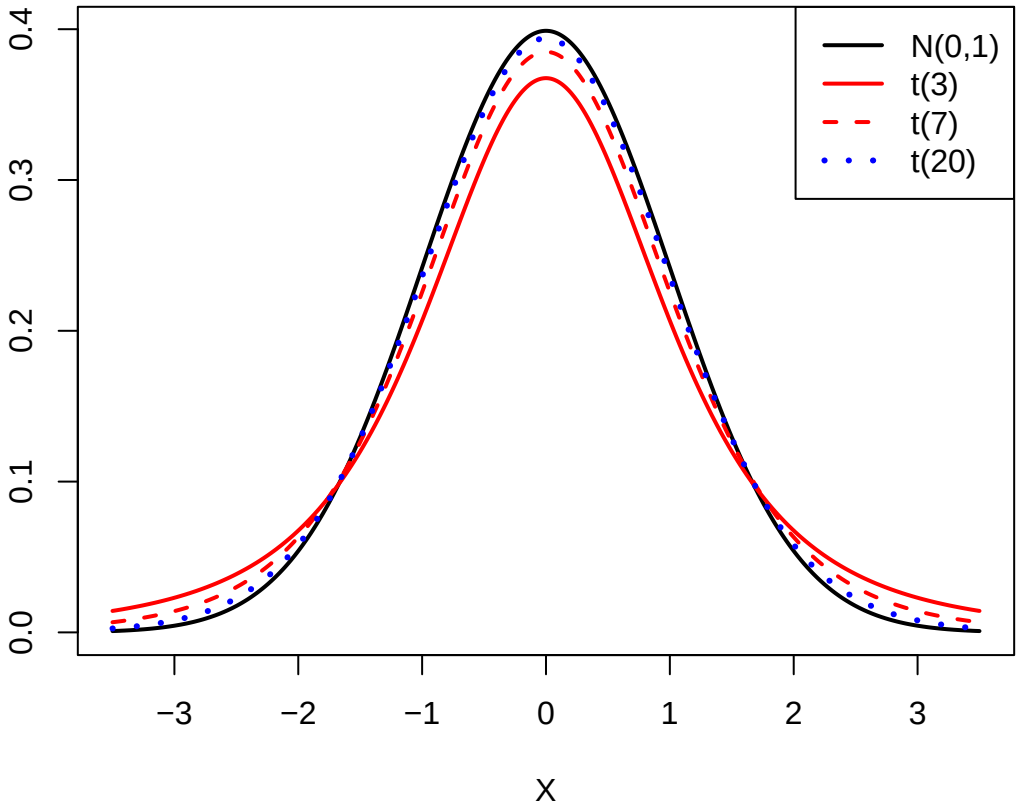


Figure 4.4: The higher the degree of freedom, the closer the distribution of $t(\nu)$ gets to the normal distribution. (Convergence in distribution.)

```
library(AEC)
library(sandwich)
shp$income <- shp$i19ptotn/1000
shp$female <- 1*(shp$sex19==2)
eq <- lm(income ~ edyear19 + age19 + I(age19^2) + female,data=shp)
lmtest::coeftest(eq)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -71.9738073   5.7082456 -12.609 < 2.2e-16 ***
## edyear19      4.8442661   0.2172320  22.300 < 2.2e-16 ***
## age19         3.2386215   0.2183812  14.830 < 2.2e-16 ***
## I(age19^2)   -0.0289498   0.0020915 -13.842 < 2.2e-16 ***
## female      -31.8089006   1.4578004 -21.820 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The last two columns of the previous table give the t-statistic and the p-values associated with t-tests, whose size-α critical region is:

$$\left]-\infty,-\Phi^{-1}_{t(n-K)}\left(1-\frac{\alpha}{2}\right)\right]\cup\left[\Phi^{-1}_{t(n-K)}\left(1-\frac{\alpha}{2}\right),+\infty\right[.$$

We recall that the **p-value** is defined as the probability that $|Z|>|t|$, where t is the (computed) t-statistics and where $Z\sim t(n-K)$. That is, in the context of the t-test, the p-value is given by $2(1-\Phi_{t(n-K)}(|t_k|))$. See this webpage for details regarding the link between critical regions, p-value, and test outcomes.

Now, suppose we want to compute a (symmetrical) *confidence interval* $[I_{d,1-\alpha},I_{u,1-\alpha}]$ that is such that $\mathbb{P}(\beta_k\in[I_{d,1-\alpha},I_{u,1-\alpha}])=1-\alpha$. That is, we want to have: $\mathbb{P}(\beta_k<I_{d,1-\alpha})=\frac{\alpha}{2}$ and $\mathbb{P}(\beta_k>I_{u,1-\alpha})=\frac{\alpha}{2}$. Let us focus on $I_{d,1-\alpha}$ to start with. Using Eq. (4.11), i.e., $t_k=\frac{b_k-\beta_k}{\sqrt{s^2v_k}}\sim t(n-K)$, we have:

$$\begin{aligned}
\mathbb{P}(\beta_k<I_{d,1-\alpha})=\frac{\alpha}{2}\quad&\Leftrightarrow\\
\mathbb{P}\left(\frac{b_k-\beta_k}{\sqrt{s^2v_k}}>\frac{b_k-I_{d,1-\alpha}}{\sqrt{s^2v_k}}\right)=\frac{\alpha}{2}\quad&\Leftrightarrow\quad\mathbb{P}\left(t_k>\frac{b_k-I_{d,1-\alpha}}{\sqrt{s^2v_k}}\right)=\frac{\alpha}{2}\Leftrightarrow\\
1-\mathbb{P}\left(t_k\leq\frac{b_k-I_{d,1-\alpha}}{\sqrt{s^2v_k}}\right)=\frac{\alpha}{2}\quad&\Leftrightarrow\quad\frac{b_k-I_{d,1-\alpha}}{\sqrt{s^2v_k}}=\Phi^{-1}_{t(n-K)}\left(1-\frac{\alpha}{2}\right),
\end{aligned}$$

where $\Phi_{t(n-K)}(\alpha)$ is the c.d.f. of the $t(n-K)$ distribution (Table 9.2).

Doing the same for $I_{u,1-\alpha}$, we obtain:

$$[I_{d,1-\alpha}, I_{u,1-\alpha}] = \left[b_k - \Phi^{-1}_{t(n-K)}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2 v_k}, b_k + \Phi^{-1}_{t(n-K)}\left(1-\frac{\alpha}{2}\right)\sqrt{s^2 v_k}\right].$$

Using the results presented in Example 4.2, we can compute lower and upper bounds of 95% confidence intervals for the estimated parameters as follows:

```
n <- length(eq$residuals); K <- length(eq$coefficients)
lower.b <- eq$coefficients + qt(.025,df=n-K)*sqrt(diag(vcov(eq)))
upper.b <- eq$coefficients + qt(.975,df=n-K)*sqrt(diag(vcov(eq)))
cbind(lower.b,upper.b)
```

```
##                   lower.b      upper.b
## (Intercept) -83.16413225 -60.78348237
## edyear19      4.41840914   5.27012310
## age19         2.81051152   3.66673148
## I(age19^2)   -0.03304986  -0.02484977
## female      -34.66674188 -28.95105932
```

4.2.5 Testing a set of linear restrictions

We sometimes want to test if a set of restrictions are *jointly* consistent with the data at hand. Let us formalize such a set of (J) linear restrictions:

$$\begin{array}{rcl} r_{1,1}\beta_1 + \cdots + r_{1,K}\beta_K &=& q_1 \\ \vdots & & \vdots \\ r_{J,1}\beta_1 + \cdots + r_{J,K}\beta_K &=& q_J. \end{array} \tag{4.13}$$

In matrix form, we get:

$$\mathbf{R}\beta = \mathbf{q}. \tag{4.14}$$

Define the *discrepancy vector* $\mathbf{m} = \mathbf{Rb} - \mathbf{q}$. Under the null hypothesis:

$$\begin{array}{rcl} \mathbb{E}(\mathbf{m}|\mathbf{X}) &=& \mathbf{R}\beta - \mathbf{q} = 0 \quad \text{and} \\ \mathbb{V}ar(\mathbf{m}|\mathbf{X}) &=& \mathbf{R}\mathbb{V}ar(\mathbf{b}|\mathbf{X})\mathbf{R}'. \end{array}$$

With these notations, the assumption to test is:

$$\boxed{H_0 : \mathbf{R}\beta - \mathbf{q} = 0 \text{ against } H_1 : \mathbf{R}\beta - \mathbf{q} \neq 0.} \tag{4.15}$$

Under Hypotheses 4.1 to 4.4, we have $\mathbb{V}ar(\mathbf{m}|\mathbf{X}) = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ (see Prop. 4.3). If we add the normality assumption (Hypothesis 4.5), we have:

$$W = \mathbf{m}'\mathbb{V}ar(\mathbf{m}|\mathbf{X})^{-1}\mathbf{m} \sim \chi^2(J). \tag{4.16}$$

If σ^2 was known, we could then conduct a *Wald test* (directly exploiting Eq. (4.16)). But this is not the case in practice and we cannot compute W. We can, however, approximate it be replacing σ^2 by s^2 (given in Eq. (4.10)). The distribution of this new statistic is not $\chi^2(J)$ any more; it is an $\mathcal{F}$ *distribution* (whose quantiles are shown in Table 9.4), and the test is called F *test*.

Proposition 4.9. *Under Hypotheses 4.1 to 4.5 and if Eq.* (4.15) *holds, we have:*

$$F = \frac{W}{J}\frac{\sigma^2}{s^2} = \frac{\mathbf{m}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{m}}{s^2 J} \sim \mathcal{F}(J, n-K), \tag{4.17}$$

where $\mathcal{F}$ is the distribution of the F-statistic (see Def. 9.11).

Proof. According to Eq. (4.16), $W/J \sim \chi^2(J)/J$. Moreover, the denominator (s^2/σ^2) is $\sim \chi^2(n-K)$. Therefore, F is the ratio of a r.v. distributed as $\chi^2(J)/J$ and another distributed as $\chi^2(n-K)/(n-K)$. It remains to verify that these r.v. are independent. Under H_0, we have $\mathbf{m} = \mathbf{R}(\mathbf{b}-\beta) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$. Therefore $\mathbf{m}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{m}$ is of the form $\varepsilon'\mathbf{T}\varepsilon$ with $\mathbf{T} = \mathbf{D}'\mathbf{C}\mathbf{D}$ where $\mathbf{D} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{C} = (\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}$. Under Hypotheses 4.1 to 4.4, the covariance between $\mathbf{T}\varepsilon$ and $\mathbf{M}\varepsilon$ is $\sigma^2\mathbf{T}\mathbf{M} = \mathbf{0}$. Therefore, under 4.5, these variables are Gaussian variables with 0 covariance. Hence they are independent. □

For large $n-K$, the $\mathcal{F}_{J,n-K}$ distribution converges to $\mathcal{F}_{J,\infty} = \chi^2(J)/J$. This implies that, in large samples, the F-statistic approximately has a χ^2 distribution. In other words, one can approximately employ Eq. (4.16) to perform a Wald test (one just has to replace σ^2 with s^2 when computing $\mathbb{V}ar(\mathbf{m}|\mathbf{X})$).

The following proposition proposes another equivalent computation of the F-statistic, based on the R^2 of the restricted and unrestricted linear models.

Proposition 4.10. *The F-statistic defined by Eq.* (4.17) *is also equal to:*

$$F = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)} = \frac{(SSR_{restr} - SSR_{unrestr})/J}{SSR_{unrestr}/(n-K)}, \tag{4.18}$$

where R_*^2 is the coef. of determination (Eq. (4.6)*) of the "restricted regression"* (SSR: sum of squared residuals.)

Proof. Let's denote by $\mathbf{e}_* = \mathbf{y} - \mathbf{X}\mathbf{b}_*$ the vector of residuals associated to the *restricted regression* (i.e. $\mathbf{R}\mathbf{b}_* = \mathbf{q}$). We have $\mathbf{e}_* = \mathbf{e} - \mathbf{X}(\mathbf{b}_* - \mathbf{b})$. Using $\mathbf{e}'\mathbf{X} = 0$, we get $\mathbf{e}_*'\mathbf{e}_* = \mathbf{e}'\mathbf{e} + (\mathbf{b}_* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}_* - \mathbf{b}) \geq \mathbf{e}'\mathbf{e}$.

By Proposition 9.5 (in Appendix 9.2), we have: $\mathbf{b}_* - \mathbf{b} = -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$. Therefore:

$$\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e} = (\mathbf{R}\mathbf{b} - \mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}).$$

This implies that the F statistic defined in Prop. 4.9 is also equal to:

$$\frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n-K)},$$

which leads to the result. □

The null hypothesis H_0 (Eq. (4.15)) of the F-test is rejected if F —defined by Eq. (4.17) or (4.18)— is higher than $\mathcal{F}_{1-\alpha}(J, n-K)$. (Hence, this test is a one-sided test.)

4.2.6 Large Sample Properties

Even if we relax the normality assumption (Hypothesis 4.5), we can approximate the finite-sample behavior of the estimators by using *large-sample* or *asymptotic properties*.

To begin with, we proceed under Hypothesis 4.1 to 4.4. (We will see, later on, how to deal with —partial— relaxations of Hypothesis 4.3 and 4.4.)

Under regularity assumptions, and under Hypotheses 4.1 to 4.4, even if the residuals are not normally-distributed, the least square estimators can be *asymptotically normal* and inference can be performed in the same way as in small samples when Hypotheses 4.1 to 4.5 hold. This derives from Prop. 4.11 (below). The F-test (Prop. 4.10) and the t-test (Eq. (4.11)) can then be performed.

Proposition 4.11. *Under Assumptions 4.1 to 4.4, and assuming further that:*

$$Q = plim_{n\to\infty} \frac{\mathbf{X}'\mathbf{X}}{n}, \tag{4.19}$$

and that the $(\mathbf{x}_i, \varepsilon_i)$'s are independent (across entities i), we have:

$$\sqrt{n}(\mathbf{b} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 Q^{-1}\right). \tag{4.20}$$

Proof. Since $\mathbf{b} = \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$, we have: $\sqrt{n}(\mathbf{b} - \beta) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon$. Since $f : A \to A^{-1}$ is a continuous function (for $A \neq \mathbf{0}$), $\text{plim}_{n\to\infty}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = \mathbf{Q}^{-1}$ (see Prop. 9.12). Let us denote by V_i the vector $\mathbf{x}_i\varepsilon_i$. Because the $(\mathbf{x}_i, \varepsilon_i)$'s are independent, the V_i's are independent as well. Their covariance matrix is $\sigma^2\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') = \sigma^2 Q$. Applying the multivariate central limit theorem on vectors V_i gives $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i\right) = \left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q)$. An application of Slutsky's theorem (Prop. 9.12) then leads to the results. □

In practice, σ^2 is approximated by $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$ (Eq. (4.10)) and $\mathbf{Q}^{-1}$ by $\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$. That is, the covariance matrix of the estimator is approximated by:

$$\boxed{\widehat{\mathbb{V}ar}(\mathbf{b}) = s^2(\mathbf{X}'\mathbf{X})^{-1}.} \tag{4.21}$$

Eqs. (4.19) and (4.20) respectively correspond to convergences in probability and in distribution (see Definitions 9.16 and 9.19, respectively).

4.3 Common pitfalls in linear regressions

4.3.1 Multicollinearity

Consider the model: $y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$, where all variables are zero-mean and $\mathbb{V}ar(\varepsilon_i) = \sigma^2$. We have:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum_i x_{i,1}^2 & \sum_i x_{i,1}x_{i,2} \\ \sum_i x_{i,1}x_{i,2} & \sum_i x_{i,2}^2 \end{bmatrix},$$

therefore:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum_i x_{i,1}^2 \sum_i x_{i,2}^2 - (\sum_i x_{i,1}x_{i,2})^2} \begin{bmatrix} \sum_i x_{i,2}^2 & -\sum_i x_{i,1}x_{i,2} \\ -\sum_i x_{i,1}x_{i,2} & \sum_i x_{i,1}^2 \end{bmatrix}.$$

The inverse of the upper-left parameter of $(\mathbf{X}'\mathbf{X})^{-1}$ is:

$$\sum_i x_{i,1}^2 - \frac{(\sum_i x_{i,1}x_{i,2})^2}{\sum_i x_{i,2}^2} = \sum_i x_{i,1}^2(1 - correl_{1,2}^2), \tag{4.22}$$

where $correl_{1,2}$ is the sample correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$.

Hence, the closer to one $correl_{1,2}$, the higher the variance of b_1 (recall that the variance of b_1 is the upper-left component of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$). That is, if some of our regressors are close to a linear conbination of the other ones, then the confidence intervals will tend to be wide, which typically reduces the power of the t-test (we tend to fail to reject the null hypothesis that the coefficients are different from zero).

4.3.2 Omitted variables

Consider the following model (the "True model"):

$$\mathbf{y} = \underbrace{\mathbf{X}_1}_{n\times K_1} \underbrace{\beta_1}_{K_1\times 1} + \underbrace{\mathbf{X}_2}_{n\times K_2} \underbrace{\beta_2}_{K_2\times 1} + \varepsilon$$

If one computes $\mathbf{b}_1$ by regressing $\mathbf{y}$ on $\mathbf{X}_1$ only, one gets:

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon.$$

This results in the omitted-variable formula:

$$\mathbb{E}(\mathbf{b}_1|\mathbf{X}) = \beta_1 + \underbrace{(\mathbf{X}_1'\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{X}_2)}_{K_1\times K_2}\beta_2.$$

(Each column of $(\mathbf{X}_1'\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{X}_2)$ are the OLS regressors obtained when regressing the columns of $\mathbf{X}_2$ on $\mathbf{X}_1$.) Unless the variables included in $\mathbf{X}_1$ are orthogonal to those in $\mathbf{X}_2$, we obtain a bias. A way to address this potential pitfall is to introduce "controls" in the specification.

Example 4.3. Let us use the California Test Score dataset (in the package `AER`). Assume we want to measure the effect of the students-to-teacher ratio (`str`) on student test scores (`testscr`). The following regressions show that the effect is lower when controls are added.

```
library(AER); data("CASchools")
CASchools$str <- CASchools$students/CASchools$teachers
CASchools$testscr <- .5 * (CASchools$math + CASchools$read)
eq1 <- lm(testscr~str,data=CASchools)
eq2 <- lm(testscr~str+lunch,data=CASchools)
eq3 <- lm(testscr~str+lunch+english,data=CASchools)
stargazer::stargazer(eq1,eq2,eq3,type="text",
                     no.space = TRUE,omit.stat=c("f","ser"))
```

```
##
## ================================================
##                        Dependent variable:
##                    ----------------------------
##                               testscr
##                      (1)         (2)         (3)
## ------------------------------------------------
## str                -2.280***   -1.117***   -0.998***
##                     (0.480)     (0.240)     (0.239)
## lunch                          -0.600***   -0.547***
##                                 (0.017)     (0.022)
## english                                    -0.122***
##                                             (0.032)
## Constant          698.933***  702.911***  700.150***
##                     (9.467)     (4.700)     (4.686)
## ------------------------------------------------
## Observations         420         420         420
## R2                  0.051       0.767       0.775
## Adjusted R2         0.049       0.766       0.773
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

4.3.3 Irrelevant variable

Consider the *true model*:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \varepsilon,$$

while the *estimated model* is:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$$

The estimates are unbiased. However, adding irrelevant explanatory variables increases the variance of the estimate of β_1 (compared to the case where one uses the correct explanatory variables). This is the case unless the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ is null, see Eq. (4.22).

In other words, the estimator is *inefficient*, i.e., there exists an alternative consistent estimator whose variance is lower. The inefficiency problem can have serious consequences when testing hypotheses such as $H_0 : \beta_1 = 0$. Due to the loss of power, we might wrongly infer that the $\mathbf{X}_1$ variables are not "relevant" (*Type-II error, False Negative*).

4.4 Instrumental Variables

The conditional mean zero assumption (Hypothesis 4.2), according to which $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$ —which implies in particular that $\mathbf{x}_i$ and ε_i are uncorrelated— is sometimes not consistent with the considered economic framework. When it is the case, the parameters of interest may still be estimated consistently by resorting to instrumental variable techniques.

Consider the following model:

$$y_i = \mathbf{x_i}'\beta + \varepsilon_i, \quad \text{where } \mathbb{E}(\varepsilon_i) = 0 \text{ and } \mathbf{x_i} \not\perp \varepsilon_i. \tag{4.23}$$

Let us illustrate how this situation may result in biased OLS estimate. Consider for instance the situation where:

$$\mathbb{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathbb{E}(\varepsilon_i \mathbf{x_i}) = \gamma, \tag{4.24}$$

in which case we have $\mathbf{x}_i \not\perp \varepsilon_i$ (consistently with Eq. (4.23)).

By the law of large numbers, $\text{plim}_{n\to\infty} \mathbf{X}'\varepsilon/n = \gamma$. If $\mathbf{Q}_{xx} := \text{plim } \mathbf{X}'\mathbf{X}/n$, the OLS estimator is not consistent because

$$\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \xrightarrow{p} \beta + \mathbf{Q}_{xx}^{-1}\gamma \neq \beta.$$

Let us now introduce the notion of instruments.

Definition 4.2 (Instrumental variables)**.** The L-dimensional random variable $\mathbf{z}_i$ is a valid set of instruments if:

a. $\mathbf{z}_i$ is correlated to $\mathbf{x}_i$;
b. we have $\mathbb{E}(\varepsilon|\mathbf{Z}) = 0$ and
c. the orthogonal projections of the $\mathbf{x}_i$'s on the $\mathbf{z}_i$'s are not multicollinear.

Point c implies in particular that the dimension of $\mathbf{z}_i$ has to be at least as large as that of $\mathbf{x}_i$. If $\mathbf{z}_i$ is a valid set of instruments, we have:

$$\text{plim}\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \text{plim}\left(\frac{\mathbf{Z}'(\mathbf{X}\beta + \varepsilon)}{n}\right) = \text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\beta.$$

Indeed, by the law of large numbers, $\frac{\mathbf{Z}'\varepsilon}{n} \xrightarrow{p} \mathbb{E}(\mathbf{z}_i\varepsilon_i) = 0$.

If $L = K$, the matrix $\frac{\mathbf{Z}'\mathbf{X}}{n}$ is of dimension $K \times K$ and we have:

$$\left[\text{plim } \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1} \text{plim } \left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \beta.$$

By continuity of the inverse function (everywhere but at 0): $\left[\text{plim } \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1} = \text{plim } \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}$. The Slutsky Theorem (Prop. 9.12) further implies that:

$$\text{plim } \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \text{plim } \left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = \text{plim } \left(\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \frac{\mathbf{Z}'\mathbf{y}}{n}\right).$$

Hence $\mathbf{b}_{iv}$ is consistent if it is defined by:

$$\boxed{\mathbf{b}_{iv} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}.}$$

Proposition 4.12 (Asymptotic distribution of the IV estimator)**.** *If $\mathbf{z}_i$ is a L-dimensional random variable that constitutes a valid set of instruments (see Def. 4.2) and if $L = K$, then the asymptotic distribution of $\mathbf{b}_{iv}$ is:*

$$\mathbf{b}_{iv} \xrightarrow{d} \mathcal{N}\left(\beta, \frac{\sigma^2}{n}\left[Q_{xz}Q_{zz}^{-1}Q_{zx}\right]^{-1}\right)$$

where plim $\mathbf{Z}'\mathbf{Z}/n =: \mathbf{Q}_{zz}$, plim $\mathbf{Z}'\mathbf{X}/n =: \mathbf{Q}_{zx}$, plim $\mathbf{X}'\mathbf{Z}/n =: \mathbf{Q}_{xz}$.

Proof. The proof is very similar to that of Prop. 4.11, the starting point being that $\mathbf{b}_{iv} = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\varepsilon$. □

When $L = K$, we have:

$$\left[Q_{xz}Q_{zz}^{-1}Q_{zx}\right]^{-1} = Q_{zx}^{-1}Q_{zz}Q_{xz}^{-1}.$$

In practice, to estimate $\mathbb{V}ar(\mathbf{b}_{iv}) = \frac{\sigma^2}{n}Q_{zx}^{-1}Q_{zz}Q_{xz}^{-1}$, we replace σ^2 by:

$$s_{iv}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b}_{iv})^2.$$

What about when $L > K$? In this case, we proceed as follows:

1. Regress $\mathbf{X}$ on the space spanned by $\mathbf{Z}$ and
2. Regress $\mathbf{y}$ on the fitted values $\hat{\mathbf{X}} := \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$.

These two-step approach is called **Two-Stage Least Squares (2SLS)**. It results in:

$$\boxed{\mathbf{b}_{iv} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}.} \tag{4.25}$$

In this case, Prop. 4.12 still holds, with $\mathbf{b}_{iv}$ given by Eq. (4.25).

If the instruments do not properly satisfy Condition (a) in Def. 4.2 (i.e. if $\mathbf{x}_i$ and $\mathbf{z}_i$ are only loosely related), the instruments are said to be **weak** (see, e.g., Stock and Yogo (2005), available here or Andrews et al. (2019)). A simple standard way to test for weak instruments consist in looking at the F-statistic associated with the first stage of the estimation. The easier it is to reject the null hypothesis (large test statistic), the less weak —or the stronger— the instruments.

The Durbin-Wu-Hausman test (Durbin (1954), Wu (1973), Hausman (1978)) can be used to test if IV necessary. (IV techniques are required if $\text{plim}_{n\to\infty}\mathbf{X}'\varepsilon/n \neq 0$.) Hausman (1978) proposes a test of the efficiency of estimators. Under the null hypothesis two estimators, $\mathbf{b}_0$ and $\mathbf{b}_1$, are consistent but $\mathbf{b}_0$ is (asymptotically) efficient relative to $\mathbf{b}_1$. Under the alternative hypothesis, $\mathbf{b}_1$ (IV in the present case) remains consistent but not $\mathbf{b}_0$ (OLS in the present case). That is, when we reject the null hypothesis, it means that the OLS estimator is not consistent, potentially due to endogeneity issue.

The test statistic is:

$$H = (\mathbf{b}_1 - \mathbf{b}_0)' MPI(\mathbb{V}ar(\mathbf{b}_1) - \mathbb{V}ar(\mathbf{b}_0))(\mathbf{b}_1 - \mathbf{b}_0),$$

where MPI is the Moore-Penrose pseudo-inverse. Under the null hypothesis, $H \sim \chi^2(q)$, where q is the rank of $\mathbb{V}ar(\mathbf{b}_1) - \mathbb{V}ar(\mathbf{b}_0)$.

Example 4.4 (Estimation of price elasticity)**.** See e.g. WHO and estimation of tobacco price elasticity of demand.

We want to estimate what is the effect on demand of an *exogenous increase* in prices of cigarettes (say).

The model is:

$$\begin{array}{rcl} \underbrace{q_t^d}_{\log(\text{demand})} & = & \alpha_0 + \alpha_1 \underbrace{\times p_t}_{\log(\text{price})} + \alpha_2 \underbrace{\times w_t}_{\text{income}} + \varepsilon_t^d \\ \underbrace{q_t^s}_{\log(\text{supply})} & = & \gamma_0 + \gamma_1 \times p_t + \gamma_2 \underbrace{\times \mathbf{y}_t}_{\text{cost factors}} + \varepsilon_t^s, \end{array}$$

where $\mathbf{y}_t$, w_t, $\varepsilon_t^s \sim \mathcal{N}(0, \sigma_s^2)$ and $\varepsilon_t^d \sim \mathcal{N}(0, \sigma_d^2)$ are independent.

Equilibrium: $q_t^d = q_t^s$. This implies that prices are **endogenous**:

$$p_t = \frac{\alpha_0 + \alpha_2 w_t + \varepsilon_t^d - \gamma_0 - \gamma_2 \mathbf{y}_t - \varepsilon_t^s}{\gamma_1 - \alpha_1}.$$

In particular we have $\mathbb{E}(p_t \varepsilon_t^d) = \frac{\sigma_d^2}{\gamma_1 - \alpha_1} \neq 0 \Rightarrow$ Regressing by OLS q_t^d on p_t gives biased estimates (see Eq. (4.24)).

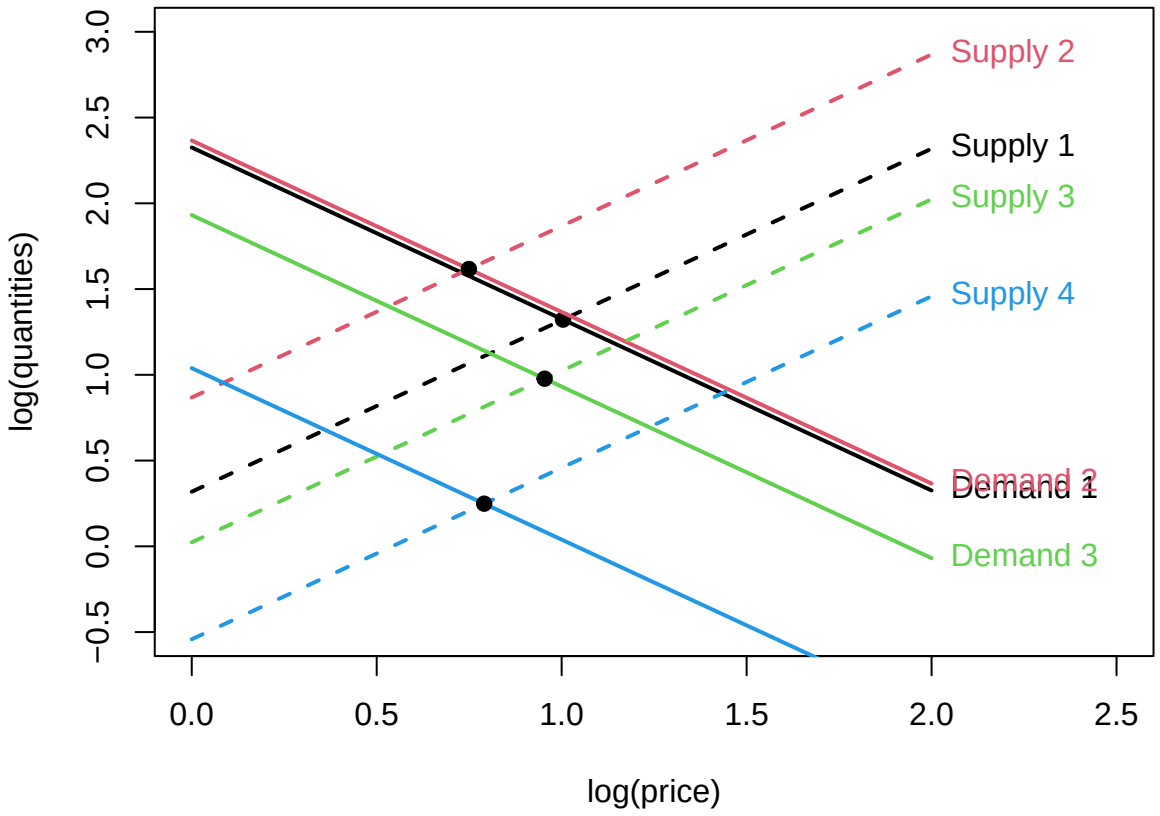


Figure 4.5: This figure illustrates the situation prevailing when estimating a price-elasticity (and the price is endogenous).

Let us use IV regressions to estimate the price elasticity of cigarette demand. For that purpose, we use the `CigarettesSW` dataset of package `AER` (these data are used by Stock and Watson (2003)). This panel dataset documents cigarette consumption for the 48 continental US States from 1985–1995. The instrument is the real tax on cigarettes arising from the state's general sales tax. The rationale is that larger general sales tax drives cigarette prices up, but the general tax is not determined by other forces affecting ε_t^d.

```
data("CigarettesSW", package = "AER")
CigarettesSW$rprice  <- with(CigarettesSW, price/cpi)
CigarettesSW$rincome <- with(CigarettesSW, income/population/cpi)
CigarettesSW$tdiff   <- with(CigarettesSW, (taxs - tax)/cpi)

## model
eq.IV1 <- ivreg(log(packs) ~ log(rprice) + log(rincome) |
                  log(rincome) + tdiff + I(tax/cpi),
                data = CigarettesSW, subset = year == "1995")
eq.IV2 <- ivreg(log(packs) ~ log(rprice) | tdiff,
                data = CigarettesSW, subset = year == "1995")
eq.no.IV <- lm(log(packs) ~ log(rprice) + log(rincome),
               data = CigarettesSW, subset = year == "1995")
stargazer::stargazer(eq.no.IV,eq.IV1,eq.IV2,type="text",no.space = TRUE,
                     omit.stat=c("f","ser"))
```

```
##
## ============================================
##                     Dependent variable:
##                 ---------------------------
##                           log(packs)
##                   OLS       instrumental
##                                variable
##                   (1)       (2)       (3)
## --------------------------------------------
## log(rprice)   -1.407*** -1.277*** -1.084***
##                (0.251)   (0.263)   (0.317)
## log(rincome)    0.344     0.280
##                (0.235)   (0.239)
## Constant      10.342*** 9.895***  9.720***
##                (1.023)   (1.059)   (1.514)
## --------------------------------------------
## Observations     48        48        48
## R2              0.433     0.429     0.401
## Adjusted R2     0.408     0.404     0.388
## ============================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

```
summary(eq.IV1,diagnostics = TRUE)$diagnostics
```

```
##                   df1 df2   statistic      p-value
## Weak instruments   2  44 244.7337536 1.444054e-24
## Wu-Hausman         1  44   3.0678163 8.682505e-02
## Sargan             1  NA   0.3326221 5.641191e-01
```

The last three tests are interpreted as follows:

- Since the p-value of the first test is small, we reject the null hypothesis according to which the instrument is weak.
- The small p-value of the Wu-Hausman test implies that we reject the null hypothesis according to which the OLS estimates are consistent (at the 10% level only, though).
- No over-identification (misspecification) is detected by the Sargan test (large p-value).

Example 4.5 (Education and wage)**.** In this example, we make use of another dataset proposed by Stock and Watson (2003), namely the `CollegeDistance` dataset.[3] the objective is to estimate the effect of education on wages. Education choice is suspected to be an endogenous variable, which calls for an IV strategy. The instrumental variable is the distance to college (see, e.g., Dee (2004)).

```
library(sem)
data("CollegeDistance", package = "AER")
eq.1st.stage <- lm(education ~ urban + gender + ethnicity + unemp + distance,
                   data = CollegeDistance)
CollegeDistance$ed.pred<- predict(eq.1st.stage)
eq.2nd.stage <- lm(wage ~ urban + gender + ethnicity + unemp + ed.pred,
                   data = CollegeDistance)
eqOLS <- lm(wage ~ urban + gender + ethnicity + unemp + education,
            data=CollegeDistance)
eq2SLS <- ivreg(wage ~ urban + gender + ethnicity + unemp + education|
                  urban + gender + ethnicity + unemp + distance,
                data=CollegeDistance)
stargazer::stargazer(eq.1st.stage,eq.2nd.stage,eq2SLS,eqOLS,
                     type="text",no.space = TRUE,omit.stat = c("f","ser"))
```

```
##
## ============================================================
##                               Dependent variable:
##                   ------------------------------------------
```

[3]Cross-section data from the High School and Beyond survey conducted by the Department of Education in the 80s. The survey includes students from approximately 1,100 high schools.

```
##                    education              wage
##                       OLS       OLS    instrumental    OLS
##                                          variable
##                        (1)       (2)        (3)         (4)
## ------------------------------------------------------------
## urbanyes             -0.092     0.046      0.046       0.070
##                      (0.065)   (0.045)    (0.060)     (0.045)
## genderfemale         -0.025    -0.071*     -0.071    -0.085**
##                      (0.052)   (0.037)    (0.050)     (0.037)
## ethnicityafam      -0.524*** -0.227***   -0.227**   -0.556***
##                      (0.072)   (0.073)    (0.099)     (0.052)
## ethnicityhispanic -0.275*** -0.351***  -0.351***   -0.544***
##                      (0.068)   (0.057)    (0.077)     (0.049)
## unemp                 0.010   0.139***    0.139***   0.133***
##                      (0.010)   (0.007)    (0.009)     (0.007)
## distance           -0.087***
##                      (0.012)
## ed.pred                       0.647***
##                               (0.101)
## education                                 0.647***     0.005
##                                           (0.136)     (0.010)
## Constant           14.061***  -0.359      -0.359    8.641***
##                      (0.083)   (1.412)    (1.908)     (0.157)
## ------------------------------------------------------------
## Observations         4,739     4,739      4,739       4,739
## R2                   0.023     0.117      -0.612      0.110
## Adjusted R2          0.022     0.116      -0.614      0.109
## ============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

4.5 General Regression Model (GRM) and robust covariance matrices

The statistical inference presented above relies on strong assumptions regarding the stochastic properties of the errors. Namely, they are assumed to be mutually uncorrelated (Hypothesis 4.4) and homoskedastic (Hypothesis 4.3).

The objective of this section is to present approaches aimed at adjusting the estimate of the covariance matrix of the OLS estimator ($(\mathbf{X}'\mathbf{X})^{-1}s^2$, see Eq. (4.21)), when the previous hypotheses do not hold.

4.5.1 Presentation of the General Regression Model (GRM)

It will prove useful to introduce the following notation:

$$\mathbb{V}ar(\varepsilon|\mathbf{X}) = \mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) \quad = \quad \Sigma. \tag{4.26}$$

Note that Eq. (4.26) is more general than Hypothesis 4.3 and 4.4 because the diagonal entries of Σ may be different (as opposed to under Hypothesis 4.3), and the non-diagonal entries of Σ can be non-null (as opposed to under Hypothesis 4.4).

Definition 4.3 (General Regression Model (GRM))**.** Hypothesis 4.1 and 4.2, together with Eq. (4.26), form the General Regression Model (GRM) framework.

Naturally, a regression model where Hypotheses 4.1 to 4.4 hold is a specific case of the GRM framework.

The GRM context notably encompasses situations of heteroskedasticity and autocorrelation:

- Heteroskedasticity:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & ... & 0 & \sigma_n^2 \end{bmatrix}. \tag{4.27}$$

- Autocorrelation:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_{2,1} & ... & \rho_{n,1} \\ \rho_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \rho_{n,n-1} \\ \rho_{n,1} & \rho_{n,2} & ... & 1 \end{bmatrix}. \tag{4.28}$$

Example 4.6 (Auto-regressive processes)**.** Autocorrelation is common in time-series contexts (see Section 8). In a time-series context, subscript i refers to a date.

Assume for instance that:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i \tag{4.29}$$

with

$$\varepsilon_i = \rho\varepsilon_{i-1} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2). \tag{4.30}$$

In this case, we are in the GRM context, with:

$$\Sigma = \frac{\sigma_v^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & ... & \rho^{n-1} \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & ... & 1 \end{bmatrix}. \tag{4.31}$$

In some cases —in particular when one assumes a parametric formulation for Σ— one can determine a better (more accurate) estimator than the OLS one. This approach is called Generalized Least Squares (GLS), which we present below.

4.5.2 Generalized Least Squares

Assume Σ is known ("feasible GLS"). Because Σ is symmetric positive, it admits a spectral decomposition of the form $\Sigma = \mathbf{C}\Lambda\mathbf{C}'$, where $\mathbf{C}$ is an orthogonal matrix (i.e. $\mathbf{C}\mathbf{C}' = Id$) and Λ is a diagonal matrix (the diagonal entries are the eigenvalues of Σ).

We have $\Sigma = (\mathbf{P}\mathbf{P}')^{-1}$ with $\mathbf{P} = \mathbf{C}\Lambda^{-1/2}$. Consider the transformed model:

$$\mathbf{P}'\mathbf{y} = \mathbf{P}'\mathbf{X}\beta + \mathbf{P}'\varepsilon \quad \text{or} \quad \mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*.$$

The variance of ε^* is the identity matrix Id. In the transformed model, OLS is BLUE (Gauss-Markow Theorem 4.1).

The **Generalized least squares** estimator of β is:

$$\boxed{\mathbf{b}_{GLS} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}.} \tag{4.32}$$

We have:

$$\mathbb{V}ar(\mathbf{b}_{GLS}|\mathbf{X}) = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}.$$

However, in general, Σ is unknown. The GLS estimator is then said to be *infeasible*. Some structure is required. Assume Σ admits a parametric form $\Sigma(\theta)$. The estimation becomes *feasible* (FGLS) if one replaces $\Sigma(\theta)$ by $\Sigma(\hat{\theta})$, where $\hat{\theta}$ is a consistent estimator of θ. In that case, the FGLS is asymptotically efficient (see Example 4.7).

When Σ has no obvious structure: the OLS (or IV) is the only estimator available. Under regularity assumptions, it remains unbiased, consistent, and asymptotically normally distributed, but not efficient. Standard inference procedures are no longer appropriate.

Example 4.7 (GLS in the auto-correlation case)**.** Consider the case presented in Example 4.6. Because the OLS estimate $\mathbf{b}$ of β is consistent, the estimates e_i of the ε_i's also are. Consistent estimators of ρ and σ_v are then obtained by regressing the e_i's on the e_{i-1}'s. Using these estimates in Eq. (4.31) provides a consistent estimate of Σ. Applying these steps recursively gives an efficient estimator of β (Cochrane and Orcutt (1949)).

4.5.3 Asymptotic properties of the OLS estimator in the GRM framework

Since $\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$ and $\mathbb{V}ar(\varepsilon|\mathbf{X}) = \Sigma$, we have:

$$\mathbb{V}ar(\mathbf{b}|\mathbf{X}) = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{4.33}$$

Therefore, the conditional covariance matrix of the OLS estimator is not $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ any longer, and using $s^2(\mathbf{X}'\mathbf{X})^{-1}$ for inference may be misleading. Below, we will see how to construct appropriate estimates of the covariance matrix of $\mathbf{b}$. Before that, let us prove that the OLS estimator remains consistent in the GRM framework.

Proposition 4.13 (Consistency of the OLS estimator in the GRM framework)**.** *If plim* $(\mathbf{X}'\mathbf{X}/n)$ *and plim* $(\mathbf{X}'\Sigma\mathbf{X}/n)$ *are finite positive definite matrices, then plim* $(\mathbf{b}) = \beta$*.*

Proof. We have $\mathbb{V}ar(\mathbf{b}) = \mathbb{E}[\mathbb{V}ar(\mathbf{b}|\mathbf{X})] + \mathbb{V}ar[\mathbb{E}(\mathbf{b}|\mathbf{X})]$. Since $\mathbb{E}(\mathbf{b}|\mathbf{X}) = \beta$, $\mathbb{V}ar[\mathbb{E}(\mathbf{b}|\mathbf{X})] = 0$. Eq. (4.33) implies that $\mathbb{V}ar(\mathbf{b}|\mathbf{X}) \to 0$. Hence $\mathbf{b}$ converges in mean square, and therefore in probability (see Prop. 9.13). □

Prop. 4.14 gives the asymptotic distribution of the OLS estimator in the GRM framework.

Proposition 4.14 (Asymptotic distribution of the OLS estimator in the GRM framework)**.** *If* $Q_{xx} = plim\ (\mathbf{X}'\mathbf{X}/n)$ *and* $Q_{x\Sigma x} = plim\ (\mathbf{X}'\Sigma\mathbf{X}/n)$ *are finite positive definite matrices, then:*

$$\sqrt{n}(\mathbf{b} - \beta) \xrightarrow{d} \mathcal{N}(0, Q_{xx}^{-1} Q_{x\Sigma x} Q_{xx}^{-1}).$$

The IV estimator also features a normal asymptotic distribution:

Proposition 4.15 (Asymptotic distribution of the IV estimator in the GRM framework)**.** *If regressors and IV variables are "well-behaved", then:*

$$\mathbf{b}_{iv} \overset{a}{\sim} \mathcal{N}(\beta, \mathbf{V}_{iv}),$$

where

$$\mathbf{V}_{iv} = \frac{1}{n}(\mathbf{Q}^*)\ plim\ \left(\frac{1}{n}\mathbf{Z}'\Sigma\mathbf{Z}\right)(\mathbf{Q}^*)',$$

with

$$\mathbf{Q}^* = [\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}\mathbf{Q}_{zx}]^{-1}\mathbf{Q}_{xz}\mathbf{Q}_{zz}^{-1}.$$

For practical purposes, one needs to have estimates of Σ in Props. 4.14 or 4.15. The complication comes from the fact that Σ is of dimension $n \times n$, and its estimation —based on a sample of length n— is therefore infeasible in the general case. Notwithstanding, looking at Eq. (4.33), it appears that one can focus on the estimation of $Q_{x\Sigma x} = \text{plim}\ (\mathbf{X}'\Sigma\mathbf{X}/n)$ (or plim $\left(\frac{1}{n}\mathbf{Z}'\Sigma\mathbf{Z}\right)$ in the IV case). This matrix being of dimension $K \times K$, its estimation is easier.

We have:

$$\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'. \tag{4.34}$$

The so-called **robust covariance matrices** are estimates of the previous matrix. Their computation is based on the fact that if $\mathbf{b}$ is consistent, then the e_i's are consistent estimators of the ε_i's.

In the following sections (4.5.4 and 4.5.5), we present two types of robust covariance matrices.

4.5.4 HAC-robust covariance matrices

When only heteroskedasticity prevails, i.e., when matrix Σ is as in Eq. (4.27), then one can use the formula proposed by White (1980) to estimate $\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}$ (see Example 4.8). When the residuals feature both heteroskedasticity and auto-correlation, then one can use the Newey and West (1987) approach (see Example 4.9).

Example 4.8 (Heteroskedasticity)**.** This is the case of Eq. (4.27). We have $\sigma_{i,j} = 0$ for $i \neq j$. Hence, in this case, we then need to estimate $\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'$. White (1980) has shown that, under general conditions:

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'\right). \tag{4.35}$$

The estimator of $\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}$ therefore is:

$$M_{HC0} = \frac{1}{n}\mathbf{X}'\begin{bmatrix} e_1^2 & 0 & \dots & 0 \\ 0 & e_2^2 & & \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & e_n^2 \end{bmatrix}\mathbf{X}. \tag{4.36}$$

where the e_i are the OLS residuals of the regression. The previous estimator is often called **HC0**. The **HC1** estimator, due to MacKinnon and White (1985), is obtained by applying an adjustment factor $n/(n-K)$ for the number of degrees of freedom (as in Prop. 4.6). That is:

$$M_{HC1} = \frac{n}{n-K}M_{HC0}. \tag{4.37}$$

We can illustrate the influence of heteroskedasticity using simulations. Consider the following model:

$$y_i = x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, x_i^2),$$

where the x_i's are i.i.d. $t(6)$.

Here is a simulated sample ($n = 200$) of this model:

```
n <- 200
x <- rt(n,df=6)
y <- x + x*rnorm(n)
par(plt=c(.1,.95,.1,.95))
plot(x,y,pch=19)
```

We simulate 1000 samples of the same model with $n = 200$. For each sample, we compute the OLS estimate of β ($= 1$). For each of the 1000 OLS estimations, we employ (a) the standard OLS variance formula ($s^2(\mathbf{X}'\mathbf{X})^{-1}$) and (b) the White formula to estimate the variance of b. For each formula, we compute the average of the 1000 resulting standard deviations and compare these with the standard deviation of the 1000 OLS estimate of β.

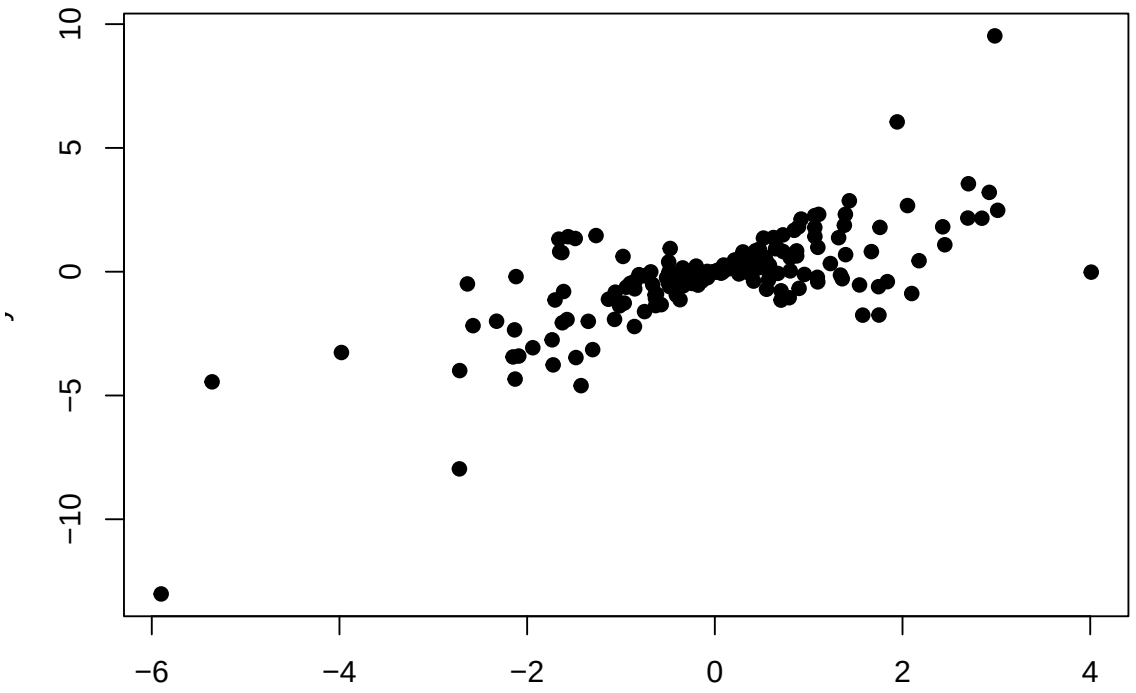


Figure 4.6: Situation of heteroskedasticity. The model is $y_i = x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, x_i^2)$, where the x_i's are i.i.d. $t(6)$.

```
n <- 200 # sample size
N <- 1000 # number of simulated samples
XX <- matrix(rt(n*N,df=6),n,N)
YY <- matrix(XX + XX*rnorm(n),n,N)
all_b       <- NULL;all_V_OLS   <- NULL;all_V_White <- NULL
for(j in 1:N){
  Y <- matrix(YY[,j],ncol=1)
  X <- matrix(XX[,j],ncol=1)
  b <- solve(t(X)%*%X) %*% t(X)%*%Y
  e <- Y - X %*% b
  S <- 1/n * t(X) %*% diag(c(e^2)) %*% X
  V_OLS   <- solve(t(X)%*%X) * var(e)
  V_White <- 1/n * (solve(1/n*t(X)%*%X)) %*% S %*% (solve(1/n*t(X)%*%X))
  all_b       <- c(all_b,b)
  all_V_OLS   <- c(all_V_OLS,V_OLS)
  all_V_White <- c(all_V_White,V_White)
}
c(sd(all_b),mean(sqrt(all_V_OLS)),mean(sqrt(all_V_White)))
```

```
## [1] 0.14024423 0.06748804 0.13973431
```

The results show that the White formula yields, on average, an estimated standard deviation that is much closer to the "true" value than the standard OLS formula. The latter

underestimate the standard deviation of b.

In the following example, we regress GDP growth rates from the Jordà et al. (2017) database on a systemic financial crisis dummy. We compute the HC0- and HC1-based standard deviations of the parameter estimate, and compare it to the one based on the standard OLS formula. The adjusted standard deviations are close to the one provided by the non-adjusted OLS formula.

```
library(lmtest)
library(sandwich)
nT <- dim(JST)[1]
JST$growth <- NaN
JST$growth[2:nT] <- log(JST$rgdpbarro[2:nT]/JST$rgdpbarro[1:(nT-1)])
JST.red <- subset(JST,year>1950)
JST.red$iso <- as.factor(JST.red$iso)
JST.red$year <- as.factor(JST.red$year)
eq <- lm(growth~crisisJST+iso,data=JST.red)
rbind(coeftest(eq)[2,],
      coeftest(eq, vcov = vcovHC(eq, type = "HC0"))[2,],
      coeftest(eq, vcov = vcovHC(eq, type = "HC1"))[2,])
```

```
##         Estimate  Std. Error   t value     Pr(>|t|)
## [1,] -0.02481424 0.005490411 -4.519560 6.789284e-06
## [2,] -0.02481424 0.005605452 -4.426804 1.040602e-05
## [3,] -0.02481424 0.005648268 -4.393247 1.212105e-05
```

Example 4.9 (Heteroskedasticity and Autocorrelation (HAC))**.** Newey and West (1987) have proposed a formula to address both heteroskedasticity and auto-correlation of the residuals (Eqs. (4.27) and (4.28)). They show that, if the correlation between terms i and j gets sufficiently small when $|i-j|$ increases:

$$\begin{aligned} &\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'\right) \approx \\ &\text{plim}\left(\frac{1}{n}\sum_{t=1}^{n}e_t^2\mathbf{x}_t\mathbf{x}_t' + \frac{1}{n}\sum_{\ell=1}^{L}\sum_{t=\ell+1}^{n}w_\ell e_t e_{t-\ell}(\mathbf{x}_t\mathbf{x}_{t-\ell}' + \mathbf{x}_{t-\ell}\mathbf{x}_t')\right), \end{aligned} \tag{4.38}$$

where $w_\ell = 1 - \ell/(L+1)$ (with L large).

Let us illustrate the influence of autocorrelation using simulations. We consider the following model:

$$y_i = x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, x_i^2), \tag{4.39}$$

where the x_i's and the ε_i's are such that:

$$x_i = 0.8x_{i-1} + u_i \quad \textit{and} \quad \varepsilon_i = 0.8\varepsilon_{i-1} + v_i, \tag{4.40}$$

where the u_i's and the v_i's are i.i.d. $\mathcal{N}(0,1)$.

We simulate 500 samples of the same model with $n = 200$. For each sample, we compute the OLS estimate of β (=1). For each of the 1000 OLS estimations, we employ (a) the standard OLS variance formula ($s^2(\mathbf{X}'\mathbf{X})^{-1}$), (b) the White formula, and (c) the Newey-West formula to estimate the variance of b. For each formula, we compute the average of the 500 resulting standard deviations and compare these with the standard deviation of the 500 OLS estimate of β.

```
library(AEC)
n <- 100 # sample length
nb.sim <- 500 # number of simulated samples
all.b <- NULL;all.OLS.stdv.b <- NULL
all.Whi.stdv.b <- NULL;all.NW.stdv.b <- NULL
for(i in 1:nb.sim){
  eps <- rnorm(n);x <- rnorm(n)
  for(i in 2:n){
    eps[i] <- eps[i] + .8*eps[i-1]
    x[i]   <- x[i]   + .8*x[i-1]
  }
  y <- x + eps
  eq <- lm(y~x)
  all.b <- c(all.b,eq$coefficients[2])
  all.OLS.stdv.b <- c(all.OLS.stdv.b,summary(eq)$coefficients[2,2])
  X <- cbind(rep(1,n),x)
  XX <- t(X) %*% X;XX_1 <- solve(XX)
  E2 <- diag(eq$residuals^2)
  White.V <- XX_1 %*% (t(X) %*% E2 %*% X) %*% XX_1
  all.Whi.stdv.b <- c(all.Whi.stdv.b,sqrt(White.V[2,2]))
  # HAC:
  U <- X * cbind(eq$residuals,eq$residuals)
  XSigmaX <- NW.LongRunVariance(U,5)
  NW.V <- 1/n * (n*XX_1) %*% XSigmaX %*% (n*XX_1)
  all.NW.stdv.b <- c(all.NW.stdv.b,sqrt(NW.V[2,2]))
}
cbind(sd(all.b),mean(all.OLS.stdv.b),
      mean(all.Whi.stdv.b),mean(all.NW.stdv.b))
```

```
##          [,1]      [,2]      [,3]     [,4]
## [1,] 0.201172 0.1013048 0.0962689 0.146974
```

The results show that the Newey-West formula yields, on average, an estimated standard deviation that is closer to the "true" value than the standard OLS formula. The latter underestimate the standard deviation of b.

What precedes suggest that, when the residuals feature autocorrelation, it is important to use appropriately adjusted covariance matrices to make statistical inference. How to detect autocorrelation in residuals? A popular test has been proposed by Durbin and Watson (1950) and Durbin and Watson (1951). The Durbin-Watson test statistic is:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = 2(1-r) - \underbrace{\frac{e_1^2 + e_n^2}{\sum_{i=1}^{n} e_i^2}}_{\xrightarrow{p} 0},$$

where r is the slope in the regression of the e_i's on the e_{i-1}'s, i.e.:

$$r = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n-1} e_i^2}.$$

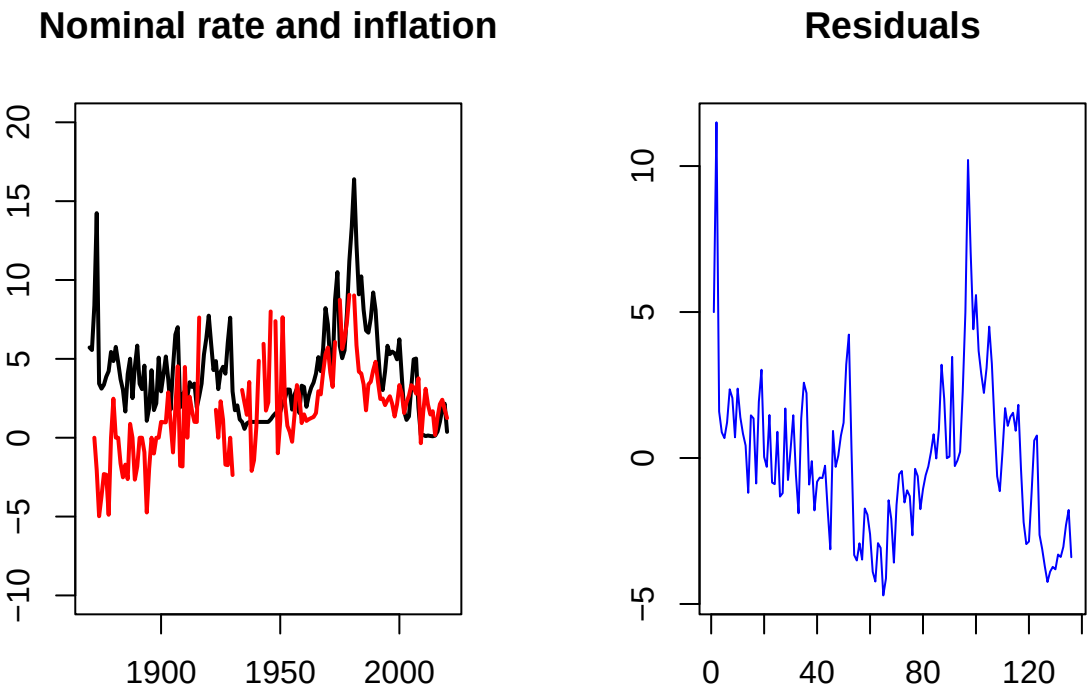
(r is a consistent estimator of $\mathbb{C}or(\varepsilon_i, \varepsilon_{i-1})$, i.e. ρ in Eq. (4.30).)

The one-sided test for H_0: $\rho = 0$ against H_1: $\rho > 0$ is carried out by comparing DW to values $d_L(T,K)$ and $d_U(T,K)$:

$$\begin{cases} \text{If } DW < d_L, & \text{the null hypothesis is rejected;} \\ \text{if } DW > d_U, & \text{the hypothesis is not rejected;} \\ \text{If } d_L \leq DW \leq d_U, & \text{no conclusion is drawn.} \end{cases}$$

Example 4.10 (Durbin-Watson test)**.** We regress the short-term nominal US interest rate on inflation. We then employ the Durbin-Watson test to see whether the residuals are auto-correlated (which is quite obviously the case).

```
library(car)
data <- subset(JST,iso=="USA");T <- dim(data)[1]
data$infl <- c(NaN,100*log(data$cpi[2:T]/data$cpi[1:(T-1)]))
data$infl[(data$infl< -5)|(data$infl>10)] <- NaN
par(mfrow=c(1,2))
plot(data$year,data$stir,ylim=c(-10,20),type="l",lwd=2,xlab="",
     ylab="",main="Nominal rate and inflation")
lines(data$year,data$infl,col="red",lwd=2)
eq <- lm(stir~infl,data=data)
plot(eq$residuals,type="l",col="blue",main="Residuals",xlab="",ylab="")
```




```
durbinWatsonTest(eq)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7321902     0.4984178       0
##  Alternative hypothesis: rho != 0
```

4.5.5 Cluster-robust covariance matrices

The present section is based on MacKinnon et al. (2022); another useful reference is Cameron and Miller (2014). We will see how one can approximate $Q_{x\Sigma x} = \text{plim}\,(\mathbf{X}'\Sigma\mathbf{X}/n)$ (see Prop. 4.14) when the dataset can be decomposed into *clusters*. These clusters constitute a partition of the sample, and they are such that the error terms may be correlated within a given cluster, but not across different clusters. A cluster may, e.g., gathers entities from a given geographical area, from the same industry, or same age cohort.

The OLS estimator satisfies:

$$\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon. \tag{4.41}$$

Consider a set $\{n_1, n_2, \dots, n_G\}$ s.t. $n = \sum_g n_g$, on which is based the following decomposition of $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_G \end{bmatrix}.$$

With these notations, Eq. (4.41) rewrites:

$$\mathbf{b} - \beta = \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{X}_g \right)^{-1} \sum_{g=1}^{G} \mathbf{s}_g, \tag{4.42}$$

where

$$\mathbf{s}_g = \mathbf{X}_g' \varepsilon_g \tag{4.43}$$

denotes the score vector (of dimension $K \times 1$) associated with the g^{th} cluster.

If the model is correctly specified then $\mathbb{E}(\mathbf{s}_g) = 0$ for all clusters g. Note that Eq. (4.42) is valid for any partition of $\{1, ..., n\}$. Dividing the sample into clusters becomes meaningful if we further assume that the following hypothesis holds:

Hypothesis 4.6 (Clusters)**.** We have:

$$(i)\ \mathbb{E}(\mathbf{s}_g \mathbf{s}_g') = \Sigma_g, \quad (ii)\ \mathbb{E}(\mathbf{s}_g \mathbf{s}_q') = 0,\ g \neq q,$$

where s_g is defined in Eq. (4.43).

The real assumption here is (ii); the first one simply gives a notation for the covariance matrix of the score associated with the g^{th} cluster. Remark that these covariance matrices can differ across clusters. That is, *cluster-based inference is robust against both heteroskedasticity and intra-cluster dependence without imposing any restrictions on the (unknown) form of either of them.*

Naturally, matrix Σ_g depends on the covariance structure of the ε's. In particular, if $\Omega_g = \mathbb{E}(\varepsilon_g \varepsilon_g' | \mathbf{X}_g)$, then we have $\Sigma_g = \mathbb{E}(\mathbf{X}_g' \Omega_g \mathbf{X}_g)$.

Under Hypothesis 4.6, it comes that the conditional covariance matrix of $\mathbf{b}$ is:

$$(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^{G} \Sigma_g \right) (\mathbf{X}'\mathbf{X})^{-1} \tag{4.44}$$

Let us denote by $\varepsilon_{g,i}$ the error associated with the i^{th} component of vector ε_g. Consider the special case where $\mathbb{E}(\varepsilon_{g,i} \varepsilon_{g,j} | \mathbf{X}_g) = \sigma^2 \mathbb{I}_{\{i=j\}}$, then Eq. (4.44) gives the standard expression $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ (see Eq. (4.9)).

If we have $\mathbb{E}(\varepsilon_{gi} \varepsilon_{gj} | \mathbf{X}_g) = \sigma_{gi}^2 \mathbb{I}_{\{i=j\}}$, then we fall in the case addressed by the White formula (see Example 4.8). That is, in this case, the conditional covariance matrix of $\mathbf{b}$ is:

$$(\mathbf{X}'\mathbf{X})^{-1} \left(\mathbf{X}' \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & ... & 0 & \sigma_n^2 \end{bmatrix} \mathbf{X} \right) (\mathbf{X}'\mathbf{X})^{-1}.$$

As in White (1980), the natural way to approach the conditional covariance given in Eq. (4.44) consists in replacing the Σ_g matrices by their sample equivalent, i.e. $\widehat{\Sigma}_g = \mathbf{X}_g' \mathbf{e}_g \mathbf{e}_g' \mathbf{X}_g$.

Adding corrections for the number of degrees of freedom, this leads to the following estimate of the covariance matrix of $\mathbf{b}$:

$$\frac{G(n-1)}{(G-1)(n-K)}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\widehat{\Sigma}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{4.45}$$

The previous estimate is CRCV1 in MacKinnon et al. (2022). Note that we indeed find the White-MacKinnon estimator (Eq. (4.37)) when $G = n$.

Remark that if there was only one cluster ($G = 1$), and neglecting the degree-of-freedom correction, we would have:

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} = 0$$

because $\mathbf{X}'\mathbf{e} = 0$. Hence, having large clusters does not necessarily increase variance.

Often, when working with panel data (see Chapter 5), we want to cluster in different dimensions. A typical case is when the data are indexed by both individuals (i) and time (t). In that case, we may indeed suspect that: (a) the residuals are correlated across clusters of dates (e.g., with monthly data, a cluster may be one year) and (b) the residuals are correlated across clusters of individuals (e.g., with data at the county level a cluster may be a state). In this case, one can employ **two-way clustering**.

Formally, consider two distinct partitions of the data: one through index g, with $g \in \{1, \dots, G\}$, and the other through index h, with $h \in \{1, \dots, H\}$. Accordingly, we denote by $\mathbf{X}_{g,h}$ the submatrix of $\mathbf{X}$ that contains the explanatory variables corresponding to clusters g and h (e.g., the firms of a given country g at a given date h). We also denote by $\mathbf{X}_{g,\bullet}$ (respectively $\mathbf{X}_{\bullet,h}$) the submatrix of $\mathbf{X}$ containing all explanatory variables pertaining to cluster g, for all possible values of h (resp. to cluster h, for all possible values of g).

We will make the following assumption:

Hypothesis 4.7 (Two-way clusters)**.** We have:

$$\begin{aligned}&\mathbb{E}(\mathbf{s}_{g,\bullet}\mathbf{s}'_{g,\bullet}) = \Sigma_g, \quad \mathbb{E}(\mathbf{s}_{\bullet,h}\mathbf{s}'_{\bullet,h}) = \Sigma^*_h, \quad \mathbb{E}(\mathbf{s}_{g,h}\mathbf{s}'_{g,h}) = \Sigma_{g,h},\\ &\mathbb{E}(\mathbf{s}_{g,h}\mathbf{s}'_{q,k}) = 0 \text{ if } g \neq q \text{ and } h \neq k.\end{aligned}$$

Proposition 4.16 (Covariance of scores in the two-way-cluster setup)**.** *Under this assumption, the matrix of covariance of the scores is given by:*

$$\Sigma = \sum_{g=1}^{G}\Sigma_g + \sum_{h=1}^{H}\Sigma^*_h - \sum_{g=1}^{G}\sum_{h=1}^{H}\Sigma_{g,h}.$$

(The last term on the right-hand side must be subtracted in order to avoid double counting.)

Proof. We have:

$$\begin{aligned}\Sigma &= \sum_{g=1}^{G}\sum_{q=1}^{G}\sum_{h=1}^{H}\sum_{k=1}^{H}\mathbf{s}_{g,h}\mathbf{s}_{q,k}' \\ &= \sum_{g=1}^{G}\underbrace{\left(\sum_{h=1}^{H}\sum_{k=1}^{H}\mathbf{s}_{g,h}\mathbf{s}_{g,k}'\right)}_{=\Sigma_g} + \sum_{h=1}^{H}\underbrace{\left(\sum_{g=1}^{G}\sum_{q=1}^{G}\mathbf{s}_{g,h}\mathbf{s}_{q,h}'\right)}_{=\Sigma_h^*} - \sum_{g=1}^{G}\sum_{h=1}^{H}\mathbf{s}_{g,h}\mathbf{s}_{g,h}',\end{aligned}$$

which gives the result. □

The asymptotic theory can be based on two different approaches: (i) large number of clusters (common case), and (ii) fixed number of clusters but large number of observations in each cluster (see Subsections 4.1 and 4.2 in MacKinnon et al. (2022)). The more variable the N_g's (clusters' sizes are heterogeneous in terms of size), the less reliable asymptotic inference based on Eq. (4.45), especially when a very few clusters are unusually large, or when the distribution of the data is heavy-tailed. These issues are somehow mitigated when the clusters have an approximate factor structure.

In practice, Σ is estimated by:

$$\widehat{\Sigma} = \sum_{g=1}^{G}\hat{\mathbf{s}}_{g,\bullet}\hat{\mathbf{s}}_{g,\bullet}' + \sum_{h=1}^{H}\hat{\mathbf{s}}_{\bullet,h}\hat{\mathbf{s}}_{\bullet,h} - \sum_{g=1}^{G}\sum_{h=1}^{H}\hat{\mathbf{s}}_{g,h}\hat{\mathbf{s}}_{g,h}',$$

and we use:

$$\widehat{\mathbb{V}ar}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\,\widehat{\Sigma}\,(\mathbf{X}'\mathbf{X})^{-1}.$$

As an alternative to the asymptotic approximation to the distribution of a statistic of interest, one can resort to bootstrap approximation (see Section 5 of MacKinnon et al. (2022)). In R, the package `fwildclusterboot` allows to implement such approaches.[4]

Let us come back to the analysis of the effect of systemic financial crises on GDP growth. Clustering the data at the country level and, further, at both the country and time levels gives the following:

```
eq <- lm(growth~crisisJST+iso,data=JST.red)
rbind(coeftest(eq)[2,],
      coeftest(eq, vcov = vcovHC(eq, type = "HC1"))[2,],
      coeftest(eq, vcov = vcovCL(eq, cluster = JST.red[, c("iso")]))[2,],
      coeftest(eq, vcov = vcovCL(eq, cluster = JST.red[, c("iso","year")]))[2,])
```

```
##         Estimate  Std. Error   t value     Pr(>|t|)
## [1,] -0.02481424 0.005490411 -4.519560 6.789284e-06
## [2,] -0.02481424 0.005648268 -4.393247 1.212105e-05
## [3,] -0.02481424 0.005847708 -4.243413 2.365508e-05
## [4,] -0.02481424 0.006546931 -3.790209 1.577572e-04
```

[4] See, e.g., this tutorial by Alexander Fischer.

4.6 Shrinkage methods

Chosing the appropriate explanatory variables is often complicated, especially in the presence of many potentially relevant covariates. Keeping a large number of covariates results in large standard deviations for the estimated parameters (see Section 4.3.3). In order to address this issue, shrinkage methods have been designed. The objective of these methods is to help select a limited number of variables by shrinking the regression coefficients of the less useful variables towards zero. The two best-known shrinkage techniques are **ridge regression** and the **lasso** approach.[5] In both cases (ridge and lasso), the OLS minimization problem (see Section 4.2), i.e.,

$$\mathbf{b} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 \tag{4.46}$$

is replaced with the following:

$$\mathbf{b}_\lambda = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2 + \lambda f(\beta), \tag{4.47}$$

where $\lambda f(\beta)$ is a penalty term that positively depends on the "size" of the components of β. This term is called the *shrinkage penalty* term.

Specifically, assuming that the vector $\mathbf{x}_i$ of potential covariates is of dimension $K \times 1$, we have:

$$\begin{aligned} f(\beta) &= \sum_{j=1}^{K} \beta_j^2 \quad \text{in the ridge case } (\ell_2 \text{ norm}), \\ f(\beta) &= \sum_{j=1}^{K} |\beta_j| \quad \text{in the lasso case } (\ell_1 \text{ norm}). \end{aligned}$$

In most cases, we do not want to involve the intercept in the set of parameters to shrink, and the preceding equations are respectively replaced with:

$$\begin{aligned} f(\beta) &= \sum_{j=2}^{K} \beta_j^2 \quad \text{(ridge)}, \\ f(\beta) &= \sum_{j=2}^{K} |\beta_j| \quad \text{(lasso)}. \end{aligned}$$

The nature of the penalty —based on either the ℓ_1 or the ℓ_2 norm— implies a different behaviour of the parameter estimates when λ —the *tuning parameter*— grows. In the ridge regression, coefficient estimates go to zero (shrinkage); in the lasso case, some coefficients reach zero when λ reach some values. In other words, while ridge regression achieve shrinkage, lasso regressions achieve shrinkage *and* variable selection.

[5] See Tibshirani (2011) for a review of the lasso approach. See also Section 6.2 of James et al. (2013).

Parameter λ has to be determined separately from the minimization problem of Eq. (4.47). One can combine standard criteria (e.g., BIC or Akaike) for this purpose.

In R, one can use the `glmnet` package to run ridge and lasso regressions. In the following example, we employ this package to model interest rates proposed to debtors, using data extracted from the Lending Club platform.

To begin with, let us define the variables we want to consider:

```
library(AEC)
library(glmnet)
credit$owner <- 1*(credit$home_ownership=="OWN")
credit$renter <- 1*(credit$home_ownership=="MORTGAGE")
credit$verification_status <- 1*(credit$verification_status=="Not Verified")
credit$emp_length_10 <- 1*(credit$emp_length_10)
credit$log_annual_inc <- log(credit$annual_inc)
credit$log_funded_amnt <- log(credit$funded_amnt)
credit$annual_inc2 <- (credit$annual_inc)^2
credit$funded_amnt2 <- (credit$funded_amnt)^2
x <- subset(credit,
            select = c(delinq_2yrs,annual_inc,annual_inc2,log_annual_inc,
                       dti,installment,verification_status,funded_amnt,
                       funded_amnt2,log_funded_amnt,pub_rec,emp_length_10,
                       owner,renter,pub_rec_bankruptcies,revol_util,revol_bal))
```

Let us standardize the data:

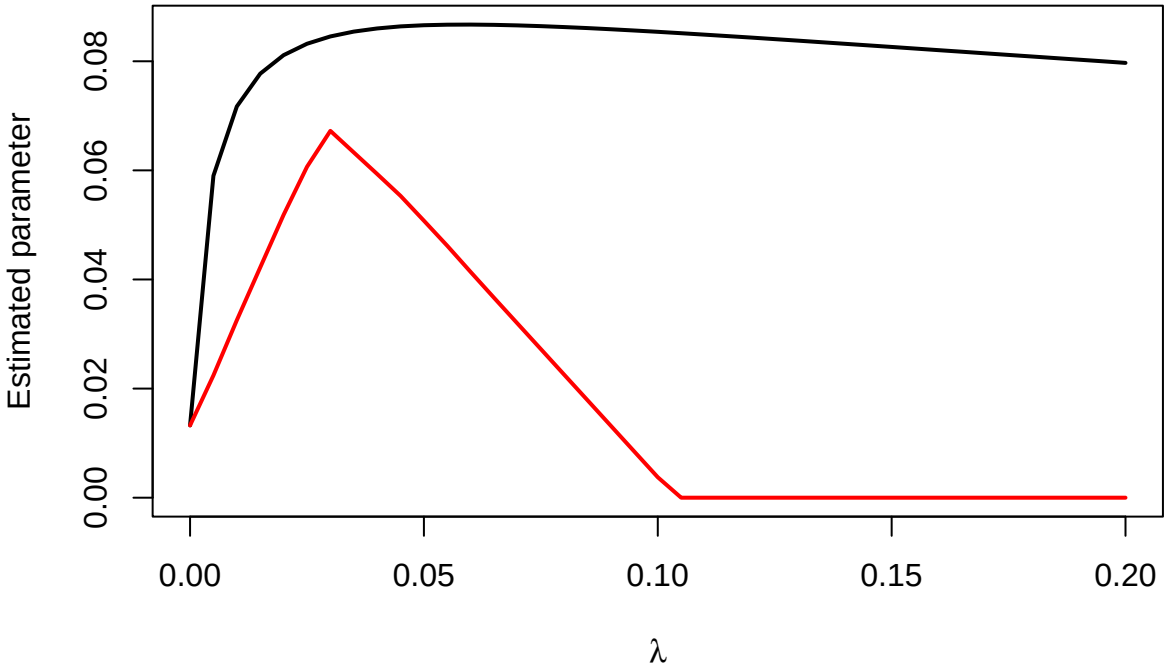
```
y <- scale(credit$int_rate)
x <- scale(x)
```

Next, we define the set of λ we will use, and run the ridge and lasso regressions:

```
grid.lambda <- seq(0,.2,by=.005)
result.ridge <- glmnet(x, y, alpha = 0, lambda = grid.lambda)
result.lasso <- glmnet(x, y, alpha = 1, lambda = grid.lambda)
```

The following figure shows how estimated parameters depend on λ:

```
variab <- 6
plot(result.ridge$lambda,coef(result.ridge)[variab,],type="l",
     ylim=c(min(coef(result.ridge)[variab,],coef(result.lasso)[variab,]),
            max(coef(result.ridge)[variab,],coef(result.lasso)[variab,])),
     xlab=expression(lambda),ylab="Estimated parameter",lwd=2)
lines(result.lasso$lambda,coef(result.lasso)[variab,],col="red",lwd=2)
```



Let us take two values of λ and see the associated estimated parameters in the context of lasso regressions:

```
i <- 20; j <- 40
cbind(result.lasso$lambda[i],result.lasso$lambda[j])
```

```
##       [,1]  [,2]
## [1,] 0.105 0.005
```

```
cbind(coef(result.lasso)[,i],coef(result.lasso)[,j])
```

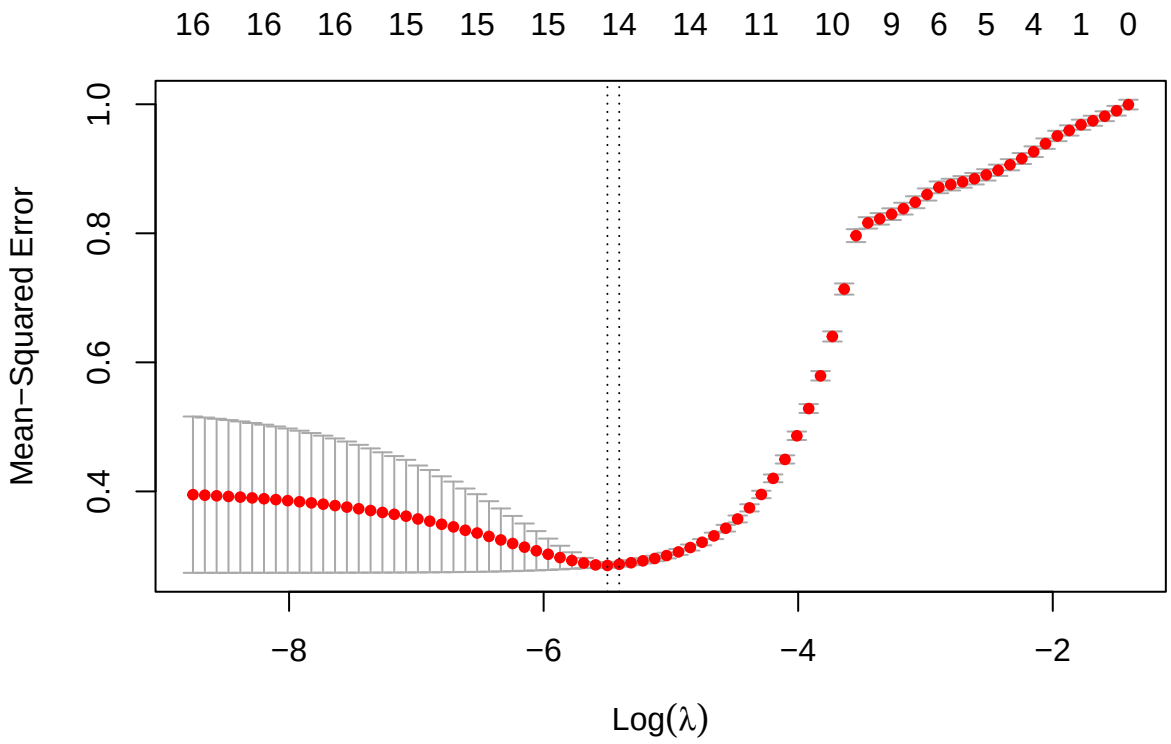
```
##                              [,1]          [,2]
## (Intercept)          -1.044971e-15  1.088731e-14
## delinq_2yrs           6.308870e-02  6.893527e-02
## annual_inc            0.000000e+00  4.595653e-03
## annual_inc2           0.000000e+00  0.000000e+00
## log_annual_inc        0.000000e+00 -3.612382e-02
## dti                   0.000000e+00  2.242246e-02
## installment           1.476796e-01  8.228729e+00
## verification_status   0.000000e+00 -9.750047e-04
## funded_amnt           0.000000e+00 -7.309169e+00
## funded_amnt2          0.000000e+00 -4.711846e-01
## log_funded_amnt       0.000000e+00 -2.460932e-01
## pub_rec               3.390816e-02  5.997252e-02
## emp_length_10         0.000000e+00 -1.924941e-02
## owner                 0.000000e+00 -2.444599e-02
```

```
## renter                -3.882640e-02 -6.243087e-02
## pub_rec_bankruptcies   0.000000e+00  0.000000e+00
## revol_util             0.000000e+00  0.000000e+00
## revol_bal              0.000000e+00  2.402268e-03
```

```
# Compute values of y predicted by the model, for all lambdas:
pred1 <- predict(result.lasso,as.matrix(x))
# Compute values of y predicted by the model, for a specific value:
pred2 <- predict(result.lasso,as.matrix(x),s=0.085)
```

The `glmnet` package (see Hastie et al. (2021)) also offers tools to implement cross-validation:

```
# cross validation (cv):
cvglmnet <- cv.glmnet(as.matrix(x),y)
plot(cvglmnet)
```




```
# lambda.min: lambda that gives minimum mean cross-validated error
cvglmnet$lambda.min
```

```
## [1] 0.004091039
```

```
# lambda.1se: largest lambda s.t. cost within the one-std-dev cv-based band
cvglmnet$lambda.1se
```

```
## [1] 0.00448991
```

```
coef(cvglmnet, s = "lambda.min") # associated parameters
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                                  s1
## (Intercept)           1.128990e-14
## delinq_2yrs           6.625860e-02
## annual_inc            6.054923e-03
## annual_inc2           .
## log_annual_inc       -3.795151e-02
## dti                   2.065272e-02
## installment           8.518961e+00
## verification_status  -2.894435e-03
## funded_amnt          -7.560853e+00
## funded_amnt2         -5.041923e-01
## log_funded_amnt      -2.537567e-01
## pub_rec               5.804333e-02
## emp_length_10        -1.894666e-02
## owner                -2.484192e-02
## renter               -6.039831e-02
## pub_rec_bankruptcies  .
## revol_util            .
## revol_bal             3.093752e-03
```

```
# predicted values of y for specific values of x:
predict(cvglmnet, newx = as.matrix(x)[1:5,], s = "lambda.min")
```

```
##        lambda.min
## 21529  0.34496384
## 21547 -0.04553753
## 21579  0.56455499
## 21583 -0.20696954
## 21608 -0.12165356
```

Chapter 5

Panel regressions

5.1 Specification and notations

A standard panel situation is as follows: the sample covers a lot of "entities", indexed by $i \in \{1, \dots, n\}$, with n large, and, for each entity, we observe different variables over a small number of periods $t \in \{1, \dots, T\}$. This is a *longitudinal dataset*.

The linear panel regression model is:

$$y_{i,t} = \mathbf{x}_{i,t}' \underbrace{\beta}_{K\times 1} + \underbrace{\mathbf{z}_i'\alpha}_{\text{Individual effects}} + \varepsilon_{i,t}. \tag{5.1}$$

When running panel regressions, the usual objective is to estimate β.

Figure 5.1 illustrates a panel-data situation. The model is $y_i = \alpha_i + \beta x_{i,t} + \varepsilon_{i,t}$, $t \in \{1, 2\}$. On Panel (b), blue dots are for $t = 1$, red dots are for $t = 2$. The lines relate the dots associated with the same entity i. What is remarkable in the simulated model is that, while the unconditional correlation between y and x is negative, the conditional correlation (conditional on α_i) is positive. Indeed, the sign of this conditional correlation is the sign of β, which is positive in th simulated example ($\beta = 5$). In other words, if one did not know the panel nature of the data, that would be tempting to say that $\beta < 0$, but this is not the case, due to **fixed effects** (the α_i's) that are negatively correlated to the x_i's.

```
T <- 2; n <- 12 # 2 periods and 12 entities
alpha <- 5*rnorm(n) # draw fixed effects
x.1 <- rnorm(n) - .5*alpha # note: x_i's correlate to alpha_i's
x.2 <- rnorm(n) - .5*alpha
beta <- 5; sigma <- .3
y.1 <- alpha + x.1 + sigma*rnorm(n);y.2 <- alpha + x.2 + sigma*rnorm(n)
x <- c(x.1,x.2) # pooled x
y <- c(y.1,y.2) # pooled y
par(mfrow=c(1,2))
```

```
plot(x,y,col="black",pch=19,xlab="x",ylab="y",main="(a)")
plot(x,y,col="black",pch=19,xlab="x",ylab="y",main="(b)")
points(x.1,y.1,col="blue",pch=19);points(x.2,y.2,col="red",pch=19)
for(i in 1:n){lines(c(x.1[i],x.2[i]),c(y.1[i],y.2[i]))}
```

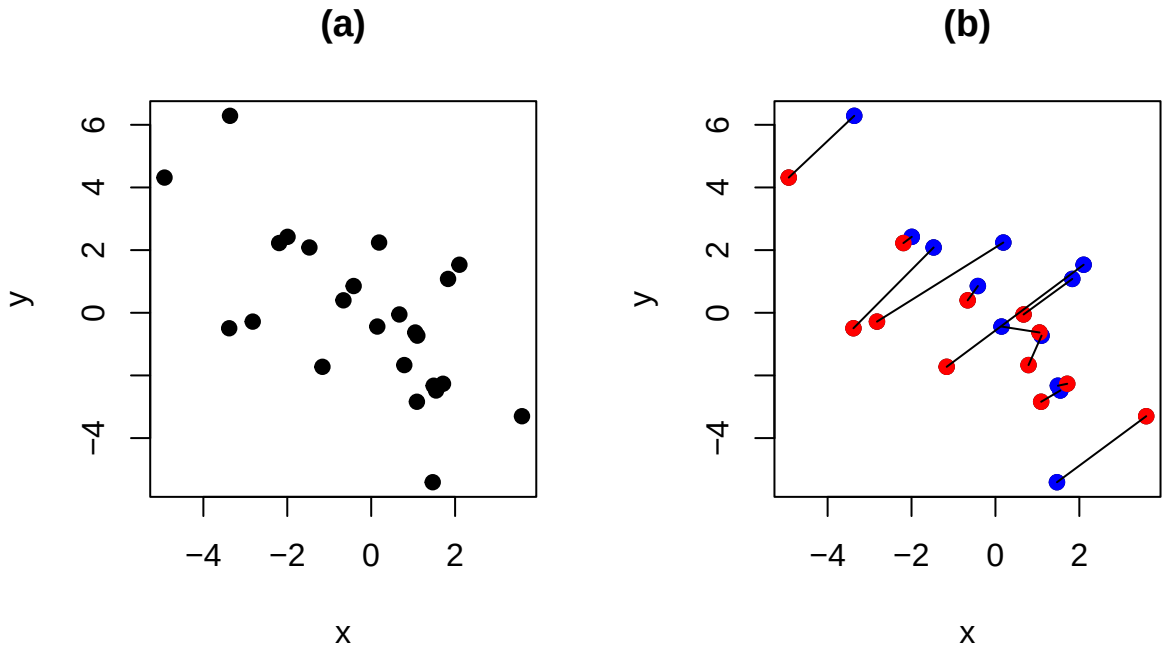


Figure 5.1: The data are the same for both panels. On Panel (b), blue dots are for $t = 1$, red dots are for $t = 2$. The lines relate the dots associated to the same entity i.

Figure 5.2 presents the same type of plot based on the Cigarette Consumption Panel dataset (`CigarettesSW` dataset, used in Stock and Watson (2003)). This dataset documents the average consumption of cigarettes in 48 continental US states for two dates (1985 and 1995).

We will make use of the following notations:

$$\mathbf{y}_i = \underbrace{\begin{bmatrix} y_{i,1} \\ \vdots \\ y_{i,T} \end{bmatrix}}_{T\times 1}, \quad \varepsilon_i = \underbrace{\begin{bmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,T} \end{bmatrix}}_{T\times 1}, \quad \mathbf{x}_i = \underbrace{\begin{bmatrix} \mathbf{x}_{i,1}' \\ \vdots \\ \mathbf{x}_{i,T}' \end{bmatrix}}_{T\times K}, \quad \mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}}_{(nT)\times K}.$$

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} y_{i,1} - \bar{y}_i \\ \vdots \\ y_{i,T} - \bar{y}_i \end{bmatrix}, \quad \tilde{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i,1} - \bar{\varepsilon}_i \\ \vdots \\ \varepsilon_{i,T} - \bar{\varepsilon}_i \end{bmatrix},$$

$$\tilde{\mathbf{x}}_i = \begin{bmatrix} \mathbf{x}_{i,1}' - \bar{\mathbf{x}}_i' \\ \vdots \\ \mathbf{x}_{i,T}' - \bar{\mathbf{x}}_i' \end{bmatrix}, \quad \tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \vdots \\ \tilde{\mathbf{x}}_n \end{bmatrix}, \quad \tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_n \end{bmatrix},$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}, \quad \bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{i,t} \quad \text{and} \quad \bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{i,t}.$$

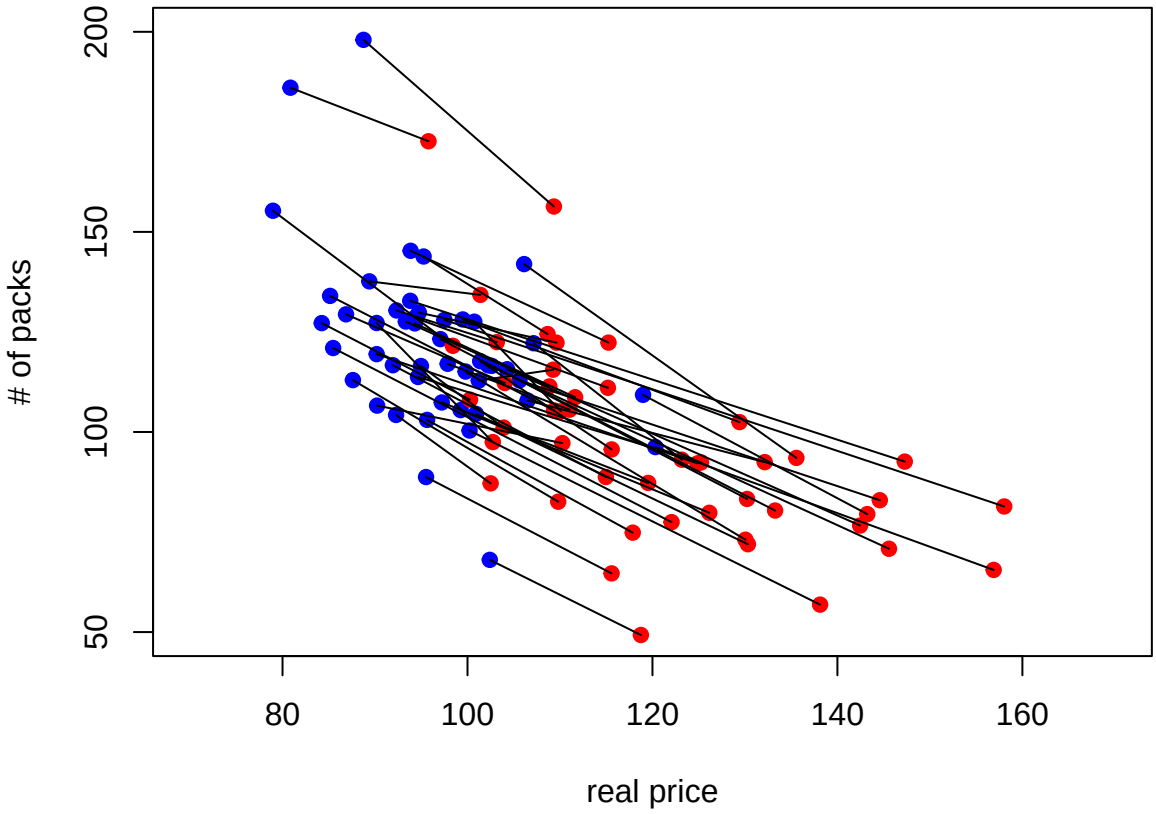


Figure 5.2: Cigarette consumption versus real price in the CigarettesSW panel dataset.

5.2 Three standard cases

There are three typical situations:

- **Pooled regression**: $\mathbf{z}_i \equiv 1$. This case amounts to the case studied in Chapter 4.
- **Fixed Effects** (Section 5.3): $\mathbf{z}_i$ is unobserved, but correlates with $\mathbf{x}_i \Rightarrow \mathbf{b}$ is biased and inconsistent in the OLS regression of $\mathbf{y}$ on $\mathbf{X}$ (omitted variable, see Section 4.3.2).
- **Random Effects** (Section 5.4): $\mathbf{z}_i$ is unobserved, but uncorrelated with $\mathbf{x}_i$. The model writes:
$$y_{i,t} = \mathbf{x}_{i,t}'\beta + \alpha + \underbrace{u_i + \varepsilon_{i,t}}_{\text{compound error}},$$
where $\alpha = \mathbb{E}(\mathbf{z}_i'\alpha)$ and $u_i = \mathbf{z}_i'\alpha - \mathbb{E}(\mathbf{z}_i'\alpha) \perp \mathbf{x}_i$. In that case, the OLS is consistent, but not efficient. GLS can be used to gain efficiencies over OLS (see Section 4.5.2 for a presentation of the GLS approach).

5.3 Estimation of Fixed-Effects Models

Hypothesis 5.1 (Fixed-effect model)**.** We assume that:

i. There is no perfect multicollinearity among the regressors.
ii. $\mathbb{E}(\varepsilon_{i,t}|\mathbf{X}) = 0$, for all i, t.
iii. We have:
$$\mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,s}|\mathbf{X}) = \begin{cases} \sigma^2 & \text{if } i = j \text{ and } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

These assumptions are analogous to those introduced in the standard linear regression:

(i) $\leftrightarrow$ Hyp. 4.1, (ii) $\leftrightarrow$ Hyp. 4.2, (iii) $\leftrightarrow$ Hyp. 4.3 + 4.4.

In matrix form, for a given i, the model writes:

$$\mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{1}\alpha_i + \varepsilon_i,$$

where $\mathbf{1}$ is a T-dimensional vector of ones.

This is the **Least Square Dummy Variable (LSDV)** model:

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{D}] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \varepsilon, \tag{5.2}$$

with:

$$\mathbf{D} = \underbrace{\begin{bmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} \end{bmatrix}}_{(nT \times n)}.$$

The linear regression (Eq. (5.2)) —with the dummy variables— satisfies the Gauss-Markov conditions (Theorem 4.1). Hence, in this context, the OLS estimator is the *best linear unbiased estimator* (BLUE).

Denoting by $\mathbf{Z}$ the matrix $[\mathbf{X} \quad \mathbf{D}]$, and by $\mathbf{b}$ and $\mathbf{a}$ the respective OLS estimates of β and of α, we have:

$$\boxed{\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y}.} \tag{5.3}$$

The asymptotical distribution of $[\mathbf{b}', \mathbf{a}']'$ derives from the standard OLS context: Prop. 4.11 can be used after having replaced $\mathbf{X}$ by $\mathbf{Z} = [\mathbf{X} \quad \mathbf{D}]$.

We have:

$$\boxed{\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} \beta \\ \alpha \end{bmatrix}, \sigma^2 \frac{Q^{-1}}{nT}\right),} \tag{5.4}$$

where

$$Q = \text{plim}_{nT\to\infty} \frac{1}{nT}\mathbf{Z}'\mathbf{Z},$$

assuming the previous limit exists.

In practice, an estimator of the covariance matrix of $[\mathbf{b}', \mathbf{a}']'$ is:

$$s^2 \left(\mathbf{Z}'\mathbf{Z}\right)^{-1} \quad with \quad s^2 = \frac{\mathbf{e}'\mathbf{e}}{nT - K - n},$$

where $\mathbf{e}$ is the $(nT) \times 1$ vector of OLS residuals.

There is an alternative way of expressing the LSDV estimators. It involves the residual-maker matrix matrix $\mathbf{M_D} = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ (see Eq. (4.4)), which acts as an operator that removes entity-specific means, i.e.:

$$\tilde{\mathbf{Y}} = \mathbf{M_D}\mathbf{Y}, \quad \tilde{\mathbf{X}} = \mathbf{M_D}\mathbf{X} \quad and \quad \tilde{\varepsilon} = \mathbf{M_D}\varepsilon.$$

With these notations, using the Frisch-Waugh theorem (Theorem 4.2), we get another expression for the estimator $\mathbf{b}$ appearing in Eq. (5.3):

$$\boxed{\mathbf{b} = [\mathbf{X}'\mathbf{M_D}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{M_D}\mathbf{y}.} \tag{5.5}$$

This amounts to regressing the $\tilde{y}_{i,t}$'s $(= y_{i,t} - \bar{y}_i)$ on the $\tilde{\mathbf{x}}_{i,t}$'s $(= \mathbf{x}_{i,t} - \bar{\mathbf{x}}_i)$.

The estimate of α is given by:

$$\boxed{\mathbf{a} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathbf{b}),} \tag{5.6}$$

which is obtained by developing the second row of

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{D} \\ \mathbf{D}'\mathbf{X} & \mathbf{D}'\mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{D}'\mathbf{Y} \end{bmatrix},$$

which are the first-order conditions resulting from the least squares problem (see Eq. (4.3)).

One can use different types of fixed effects in the same regression. Typically, one can have time and entity fixed effects. In that case, the model writes:

$$y_{i,t} = \mathbf{x}_i'\beta + \alpha_i + \gamma_t + \varepsilon_{i,t}.$$

The LSDV approach (Eq. (5.2)) can still be resorted to. It suffices to extend the $\mathbf{Z}$ matrix with additional columns (then called *time dummies*):

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{D} \quad \mathbf{C}] \begin{bmatrix} \beta \\ \alpha \\ \gamma \end{bmatrix} + \varepsilon, \tag{5.7}$$

with:

$$\mathbf{C} = \begin{bmatrix} \delta_1 & \delta_2 & \dots & \delta_{T-1} \\ \vdots & \vdots & & \vdots \\ \delta_1 & \delta_2 & \dots & \delta_{T-1} \end{bmatrix},$$

where the T-dimensional vector δ_t (the *time dummy*) is

$$[0, \dots, 0, \underbrace{1}_{\text{t}^{th} \text{ entry}}, 0, \dots, 0]'.$$

Using state and year fixed effects in the `CigarettesSW` panel dataset yields the following results:

```
CigarettesSW$rincome <- with(CigarettesSW, income/population/cpi)
eq.pooled <- lm(log(packs)~log(rprice)+log(rincome),data=CigarettesSW)
eq.LSDV <- lm(log(packs)~log(rprice)+log(rincome)+state,
              data=CigarettesSW)
CigarettesSW$year <- as.factor(CigarettesSW$year)
eq.LSDV2 <- lm(log(packs)~log(rprice)+log(rincome)+state+year,
               data=CigarettesSW)
stargazer::stargazer(eq.pooled,eq.LSDV,eq.LSDV2,type="text",no.space = TRUE,
                     omit=c("state","year"),
                     add.lines=list(c('State FE','No','Yes','Yes'),
                                    c('Year FE','No','No','Yes')),
                     omit.stat=c("f","ser"))
```

```
##
## ============================================
##                     Dependent variable:
##                 ---------------------------
##                           log(packs)
##                   (1)        (2)        (3)
```

```
## ------------------------------------------
## log(rprice)  -1.334*** -1.210*** -1.056***
##               (0.135)   (0.114)   (0.149)
## log(rincome)  0.318**    0.121     0.497
##               (0.136)   (0.190)   (0.304)
## Constant     10.067*** 9.954***  8.360***
##               (0.516)   (0.264)   (1.049)
## ------------------------------------------
## State FE        No        Yes       Yes
## Year FE         No        No        Yes
## Observations    96        96        96
## R2             0.552     0.966     0.967
## Adjusted R2    0.542     0.929     0.931
## ==========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

Example 5.1 (Housing prices and interest rates)**.** In this example, we want to estimate the effect of short and long-term interest rate on housing prices. The data come from the Jordà et al. (2017) dataset (see this website).

```
library(AEC);library(sandwich)
data(JST); JST <- subset(JST,year>1950);N <- dim(JST)[1]
JST$hpreal <- JST$hpnom/JST$cpi # real house price index
JST$dhpreal <- 100*log(JST$hpreal/c(NaN,JST$hpreal[1:(N-1)]))
# Put NA's when change in country:
JST$dhpreal[c(0,JST$iso[2:N]!=JST$iso[1:(N-1)])] <- NaN
JST$dhpreal[abs(JST$dhpreal)>30] <- NaN # remove extreme price change
JST$YEAR <- as.factor(JST$year) # to have time fixed effects
eq1_noFE <- lm(dhpreal ~ stir + ltrate,data=JST)
eq1_FE   <- lm(dhpreal ~ stir + ltrate + iso + YEAR,data=JST)
eq2_noFE <- lm(dhpreal ~ I(ltrate-stir),data=JST)
eq2_FE <- lm(dhpreal ~ I(ltrate-stir) + iso + YEAR,data=JST)
vcov_cluster1_noFE <- vcovHC(eq1_noFE, cluster = JST[, c("iso","YEAR")])
vcov_cluster1_FE   <- vcovHC(eq1_FE, cluster = JST[, c("iso","YEAR")])
vcov_cluster2_noFE <- vcovHC(eq2_noFE, cluster = JST[, c("iso","YEAR")])
vcov_cluster2_FE   <- vcovHC(eq2_FE, cluster = JST[, c("iso","YEAR")])
robust_se_FE1_noFE <- sqrt(diag(vcov_cluster1_noFE))
robust_se_FE1_FE   <- sqrt(diag(vcov_cluster1_FE))
robust_se_FE2_noFE <- sqrt(diag(vcov_cluster2_noFE))
robust_se_FE2_FE   <- sqrt(diag(vcov_cluster2_FE))
stargazer::stargazer(eq1_noFE, eq1_FE, eq2_noFE, eq2_FE, type = "text",
                     column.labels = c("no FE", "with FE", "no FE","with FE"),
                     omit = c("iso","YEAR","Constant"),keep.stat = "n",
                     add.lines=list(c('Country FE','No','Yes','No','Yes'),
                                    c('Year FE','No','Yes','No','Yes')),
```

```
                      se = list(robust_se_FE1_noFE,robust_se_FE1_FE,
                                robust_se_FE2_noFE,robust_se_FE2_FE))
```

```
##
## ===================================================
##                          Dependent variable:
##                    ---------------------------------
##                                  dhpreal
##                      no FE   with FE    no FE    with FE
##                       (1)      (2)       (3)       (4)
## ---------------------------------------------------
## stir               0.485***  0.532***
##                     (0.149)  (0.170)
##
## ltrate             -0.690*** -0.384**
##                     (0.164)  (0.182)
##
## I(ltrate - stir)                      -0.476*** -0.475***
##                                        (0.145)   (0.159)
##
## ---------------------------------------------------
## Country FE            No       Yes       No        Yes
## Year FE               No       Yes       No        Yes
## Observations         1,141    1,141     1,141     1,141
## ===================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

5.4 Estimation of random effects models

Here, the individual effects are assumed to be not correlated to other variables (the $\mathbf{x}_i$'s). In that context, the OLS estimator is consistent. However, it is not efficient. The GLS approach can be employed to gain efficiency.

Random-effect models write:

$$y_{i,t} = \mathbf{x}'_{it}\beta + (\alpha + \underbrace{u_i}_{\substack{\text{Random} \\ \text{heterogeneity}}}) + \varepsilon_{i,t},$$

with

$$\begin{aligned}
\mathbb{E}(\varepsilon_{i,t}|\mathbf{X}) &= \mathbb{E}(u_i|\mathbf{X}) = 0, \\
\mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,s}|\mathbf{X}) &= \begin{cases} \sigma_\varepsilon^2 & \text{if } i=j \text{ and } s=t, \\ 0 & \text{otherwise.} \end{cases} \\
\mathbb{E}(u_i u_j|\mathbf{X}) &= \begin{cases} \sigma_u^2 & \text{if } i=j, \\ 0 & \text{otherwise.} \end{cases} \\
\mathbb{E}(\varepsilon_{i,t}u_j|\mathbf{X}) &= 0 \quad \text{for all } i,\ j \text{ and } t.
\end{aligned}$$

Introducing the notations $\eta_{i,t} = u_i + \varepsilon_{i,t}$ and $\eta_i = [\eta_{i,1}, ..., \eta_{i,T}]'$, we have $\mathbb{E}(\eta_i|\mathbf{X}) = \mathbf{0}$ and $\mathbb{V}ar(\eta_i|\mathbf{X}) = \Gamma$, where

$$\Gamma = \begin{bmatrix} \sigma_\varepsilon^2+\sigma_u^2 & \sigma_u^2 & \sigma_u^2 & ... & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2+\sigma_u^2 & \sigma_u^2 & ... & \sigma_u^2 \\ \vdots & & \ddots & & \vdots \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & ... & \sigma_\varepsilon^2+\sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2\mathbf{I} + \sigma_u^2\mathbf{1}\mathbf{1}'.$$

Denoting by Σ the covariance matrix of $\eta = [\eta_1', ..., \eta_n']'$, we have:

$$\Sigma = \mathbf{I} \otimes \Gamma.$$

If we knew Σ, we would apply (feasible) GLS (Eq. (4.32), in Section 4.5.2):

$$\beta = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}.$$

(As explained in Section 4.5.2, this amounts to regressing ${\Sigma^{-1/2}}'\mathbf{y}$ on ${\Sigma^{-1/2}}'\mathbf{X}$.)

It can be checked that $\Sigma^{-1/2} = \mathbf{I} \otimes (\Gamma^{-1/2})$ where

$$\Gamma^{-1/2} = \frac{1}{\sigma_\varepsilon}\left(\mathbf{I} - \frac{\theta}{T}\mathbf{1}\mathbf{1}'\right), \quad \text{with} \quad \theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

Hence, if we knew Σ, we would transform the data as follows:

$$\Gamma^{-1/2}\mathbf{y}_i = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i,1} - \theta\bar{y}_i \\ y_{i,2} - \theta\bar{y}_i \\ \vdots \\ y_{i,T} - \theta\bar{y}_i \end{bmatrix}.$$

What about when Σ is unknown? One can take deviations from group means to remove heterogeneity:

$$y_{i,t} - \bar{y}_i = [\mathbf{x}_{i,t} - \bar{\mathbf{x}}_i]'\beta + (\varepsilon_{i,t} - \bar{\varepsilon}_i). \tag{5.8}$$

The previous equation can be consistently estimated by OLS. (Although the residuals are correlated across t's for the observations pertaining to a given entity, the OLS remain consistent; see Prop. 4.13.)

We have $\mathbb{E}\left[\sum_{i=1}^{T}(\varepsilon_{i,t}-\bar{\varepsilon}_i)^2\right]=(T-1)\sigma_\varepsilon^2$.

The $\varepsilon_{i,t}$'s are not observed but $\mathbf{b}$, the OLS estimator of β in Eq. (5.8), is a consistent estimator of β. Using an adjustment for the degrees of freedom, we can approximate their variance with:

$$\hat{\sigma}_e^2=\frac{1}{nT-n-K}\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{i,t}-\bar{e}_i)^2.$$

What about σ_u^2? We can exploit the fact that OLS are consistent in the pooled regression:

$$\text{plim } s_{pooled}^2=\text{plim }\frac{\mathbf{e}'\mathbf{e}}{nT-K-1}=\sigma_u^2+\sigma_\varepsilon^2,$$

and therefore use $s_{pooled}^2-\hat{\sigma}_e^2$ as an approximation to σ_u^2.

Let us come back to Example 5.1 (relationship between changes in housing prices and interest rates). In the following, we use the random effect specification; and compare the results with those obtained with the pooled regression and with the fixed-effect model. For that, we use the function `plm` of the package of the same name. (Note that `eq.FE` is similar to `eq1` in Example 5.1.)

```
library(plm);library(stargazer)
eq.RE <- plm(dhpreal ~ stir + ltrate,data=JST,index=c("iso","YEAR"),
             model="random",effect="twoways")
eq.FE <- plm(dhpreal ~ stir + ltrate,data=JST,index=c("iso","YEAR"),
             model="within",effect="twoways")
eq0   <- plm(dhpreal ~ stir + ltrate,data=JST,index=c("iso","YEAR"),
             model="pooling")
stargazer(eq0, eq.RE, eq.FE, type = "text",no.space = TRUE,
                    column.labels=c("Pooled","Random Effect","Fixed Effects"),
                    add.lines=list(c('State FE','No','Yes','Yes'),
                                   c('Year FE','No','Yes','Yes')),
                    omit.stat=c("f","ser"))
```

```
##
## ================================================
##                       Dependent variable:
##              -----------------------------------
##                             dhpreal
##               Pooled   Random Effect Fixed Effects
##                 (1)         (2)           (3)
## ------------------------------------------------
## stir          0.485***    0.456***      0.532***
##                (0.114)     (0.019)       (0.134)
## ltrate        -0.690***   -0.541***     -0.384***
##                (0.127)     (0.020)       (0.145)
```

```
## Constant      4.103***     3.341***
##                (0.421)      (0.096)
## ------------------------------------------------
## State FE          No           Yes           Yes
## Year FE           No           Yes           Yes
## Observations   1,141        1,141         1,141
## R2             0.027        0.024         0.015
## Adjusted R2    0.025        0.022        -0.067
## ================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

One can run an Hausman (1978) test in order to check whether or not the fixed-effect model is needed. Indeed, if this is not the case (i.e., if the covariates are not correlated to the disturbances), then it is preferable to use the random-effect estimation as the latter is more efficient.

```
phtest(eq.FE,eq.RE)
```

```
##
##  Hausman Test
##
## data:  dhpreal ~ stir + ltrate
## chisq = 3.8386, df = 2, p-value = 0.1467
## alternative hypothesis: one model is inconsistent
```

The p-value being high, we do not reject the null hypothesis according to which the covariates and the errors are uncorrelated. We should therefore prefer the random-effect model.

Example 5.2 (Spatial data)**.** This example makes use of Airbnb prices (Zürich, 22 June 2017), collected from Tom Slee's website. The covariates are the number of bedrooms and the number of people that can be accommodated. We consider the use of district fixed effects. Figure 5.3 shows the price to explain (the size of the circles is proportional to the prices). The white lines delineate the 12 districts of the city.

Let us regress prices on the covariates as well as on district dummies:

```
eq_noFE <- lm(price~bedrooms+accommodates,data=airbnb)
eq_FE   <- lm(price~bedrooms+accommodates+neighborhood,data=airbnb)
# Adjust standard errors:
cov_FE          <- vcovHC(eq_FE, cluster = airbnb[, c("neighborhood")])
robust_se_FE    <- sqrt(diag(cov_FE))
cov_noFE        <- vcovHC(eq_noFE, cluster = airbnb[, c("neighborhood")])
robust_se_noFE  <- sqrt(diag(cov_noFE))
stargazer::stargazer(eq_FE, eq_noFE, eq_FE, eq_noFE, type = "text",
```

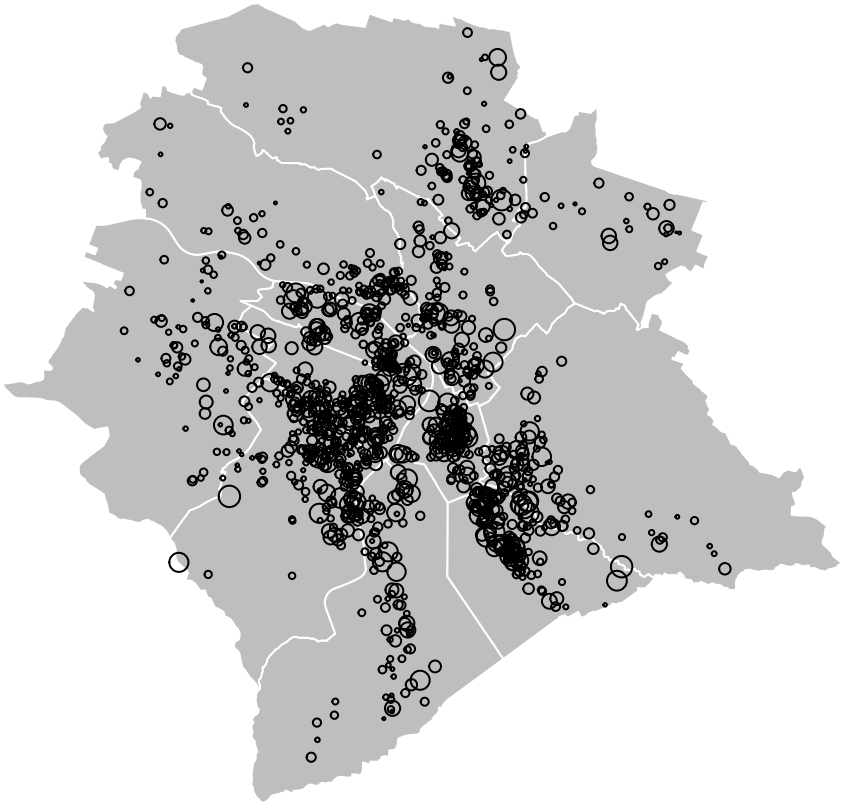


Figure 5.3: Airbnb prices for the Zurich area, 22 June 2017. The size of the circles is proportional to the prices. White lines delineate the 12 districts of the city.

```
                    column.labels = c("FE (no HAC)", "No FE (no HAC)",
                                      "FE (with HAC)", "No FE (with HAC)"),
                    omit = c("neighborhood"),no.space = TRUE,
                    omit.labels = c("District FE"),keep.stat = "n",
                    se = list(NULL, NULL, robust_se_FE, robust_se_noFE))
```

```
##
## ===============================================================
##                                 Dependent variable:
##              -------------------------------------------------
##                                        price
##              FE (no HAC) No FE (no HAC) FE (with HAC) No FE (with HAC)
##                  (1)          (2)            (3)            (4)
## ---------------------------------------------------------------
## bedrooms      7.229***     5.629**       7.229***       5.629***
##                (2.135)      (2.194)       (2.052)        (2.073)
## accommodates  16.426***    17.449***     16.426***      17.449***
##                (1.284)      (1.323)       (1.431)        (1.428)
## Constant      95.118***    68.417***     95.118***      68.417***
##                (5.323)      (3.223)       (5.664)        (3.527)
## ---------------------------------------------------------------
## District FE      Yes          No            Yes            No
## ---------------------------------------------------------------
## Observations    1,321        1,321         1,321          1,321
## ===============================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

Figure 5.4 compares the residuals with and without fixed effects. The sizes of the circles are proportional to the absolute values of the residuals, the color indicates the sign (blue for positive).

With fixed effects, the colors are better balanced within each district.

5.5 Dynamic Panel Regressions

In what precedes, it has been assumed that there is no correlation between the observations indexed by (i,t) and those indexed by (j,s) as long as $j \neq i$ or $t \neq s$. If one suspects that the errors $\varepsilon_{i,t}$ are correlated (across entities i for a given date t, or across dates for a given entity, or both), then one should employ a robust covariance matrix (see Section 4.5.5).

In several cases, auto-correlation in the variable of interest may stem from an auto-regressive specification. That is, Eq. (5.1) is then replaced by:

$$y_{i,t} = \rho y_{i,t-1} + \mathbf{x}_{i,t}' \underbrace{\beta}_{K\times 1} + \underbrace{\alpha_i}_{\text{Individual effects}} + \varepsilon_{i,t}. \tag{5.9}$$

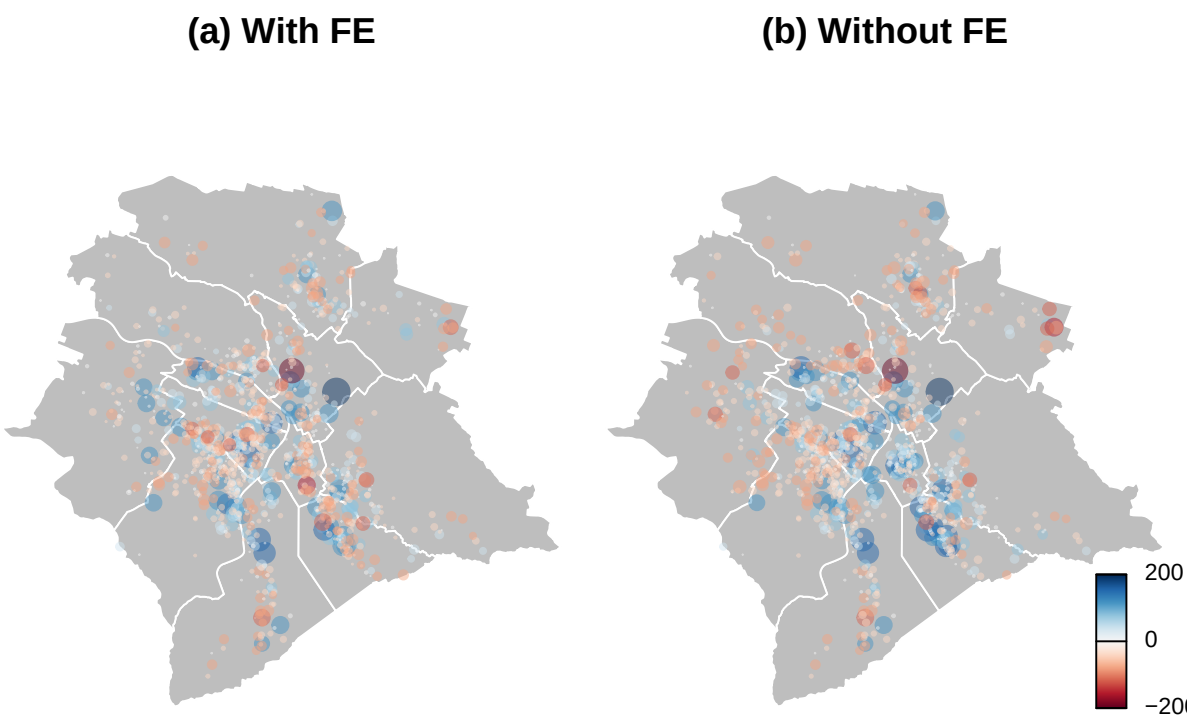


Figure 5.4: Regression residuals. The sizes of the circles are proportional to the absolute values of the residuals, the color indicates the sign (blue for negative).

In that case, even if the explanatory variables $\mathbf{x}_{i,t}$ are uncorrelated to the errors $\varepsilon_{i,t}$, we have that the additional *explanatory variable* $y_{i,t-1}$ correlates to the errors $\varepsilon_{i,t-1}, \varepsilon_{i,t-2}, \dots, \varepsilon_{i,1}$. As a result, the LSDV estimate of the model parameters $\{\rho, \beta\}$ may be biased, even if n is large. To see this, notice that the LSDV regression amounts to regressing $\tilde{\mathbf{y}}$ on $\widetilde{\mathbf{X}}$ (see Eq. (5.5)), where the elements of $\widetilde{\mathbf{X}}$ are the explanatory variables to which we subtract their within-sample means. In particular, we have:

$$\tilde{y}_{i,t-1} = y_{i,t-1} - \frac{1}{T}\sum_{s=1}^{T} y_{i,s-1},$$

which correlates to the corresponding error, that is:

$$\tilde{\varepsilon}_{i,t} = \varepsilon_{i,t} - \frac{1}{T}\sum_{s=1}^{T} \varepsilon_{i,s}.$$

The previous equation shows that the *within-group* estimator (LSDV) introduces all realizations of the $\varepsilon_{i,t}$ errors into the transformed error term ($\tilde{\varepsilon}_{i,t}$). As a result, in large-n fixed-T panels, it is consistent only if all the right-hand-side variables of the regression are strictly exogenous (i.e., do not correlate to past, present, and future errors $\varepsilon_{i,t}$).[1] This is not the case when there are lags of $y_{i,t}$ on the right-hand side of the regression formula.

[1] Although the bias may vanish for large T's, it does not if n only goes to infinity.

The following simulation illustrate this bias. The x-coordinates of the dots are the fixed effects α_i's, and the y-coordinates are their LSDV estimates. The blue line is the 45-degree line.

```
n <- 400;T <- 10
rho <- 0.8;sigma <- .5
alpha <- rnorm(n)
y <- alpha /(1-rho) + sigma^2/(1 - rho^2) * rnorm(n)
all_y <- y
for(t in 2:T){
  y <- rho * y + alpha + sigma * rnorm(n)
  all_y <- rbind(all_y,y)
}
y   <- c(all_y[2:T,]);y_1 <- c(all_y[1:(T-1),])
D <- diag(n) %x% rep(1,T-1)
Z <- cbind(c(y_1),D)
b <- solve(t(Z) %*% Z) %*% t(Z) %*% y
a <- b[2:(n+1)]
plot(alpha,a)
lines(c(-10,10),c(-10,10),col="blue")
```

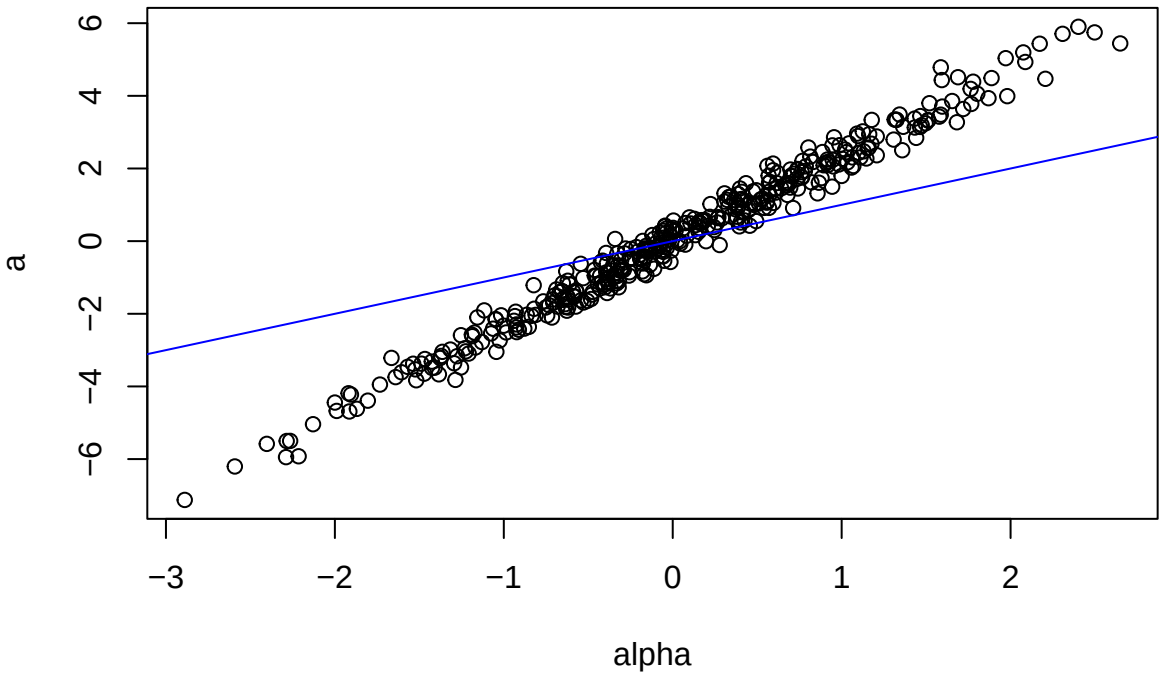


Figure 5.5: illustration of the bias pertianing to the LSDV estimation approach in the presence of auto-correlation of the depend variable.

In the previous example, the estimate of ρ (whose true value is 0.8) is 0.531.

To address this, one can resort to instrumental-variable regressions. Anderson and Hsiao (1982) have, in particular, proposed a first-differenced Two Stage Least Squares (2SLS) estimator (see Eq. (4.25) in Section 4.4). This estimation is based on the following transformation of the model:

$$\Delta y_{i,t} = \rho \Delta y_{i,t-1} + (\Delta \mathbf{x}_{i,t})'\beta + \Delta \varepsilon_{i,t}. \tag{5.10}$$

The OLS estimates of the parameters are biased because $\varepsilon_{i,t-1}$ —which is part of the error $\Delta\varepsilon_{i,t}$— is correlated to $y_{i,t-1}$ —which is part of the "explanatory variable", namely $\Delta y_{i,t-1}$. But consistent estimates can be obtained using 2SLS with instrumental variables that are correlated with $\Delta y_{i,t}$ but orthogonal to $\Delta\varepsilon_{i,t}$. One can for instance use $\{y_{i,t-2}, \mathbf{x}_{i,t-2}\}$ as instruments. Note that this approach can be implemented only if there are more than 3 time observations per entity i.

If the explanatory variables $\mathbf{x}_{i,t}$ are assumed to be predetermined (i.e., do not contemporaneous correlate with the errors $\varepsilon_{i,t}$), then $\mathbf{x}_{i,t-1}$ can be added to the instruments associated with $\Delta y_{i,t}$. Further, if these variables (the $\mathbf{x}_{i,t}$'s) are exogenous (i.e., do not contemporaneous correlate with any of the errors $\varepsilon_{i,s}$, $\forall s$), then $\mathbf{x}_{i,t}$ also constitute a valid instrument.

Using the previous (simulated) example, this approach consists in the following steps:

```
Dy    <- c(all_y[3:T,]) - c(all_y[2:(T-1),])
Dy_1 <- c(all_y[2:(T-1),]) - c(all_y[1:(T-2),])
y_2   <- c(all_y[1:(T-2),])
Z <- matrix(y_2,ncol=1)
Pz <- Z %*% solve(t(Z) %*% Z) %*% t(Z)
Dy_1hat <- Pz %*% Dy_1
rho_2SLS <- solve(t(Dy_1hat) %*% Dy_1hat) %*% t(Dy_1hat) %*% Dy
```

While the OLS estimate of ρ (whose true value is 0.8) was 0.531, we obtain here `rho_2SLS` = 0.89.

Let us come back to the general case (with covariates $\mathbf{x}_{i,k}$'s). For $t = 3$, $y_{i,1}$ (and $\mathbf{x}_{i,1}$) is the only possible instrument. However, for $t = 4$, one could use $y_{i,2}$ and $y_{i,1}$ (as well as $\mathbf{x}_{i,2}$ and $\mathbf{x}_{i,1}$). More generally, defining matrix Z_i as follows:

$$Z_i = \begin{bmatrix} \mathbf{z}'_{i,1} & 0 & \dots & & & & & & & \\ 0 & \mathbf{z}'_{i,1} & \mathbf{z}'_{i,2} & 0 & \dots & & & & & \\ 0 & 0 & 0 & \mathbf{z}_{i,1} & \mathbf{z}'_{i,2} & \mathbf{z}'_{i,3} & 0 & \dots & & \\ \vdots & & & & & & & & & \\ 0 & \dots & & & & & 0 & \mathbf{z}'_{i,1} & \dots & \mathbf{z}'_{i,T-2} \end{bmatrix},$$

where $\mathbf{z}_{i,t} = [y_{i,t}, \mathbf{x}'_{i,t}]'$, we have the moment conditions:[2]

$$\mathbb{E}(Z'_i \Delta\varepsilon_i) = 0,$$

[2] If $\mathbf{x}_{i,t}$ is predetermined (exogenous), we can use $\mathbf{z}_{i,t} = [y_{i,t}, \mathbf{x}_{i,t+1}, \mathbf{x}'_{i,t}]'$ (respectively $\mathbf{z}_{i,t} = [y_{i,t}, \mathbf{x}_{i,t+2}, \mathbf{x}_{i,t+1}, \mathbf{x}'_{i,t}]'$).

with $\Delta\varepsilon_i = [\Delta\varepsilon_{i,3}, \dots, \Delta\varepsilon_{i,T}]'$.

These restrictions are used in the GMM approach employed by Arellano and Bond (1991). Specifically, a GMM estimator of the model parameters is given by:

$$\text{argmin} \left(\frac{1}{n} \sum_{i=1}^{n} Z_i' \Delta\varepsilon_i \right)' W_n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i' \Delta\varepsilon_i \right),$$

using the weighting matrix

$$W_n = \left(\frac{1}{n} \sum_{i=1}^{n} Z_i' \widehat{\Delta\varepsilon_i \Delta\varepsilon_i}' Z_i \right)^{-1},$$

where the $\widehat{\Delta\varepsilon_i}$'s are consistent estimates of the $\Delta\varepsilon_i$'s that result from a preliminary estimation. In this sense, this estimator is a two-step GMM one.

If the disturbances are homoskedastic, then it can be shown that an asymptotically equivalent (efficient) GMM estimator can be obtained by using:

$$W_{1,n} = \left(\frac{1}{n} Z_i' H Z_i \right)^{-1},$$

where H is is $(T-2) \times (T-2)$ matrix of the form:

$$H = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{bmatrix}.$$

It is straightforward to extend these GMM methods to cases where there is more than one lag of the dependent variable on the right-hand side of the equation or in cases where disturbances feature limited moving-average serial correlation.

The `pdynmc` package allows to run these GMM approaches (see Fritsch et al. (2019)). The following lines of code allow to replicate the results of Arellano and Bond (1991):

```
library(pdynmc)
data(EmplUK, package = "plm")
dat <- EmplUK
dat[,c(4:7)]          <- log(dat[,c(4:7)])
m1 <- pdynmc(dat = dat, # name of the dataset
             varname.i = "firm", # name of the cross-section identifier
             varname.t = "year", # name of the time-series identifiers
             use.mc.diff = TRUE, # use moment conditions from equations in differences?
             use.mc.lev = FALSE, # use moment conditions from equations in levels? (i.e
```

```
            use.mc.nonlin = FALSE, # use nonlinear (quadratic) moment conditions?
            include.y = TRUE, # instruments should be derived from the lags of the
            varname.y = "emp", # name of the dependent variable in the dataset
            lagTerms.y = 2, # number of lags of the dependent variable
            fur.con = TRUE, # further control variables (covariates) are included?
            fur.con.diff = TRUE, # include further control variables in equations
            fur.con.lev = FALSE, # include further control variables in equations
            varname.reg.fur = c("wage", "capital", "output"), # covariate(s) -in th
            lagTerms.reg.fur = c(1,2,2), # number of lags of the further controls
            include.dum = TRUE, # A logical variable indicating whether dummy vari
            dum.diff = TRUE, # A logical variable indicating whether dummy variabl
            dum.lev = FALSE, # A logical variable indicating whether dummy variabl
            varname.dum = "year",
            w.mat = "iid.err", # One of the character strings c('"iid.err"', '"ide
            std.err = "corrected",
            estimation = "onestep", # One of the character strings c('"onestep"',
            opt.meth = "none" # numerical optimization procedure. When no nonlinea
)
summary(m1,digits=3)
```

```
##
## Dynamic linear panel estimation (onestep)
## Estimation steps: 1
##
## Coefficients:
##             Estimate Std.Err.rob z-value.rob Pr(>|z.rob|)
## L1.emp      0.686226    0.144594       4.746      < 2e-16 ***
## L2.emp     -0.085358    0.056016      -1.524      0.12751
## L0.wage    -0.607821    0.178205      -3.411      0.00065 ***
## L1.wage     0.392623    0.167993       2.337      0.01944 *
## L0.capital  0.356846    0.059020       6.046      < 2e-16 ***
## L1.capital -0.058001    0.073180      -0.793      0.42778
## L2.capital -0.019948    0.032713      -0.610      0.54186
## L0.output   0.608506    0.172531       3.527      0.00042 ***
## L1.output  -0.711164    0.231716      -3.069      0.00215 **
## L2.output   0.105798    0.141202       0.749      0.45386
## 1979        0.009554    0.010290       0.929      0.35289
## 1980        0.022015    0.017710       1.243      0.21387
## 1981       -0.011775    0.029508      -0.399      0.68989
## 1982       -0.027059    0.029275      -0.924      0.35549
## 1983       -0.021321    0.030460      -0.700      0.48393
## 1976       -0.007703    0.031411      -0.245      0.80646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  41 total instruments are employed to estimate 16 parameters
##  27 linear (DIF)
##  8 further controls (DIF)
##  6 time dummies (DIF)
##
## J-Test (overid restrictions):  70.82 with 25 DF, pvalue: <0.001
## F-Statistic (slope coeff):  528.06 with 10 DF, pvalue: <0.001
## F-Statistic (time dummies):  14.98 with 6 DF, pvalue: 0.0204
```

We generate novel results (`m2`) by replacing "`onestep`" with "`twostep`" (in the `estimation` field). The resulting estimated coefficients are:

```
##      L1.emp      L2.emp     L0.wage     L1.wage  L0.capital  L1.capital
##  0.62870890 -0.06518800 -0.52575951  0.31128961  0.27836190  0.01409950
##  L2.capital   L0.output   L1.output   L2.output        1979        1980
## -0.04024847  0.59192286 -0.56598515  0.10054264  0.01121551  0.02306871
##        1981        1982        1983        1976
## -0.02135806 -0.03111604 -0.01799335 -0.02336762
```

Arellano and Bond (1991) have proposed a specification test. If the model is correctly specified, then the errors of Eq. (5.10) —that is the first-difference equation— should feature non-zero first-order auto-correlations, but zero higher-order autocorrelations.

Function `mtest.fct` of package `pdynmc` implements this test. Here is its result in the present case:

```
mtest.fct(m1,order=3)
```

```
##
##  Arellano and Bond (1991) serial correlation test of degree 3
##
## data:  1step GMM Estimation
## normal = 0.045945, p-value = 0.9634
## alternative hypothesis: serial correlation of order 3 in the error terms
```

One can also implement the Hansen J-test of the over-identifying restrictions (see Section 6.1.3):

```
jtest.fct(m1)
```

```
##
##  J-Test of Hansen
```

```
##
## data:  1step GMM Estimation
## chisq = 70.82, df = 25, p-value = 2.905e-06
## alternative hypothesis: overidentifying restrictions invalid
```

```
jtest.fct(m2)
```

```
##
##  J-Test of Hansen
##
## data:  2step GMM Estimation
## chisq = 31.381, df = 25, p-value = 0.1767
## alternative hypothesis: overidentifying restrictions invalid
```

5.6 Introduction to program evaluation

This section brielfy introduces the econometrics of program evaluation. Program evaluation refer to the analysis of the causal effects of some "treatments" in a broad sense. These treatment can, e.g., correspond to the implementation (or announcement) of policy measures. A comprehensive review is proposed by Abadie and Cattaneo (2018). A seminal book on the subject is that of Angrist and Pischke (2008).

5.6.1 Presentation of the problem

To begin with, let us consider a single entity. To simplify notations, we drop the entity index (i). Let us denote by Y the outcome variable (for the variable of interest), by W is a binary variable indicating whether the considered entity has received treatment ($W = 1$) or not ($W = 0$), and by X a vector of covariates, assumed to be predetermined relative to the treatment. That is, W and X could be correlated, but the values of X have been determined before that of W (in such a way that the realization of W does not affect X). Typcally, X contains characteristics of the considered entity.

We are interested in the effect of the treatment, that is:

$$Y_1 - Y_0,$$

where Y_1 correspond to the outcome obtained under treatment, while Y_0 is the outcome obtained without it. Notice that we have:

$$Y = (1 - W)Y_0 + WY_1.$$

The problem is that observing (Y, W, X) is not sufficient to observe the treatment effect $Y_1 - Y_0$. Additional assumptions are needed to estimate it, or, more precisely, its expectations (*average treatment effect*):

$$ATE = \mathbb{E}(Y_1 - Y_0).$$

Importantly, ATE is different from the following quantity:

$$\alpha = \underbrace{\mathbb{E}(Y|W=1)}_{=\mathbb{E}(Y_1|W=1)} - \underbrace{\mathbb{E}(Y|W=0)}_{=\mathbb{E}(Y_0|W=0)},$$

that is easier to estimate. Indeed, a consistent estimate of α is the difference between the means of the outcome variables in two sub-samples: one containing only the treated entities (this gives an estimate of $\mathbb{E}(Y_1|W=1)$) and the other containing only the non-treated entities (this gives an estimate of $\mathbb{E}(Y_0|W=0)$). Coming back to ATE, the problem is that we won't have direct information regarding $\mathbb{E}(Y_0|W=1)$ and $\mathbb{E}(Y_1|W=0)$. However, these two conditional expectations are part of ATE. Indeed, $ATE = \mathbb{E}(Y_1) - \mathbb{E}(Y_0)$, and:

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_1|W=0)\mathbb{P}(W=0) + \mathbb{E}(Y_1|W=1)\mathbb{P}(W=1) \quad (5.11)$$

$$\mathbb{E}(Y_0) = \mathbb{E}(Y_0|W=0)\mathbb{P}(W=0) + \mathbb{E}(Y_0|W=1)\mathbb{P}(W=1). \quad (5.12)$$

5.6.2 Randomized controlled trials (RCTs)

In the context of Randomized controlled trials (RCTs), entities are randomly assigned to receive the treatment. As a result, we have $\mathbb{E}(Y_1) = \mathbb{E}(Y_1|W=0) = \mathbb{E}(Y_1|W=1)$ and $\mathbb{E}(Y_0) = \mathbb{E}(Y_0|W=0) = \mathbb{E}(Y_0|W=1)$. Using this into Eqs. (5.11) and (5.12) yields $ATE = \alpha$.

Therefore, in this context, estimating $\mathbb{E}(Y_1 - Y_0)$ amounts to computing the difference between two sample means, namely (a) the sample mean of the subset of Y_i's corresponding to the entities for which $W_i = 1$, and (b) the one for which $W_i = 0$.

More accurate estimates can be obtained through regressions. Assume that the model reads:

$$Y_i = W_i\beta_1 + X_i'\beta_z + \varepsilon_i,$$

where $\mathbb{E}(\varepsilon_i|X_i) = 0$ (and W_i is independent from X_i and ε_i). In this case, we obtain a consistent estimate of β_1 by regressing $\mathbf{y}$ on $\mathbf{Z} = [\mathbf{w}, \mathbf{X}]$.

5.6.3 Difference-in-Difference (DiD) approach

The DiD approach is a popular methodology implemented in cases where W cannot be considered as an independent variable. It exploits two dimensions: entities (i), and time (t). To simplify the exposition, we consider only two periods here ($t = 0$ and $t = 1$).

Consider the following model:

$$Y_{i,t} = W_{i,t}\beta_1 + \mu_i + \delta_t + \varepsilon_{i,t} \quad (5.13)$$

The parameter of interest is β_1, which is the treatment effect (recall that $W_{i,t} \in \{0,1\}$). Usually, for all entities i, we have $W_{i,t=0} = 0$. But only some of them are treated on date 1, i.e., $W_{i,1} \in \{0,1\}$.

The disturbance $\varepsilon_{i,t}$ affects the outcome, but we assume that it does not relate to the selection for treatment; therefore, $\mathbb{E}(\varepsilon_{i,t}|W_{i,t}) = 0$. By contrast, we do not exclude some correlation between $W_{i,t}$ (for $t = 1$) and μ_i; hence, μ_i may constitute a *confounder*. Finally, we suppose that the micro-variables W_i do not affect the time fixed effects δ_t, such that $\mathbb{E}(\delta_t|W_{i,t}) = \mathbb{E}(\delta_t)$.

We have:

$$\begin{array}{rcccccccc}
\mathbb{E}(Y_{i,1}|W_{i,1}=1) & = & \beta_1 & + & \mathbb{E}(\mu_i|W_{i,1}=1) & + & \mathbb{E}(\delta_1|W_{i,1}=1) & + & \mathbb{E}(\varepsilon_{i,1}) \\
\mathbb{E}(Y_{i,0}|W_{i,1}=1) & = & & & \mathbb{E}(\mu_i|W_{i,1}=1) & + & \mathbb{E}(\delta_0|W_{i,1}=1) & + & \mathbb{E}(\varepsilon_{i,0}) \\
\mathbb{E}(Y_{i,1}|W_{i,1}=0) & = & & & \mathbb{E}(\mu_i|W_{i,1}=0) & + & \mathbb{E}(\delta_1|W_{i,1}=0) & + & \mathbb{E}(\varepsilon_{i,1}) \\
\mathbb{E}(Y_{i,0}|W_{i,1}=0) & = & & & \mathbb{E}(\mu_i|W_{i,1}=0) & + & \mathbb{E}(\delta_0|W_{i,1}=0) & + & \mathbb{E}(\varepsilon_{i,0}).
\end{array}$$

and, under our assumptions, it can be checked that:

$$\beta_1 = \mathbb{E}(\Delta Y_{i,1}|W_{i,1}=1) - \mathbb{E}(\Delta Y_{i,1}|W_{i,1}=0),$$

where $\Delta Y_{i,1} = Y_{i,1} - Y_{i,0}$. Therefore, in this context, the treatment effect appears to be a difference (of two conditionnal expectations) of difference (of the outcome variable, through time).

This is illustrated by Figure 5.6, which represents the generic DiD framework.

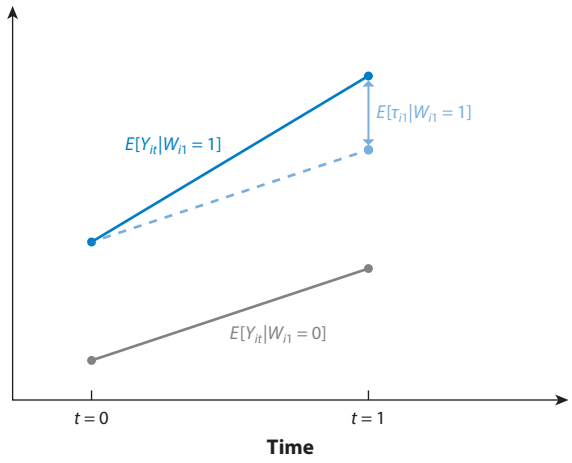


Figure 5

Identification in a difference-in-differences model. The dashed line represents the outcome that the treated units would have experienced in the absence of the treatment.

Figure 5.6: Source: Abadie et al., (1998).

In practice, implementing this approach consists in running a linear regression of the type of Eq. (5.13). These regressions also usually involve controls on top of the fixed effects μ_i. As illustrated in the next subsection, the parameter of interest (β_1) is often associated with an interaction term.

5.6.4 Application of the DiD approach

This example is based on the data used in Meyer et al. (1995). This dataset is part of the `wooldridge` package. This paper examines the effect of workers' compensation for injury on time out of work. It exploits a **natural experiment** approach of comparing individuals injured before and after increases in the maximum weekly benefit amount. Specifically, in 1980, the cap on weekly earnings covered by worker's compensation was increased in Kentucky and Michigan. Let us check whether this new policy was followed by an increase in the amount of time workers spent unemployed (for example, higher compensation may reduce workers' incentives to avoid injury).

As shown in Figure 5.7, the measure has only affected high-earning workers. The idea exploited by Meyer et al. (1995) was to compare the increase in time out of work before-after 1980 for higher-earnings workers on the one hand (entities who received the treatment) and low-earnings workers on the other hand (control group).

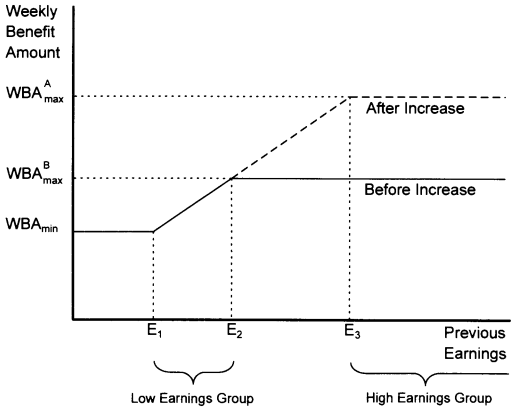


FIGURE 1. TEMPORARY TOTAL BENEFIT SCHEDULE BEFORE AND AFTER AN INCREASE IN THE MAXIMUM WEEKLY BENEFIT

Figure 5.7: Source: Meyer et al., (1995).

The next lines of codes replicate some of their results. The dependent variable is the logarithm of the duration of benefits. For more information use `?injury`, after having loaded the `wooldridge` library.

In the table of results below, the parameter of interest is the one associated with the interaction term `afchnge:highearn`. Columns 2 and 3 correspond to the first two column of Table 6 in Meyer et al. (1995).

```
library(wooldridge)
data(injury)
injury <- subset(injury,ky==1)
injury$indust <- as.factor(injury$indust)
injury$injtype <- as.factor(injury$injtype)
```

```
#names(injury)
eq1 <- lm(log(durat) ~ afchnge + highearn + afchnge*highearn,data=injury)
eq2 <- lm(log(durat) ~ afchnge + highearn + afchnge*highearn +
            lprewage*highearn + male + married + lage + ltotmed + hosp +
            indust + injtype,data=injury)
eq3 <- lm(log(durat) ~ afchnge + highearn + afchnge*highearn +
            lprewage*highearn + male + married + lage + indust +
            injtype,data=injury)
stargazer::stargazer(eq1,eq2,eq3,type="text",
                     omit=c("indust","injtype","Constant"),no.space = TRUE,
                     add.lines = list(c("industry dummy","no","yes","yes"),
                                      c("injury dummy","no","yes","yes")),
                     order = c(1,2,18,3:17,19,20),omit.stat = c("f","ser"))
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                              log(durat)
##                        (1)       (2)       (3)
## ------------------------------------------------
## afchnge                0.008    -0.004     0.016
##                       (0.045)   (0.038)   (0.045)
## highearn              0.256***  -0.595    -1.522
##                       (0.047)   (0.930)   (1.099)
## afchnge:highearn      0.191***  0.162***  0.215***
##                       (0.069)   (0.059)   (0.069)
## lprewage                        0.207**   0.258**
##                                 (0.088)   (0.104)
## male                            -0.070*   -0.072
##                                 (0.039)   (0.046)
## married                          0.055     0.051
##                                 (0.035)   (0.041)
## lage                            0.244***  0.252***
##                                 (0.044)   (0.052)
## ltotmed                         0.361***
##                                 (0.011)
## hosp                            0.252***
##                                 (0.044)
## highearn:lprewage                0.065     0.232
##                                 (0.158)   (0.187)
## ------------------------------------------------
## industry dummy          no        yes       yes
## injury dummy            no        yes       yes
```

```
## Observations        5,626     5,347     5,347
## R2                  0.021     0.319     0.049
## Adjusted R2         0.020     0.316     0.046
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Chapter 6

Estimation Methods

This chapter presents three approaches to estimate parametric models: the General Method of Moments (GMM), the Maximum Likelihood approach (ML), and the Bayesian approach. The general context is the following: You observe a sample $\mathbf{y} = \{y_1, ..., y_n\}$, you assume that these data have been generated by a model parameterized by $\theta \in \mathbb{R}^K$, and you want to estimate this vector θ_0.

6.1 Generalized Method of Moments (GMM)

6.1.1 Definition of the GMM estimator

We denote by y_i a $p \times 1$ vector of variables; by θ an $K \times 1$ vector of parameters, and by $h(y_i; \theta)$ a continuous $r \times 1$ vector-valued function.

We denote by θ_0 the true value of θ and we assume that θ_0 satisfies:

$$\mathbb{E}[h(y_i; \theta_0)] = \mathbf{0}.$$

We denote by $\underline{y_i}$ the information contained in the current and past observations of y_i, that is: $\underline{y_i} = \{y_i, y_{i-1}, ..., y_1\}$. We denote by $g(\underline{y_n}; \theta)$ the sample average of the $h(y_i; \theta)$ vectors, i.e.:

$$g(\underline{y_n}; \theta) = \frac{1}{n} \sum_{i=1}^{n} h(y_i; \theta).$$

The intuition behind the GMM estimator is the following: choose θ so as to make the sample moment as close as possible to their population values, that is 0.

Definition 6.1. A GMM estimator of θ_0 is given by:

$$\widehat{\theta}_n = \text{argmin}_\theta \quad g(\underline{y_n}; \theta)' \, W_n \, g(\underline{y_n}; \theta),$$

where W_n is a positive definite matrix (that may depend on $\underline{y_n}$).

In the specific case where $K = r$ (the dimension of θ is the same as that of $h(y_i;\theta)$ —or of $g(\underline{y_n};\theta)$— then $\hat{\theta}_n$ satisfies:

$$g(\underline{y_n};\hat{\theta}_n) = \mathbf{0}.$$

Under regularity and identification conditions, this estimator is consistent, that is $\hat{\theta}_n$ converges towards θ_0 in probability, which we denote by:

$$\text{plim}_n\, \hat{\theta}_n = \theta_0, \quad \text{or} \quad \hat{\theta}_n \xrightarrow{p} \theta_0, \tag{6.1}$$

i.e. $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta_0| > \varepsilon) = 0$ (this is Definition 9.16).

Definition 6.1 involves a positive definite matrix W_n. While one can take any positive definite matrix to have consistency (Eq. (6.1)), it can be shown that he GMM estimator achieves the minimum asymptotic variance when W_n is the inverse of matrix S, the latter being defined by:

$$S = Asy.\mathbb{V}ar\left(\sqrt{n}g(\underline{y_n};\hat{\theta}_n)\right).$$

In this case, W_n is said to be the *optimal weighting matrix*.

The intuition behind this result is the same that underlies Generalized Least Squares (see Section 4.5.2), that is: it is beneficial to use a criterion in which the weights are inversely proportional to the variances of the moments.

If $h(x_i;\theta_0)$ is not correlated to $h(x_j;\theta_0)$, for $i \neq j$, then we have:

$$S = \mathbb{V}ar(h(x_i;\theta_0)),$$

which can be approximated by

$$\hat{\Gamma}_{0,n} = \frac{1}{n}\sum_{i=1}^{n} h(x_i;\hat{\theta}_n)h(x_i;\hat{\theta}_n)'.$$

In a time series context, we often have correlation between x_i and x_{i+k}, especially for small k's. In this case, and if the time series $\{y_i\}$ is covariance stationary (see Def. 8.4), then we have:

$$S := \sum_{\nu=-\infty}^{\infty} \Gamma_\nu,$$

where $\Gamma_\nu := \mathbb{E}[h(x_i;\theta_0)h(x_{i-\nu};\theta_0)']$. Matrix S is called the **long-run variance** of process $\{y_i\}$ (see Def. 8.9).

For $\nu \geq 0$, let us define $\hat{\Gamma}_{\nu,n}$ by:

$$\hat{\Gamma}_{\nu,n} = \frac{1}{n}\sum_{i=\nu+1}^{n} h(x_i;\hat{\theta}_n)h(x_{i-\nu};\hat{\theta}_n)',$$

then S can be approximated by the Newey and West (1987) formula (similar to Eq. (8.6)):

$$\hat{\Gamma}_{0,n} + \sum_{\nu=1}^{q}\left[1 - \frac{\nu}{q+1}\right](\hat{\Gamma}_{\nu,n} + \hat{\Gamma}'_{\nu,n}). \tag{6.2}$$

6.1.2 Asymptotic distribution of the GMM estimator

We have:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, V),} \tag{6.3}$$

where $V = (DS^{-1}D')^{-1}$, with

$$D := \mathbb{E}\left(\frac{\partial h(x_i;\theta)}{\partial \theta'}\right)\Big|_{\theta=\theta_0}.$$

Matrix V can be approximated by

$$\hat{V}_n = (\hat{D}_n \hat{S}_n^{-1} \hat{D}_n')^{-1}, \tag{6.4}$$

where $\hat{S}_n$ is given by Eq. (6.2) and

$$\hat{D}_n' := \frac{\partial g(\underline{y_n};\theta)}{\partial \theta'}\Big|_{\theta=\hat{\theta}_n}.$$

In practice, the previous matrix is computed numerically.

6.1.3 Testing hypotheses in the GMM framework

A first important test is the one concerning the validity of the moment restrictions (Sargan-Hansen test; Sargan (1958) and Hansen (1982)). Assume that the number of restrictions imposed is larger than the number of parameters to estimate ($r > K$). In this case, the restrictions are said to be over-identifiying.

Under correct specification, we asymptotically have:

$$\sqrt{n} g(\underline{y_n};\theta_0) \sim \mathcal{N}(0, S).$$

As a result, it comes that:

$$J_n = \left(\sqrt{n} g(\underline{y_n};\theta_0)\right)' S^{-1} \left(\sqrt{n} g(\underline{y_n};\theta_0)\right) \tag{6.5}$$

asymptotically follows a χ^2 distribution. The number of degrees of freedom is equal to $r-K$. (Note that, for $r = K$, we have, as expected, $J = 0$.) That is, asymptotically:

$$J_n \sim \chi^2(r-K).$$

The GMM framework also allows to easily test linear restrictions on the parameters. First, given Eq. (6.3), Wald tests (see Eq. (4.16) in Section 4.2.5) are readily available. Second, one can also resort to a test equivalent to the *likelihood ratio tests* (see Definition 6.8). More precisely, consider an unconstrained model and a constrained version of this model, the number of restrictions being equal to k. If the two models are estimated by considering the same moment constraints, and the same weighting matrix —using Eq. (6.4), based on the unrestricted model—, then we have that:

$$n\left[(g(\underline{y_n};\hat{\theta}_n^*) - g(\underline{y_n};\hat{\theta}_n)\right] \sim \chi^2(k),$$

where $\hat{\theta}_n^*$ is the constrained estimate of θ_0.

6.1.4 Example: Estimation of the Stochastic Discount Factor (s.d.f.)

Under the no-arbitrage assumption, there exists a random variable $\mathcal{M}_{t,t+1}$ (a s.d.f.) such that

$$\mathbb{E}_t(\mathcal{M}_{t,t+1}R_{t+1}) = 1$$

for any (gross) asset return R_t. In the following, R_t denotes a n_r-dimensional vector of gross returns.

We consider the following specification of the s.d.f.:

$$\mathcal{M}_{t,t+1} = 1 - \mathbf{b}_M'(F_{t+1} - \mathbb{E}_t(F_{t+1})), \tag{6.6}$$

where F_t is a vector of factors. Eq. (6.6) then reads:

$$\mathbb{E}_t([1 - \mathbf{b}_M'(F_{t+1} - \mathbb{E}_t(F_{t+1}))]R_{t+1}) = 1.$$

Assume that the date-t information set is $\mathcal{I}_t = \{\mathbf{z}_t, \mathcal{I}_{t-1}\}$, where $\mathbf{z}_t$ is a vector of variables observed on date t. (We then have $\mathbb{E}_t(\bullet) \equiv \mathbb{E}(\bullet|\mathcal{I}_t)$.)

We can use $\mathbf{z}_t$ as an instrument. Indeed, we have:

$$\begin{aligned} & \mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]) \\ = \; & \mathbb{E}(\mathbb{E}_t\{z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]\}) \\ = \; & \mathbb{E}(z_{i,t}\underbrace{\mathbb{E}_t\{\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1\}}_{1-\mathbb{E}_t(\mathcal{M}_{t,t+1}R_{t+1})=0}) = 0. \end{aligned} \tag{6.7}$$

We have then converted a conditional moment condition into a unconditional one (which we need to implement the GMM approach described above). However, at that stage, we can still not directly use the GMM formulas because of the conditional expectation $\mathbb{E}_t(F_{t+1})$ that appears in $\mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]) = 0$.

To go further, let us assume that:

$$\mathbb{E}_t(F_{t+1}) = \mathbf{b}_F\mathbf{z}_t.$$

We can then easily estimate matrix $\mathbf{b}_F$ (of dimension $n_F \times n_z$) by OLS. Note here that these OLS can be seen as a special GMM case. Indeed, as was done in Eq. (6.7), we can show that, for the j^{th} component of F_t, we have:

$$\mathbb{E}([F_{j,t+1} - \mathbf{b}_{F,j}\mathbf{z}_t]\mathbf{z}_t) = 0,$$

where $\mathbf{b}_{F,j}$ denotes the j^{th} row of $\mathbf{b}_F$. This yields the OLS formula.

Equipped with $\mathbf{b}_F$, we rely on the following moment restrictions to estimate $\mathbf{b}_M$:

$$\mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbf{b}_F\mathbf{z}_t\}R_{t+1} - R_{t+1} + 1]) = 0.$$

Specifically, the number of restrictions is $n_R \times n_z$. Let us implement this approach in the U.S. context, using data extracted from the FRED database. In factor F_t, we use the changes in the VIX and in the personal consumption expenditures. The returns (R_t) are based on the Wilshire 5000 Price Index (a stock price index) and on the ICE BofA BBB US Corporate Index Total Return Index (a bond return index).

```
library(fredr)
fredr_set_key("df65e14c054697a52b4511e77fcfa1f3")
start_date <- as.Date("1990-01-01"); end_date <- as.Date("2022-01-01")
f <- function(ticker){
  fredr(series_id = ticker,
        observation_start = start_date,observation_end = end_date,
        frequency = "m",aggregation_method = "avg")
}
vix <- f("VIXCLS") # VIX
pce <- f("PCE") # Personal consumption expenditures
sto <- f("WILL5000PRFC") # Wilshire 5000 Full Cap Price Index
bdr <- f("BAMLCC0A4BBBTRIV") # ICE BofA BBB US Corp. Index Tot. Return
T <- dim(vix)[1]
dvix <- c(vix$value[3:T]/vix$value[2:(T-1)]) # change in VIX t+1
dpce <- c(pce$value[3:T]/pce$value[2:(T-1)]) # change in PCE t+1
dsto <- c(sto$value[3:T]/sto$value[2:(T-1)]) # return t+1
dbdr <- c(bdr$value[3:T]/bdr$value[2:(T-1)]) # return t+1
dvix_1 <- c(vix$value[2:(T-1)]/vix$value[1:(T-2)]) # change in VIX t
dpce_1 <- c(pce$value[2:(T-1)]/pce$value[1:(T-2)]) # change in PCE t
dsto_1 <- c(sto$value[2:(T-1)]/sto$value[1:(T-2)]) # return t
dbdr_1 <- c(bdr$value[2:(T-1)]/bdr$value[1:(T-2)]) # return t
```

Define the matrices containing the F_{t+1}, $\mathbf{z}_t$, and R_{t+1} vectors:

```
F_tp1 <- cbind(dvix,dpce)
Z     <- cbind(1,dvix_1,dpce_1,dsto_1,dbdr_1)
b_F <- t(solve(t(Z) %*% Z) %*% t(Z) %*% F_tp1)
F_innov <- F_tp1 - Z %*% t(b_F)
R_tp1 <- cbind(dsto,dbdr)
n_F <- dim(F_tp1)[2]; n_R <- dim(R_tp1)[2]; n_z <- dim(Z)[2]
```

Function `f_aux` compute the $h(x_t;\theta)$ and the $g(\underline{y_T};\theta)$; function `f2beMin` is the function to be minimized.

```
f_aux <- function(theta){
  b_M <- matrix(theta[1:n_F],ncol=1)
  R_aux <- matrix(F_innov %*% b_M,T-2,n_R) * R_tp1 - R_tp1 + 1
  H <- (R_aux %x% matrix(1,1,n_z)) * (matrix(1,1,n_R) %x% Z)
```

```
  g <- matrix(apply(H,2,mean),ncol=1)
  return(list(g=g,H=H))
}
f2beMin <- function(theta,W){# function to be minimized
  res <- f_aux(theta)
  return(t(res$g) %*% W %*% res$g)
}
```

Now, let's minimize this function, using use the BFGS numerical algorithm (part of the `optim` wrapper). We run 5 iterations (where W is updated).

```
library(AEC)
theta <- c(rep(0,n_F)) # inital value
for(i in 1:10){# recursion on W
  res <- f_aux(theta)
  W <-  solve(NW.LongRunVariance(res$H,q=6))
  res.optim <- optim(theta,f2beMin,W=W,
                     method="BFGS", # could be "Nelder-Mead"
                     control=list(trace=FALSE,maxit=200),hessian=TRUE)
  theta <- res.optim$par
}
```

Finally, let's compute the standard deviation of the parameter estimates, using Eq. (6.4):

```
eps <- .0001
g0 <- f_aux(theta)$g
D <- NULL
for(i in 1:length(theta)){
  theta.i <- theta
  theta.i[i] <- theta.i[i] + eps
  gi <- f_aux(theta.i)$g
  D <- cbind(D,(gi-g0)/eps)
}
V <- 1/T * solve(t(D) %*% W %*% D)
std.dev <- sqrt(diag(V));t.stud <- theta/std.dev
cbind(theta,std.dev,t.stud)
```

```
##            theta    std.dev     t.stud
## [1,]  -0.7180716  0.4646617 -1.5453642
## [2,] -11.2042452 17.1039449 -0.6550679
```

The Hansen statistic can be used to test the model (see Eq. (6.5)). If the model is correct, we have:

$$T g(\underline{y_T};\theta)' \, S^{-1} \, g(\underline{y_T};\theta) \sim \; i.i.d. \, \chi^2(J-K),$$

where J is the number of moment constraints ($n_z \times n_r$ here) and K is the number of estimated parameters ($= n_F$ here).

```
g <- f_aux(theta)$g
Hansen_stat <- T * t(g) %*% W %*% g
pvalue <- pchisq(q = Hansen_stat,df = n_R*n_z - n_F)
pvalue
```

```
##           [,1]
## [1,] 0.8789782
```

6.2 Maximum Likelihood Estimation

6.2.1 Intuition

Intuitively, the *Maximum Likelihood Estimation (MLE)* consists in looking for the value of θ that is such that the probability of having observed $\mathbf{y}$ (the sample at hand) is the highest possible.

To set an example, assume that the time periods between the arrivals of two customers in a shop, denoted by y_i, are i.i.d. and follow an exponential distribution, i.e. $y_i \sim i.i.d.\, \mathcal{E}(\lambda)$. You have observed these arrivals for some time, thereby constituting a sample $\mathbf{y} = \{y_1, ..., y_n\}$. You want to estimate λ (i.e. in that case, the vector of parameters is simply $\theta = \lambda$).

The density of Y (one observation) is $f(y;\lambda) = \frac{1}{\lambda}\exp(-y/\lambda)$. Fig. 6.1 represents such density functions for different values of λ.

Your 200 observations are reported at the bottom of Fig. 6.1 (red bars). You build the histogram and display it on the same chart.

What is your estimate of λ? Intuitively, one is led to take the λ for which the (theoretical) distribution is the closest to the histogram (that can be seen as an "empirical distribution"). This approach is consistent with the idea of picking the λ for which the probability of observing the values included in $\mathbf{y}$ is the highest.

Let us be more formal. Assume that you have only four observations: $y_1 = 1.1$, $y_2 = 2.2$, $y_3 = 0.7$ and $y_4 = 5.0$. What was the probability of jointly observing:

- $1.1 - \varepsilon \leq Y_1 < 1.1 + \varepsilon$,
- $2.2 - \varepsilon \leq Y_2 < 2.2 + \varepsilon$,
- $0.7 - \varepsilon \leq Y_3 < 0.7 + \varepsilon$, and
- $5.0 - \varepsilon \leq Y_4 < 5.0 + \varepsilon$?

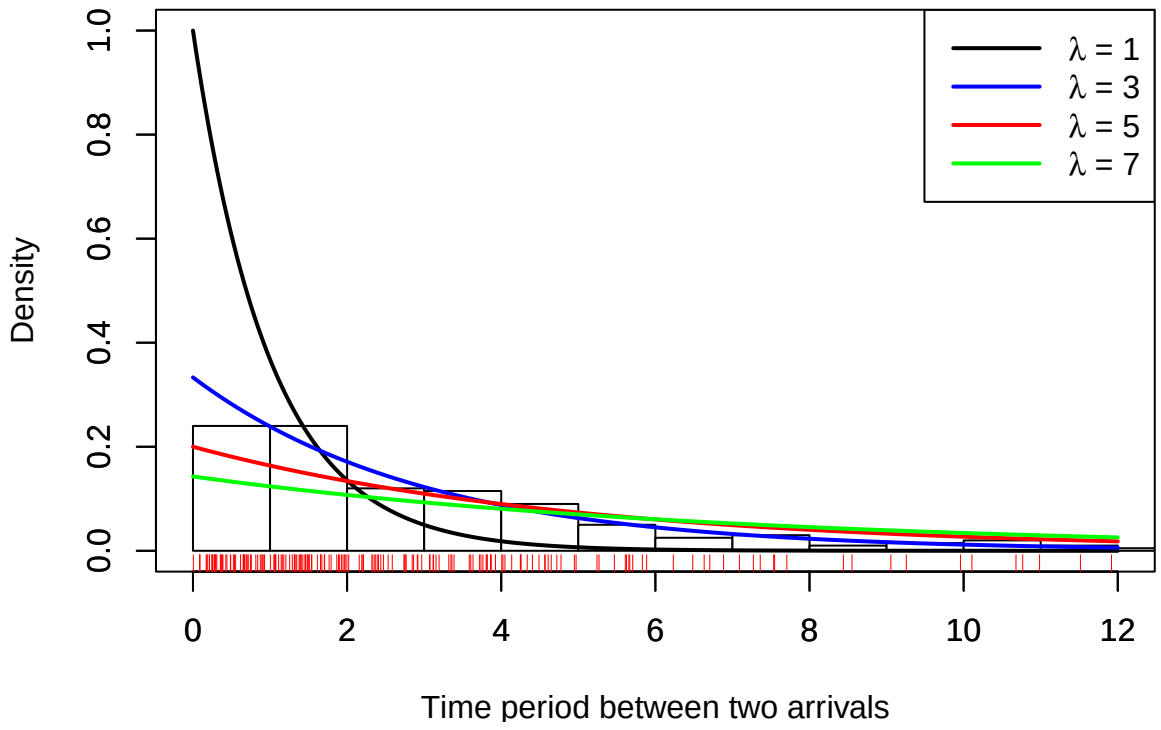


Figure 6.1: The red ticks, at the bottom, indicate observations (there are 200 of them). The historgram is based on these 200 observations

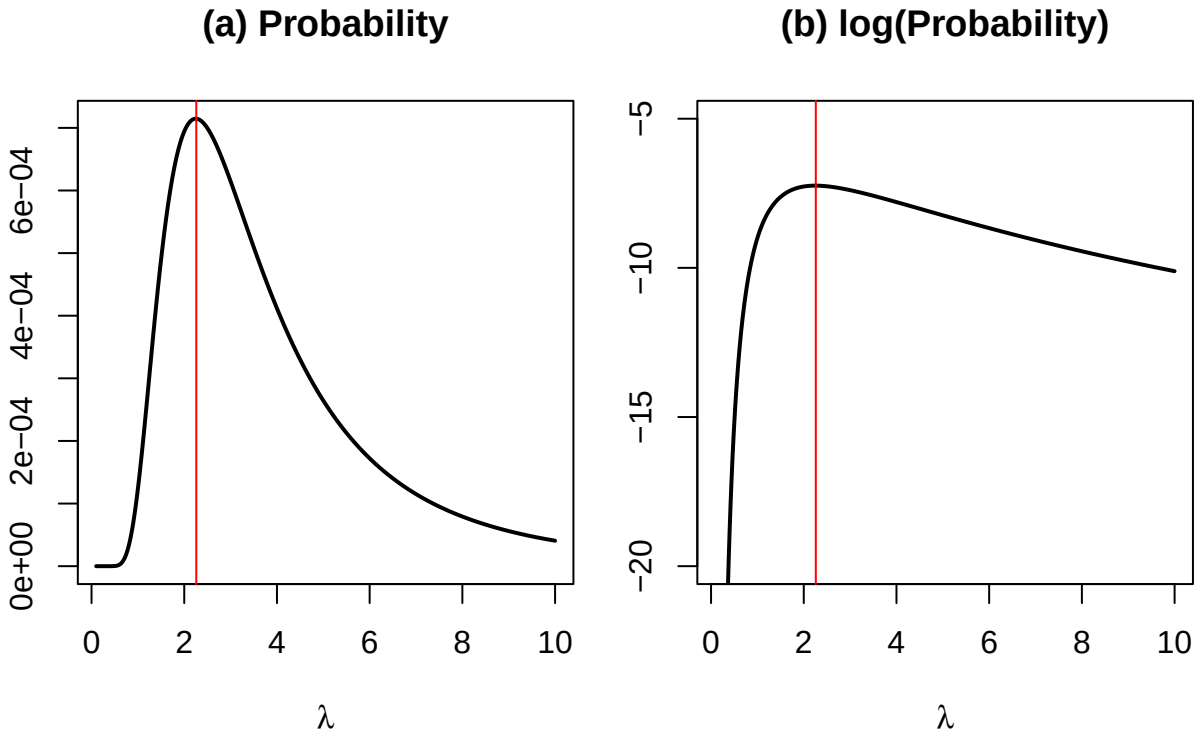


Figure 6.2: Proba. that $y_i - \varepsilon \leq Y_i < y_i + \varepsilon$, $i \in \{1, 2, 3, 4\}$. The vertical red line indicates the maximum of the function.

Because the y_i's are i.i.d., this probability is $\prod_{i=1}^{4}(2\varepsilon f(y_i, \lambda))$. The next plot shows the probability (divided by $16\varepsilon^4$, which does not depend on λ) as a function of λ.

The value of λ that maximizes the probability is 2.26.

Let us come back to the example with 200 observations:

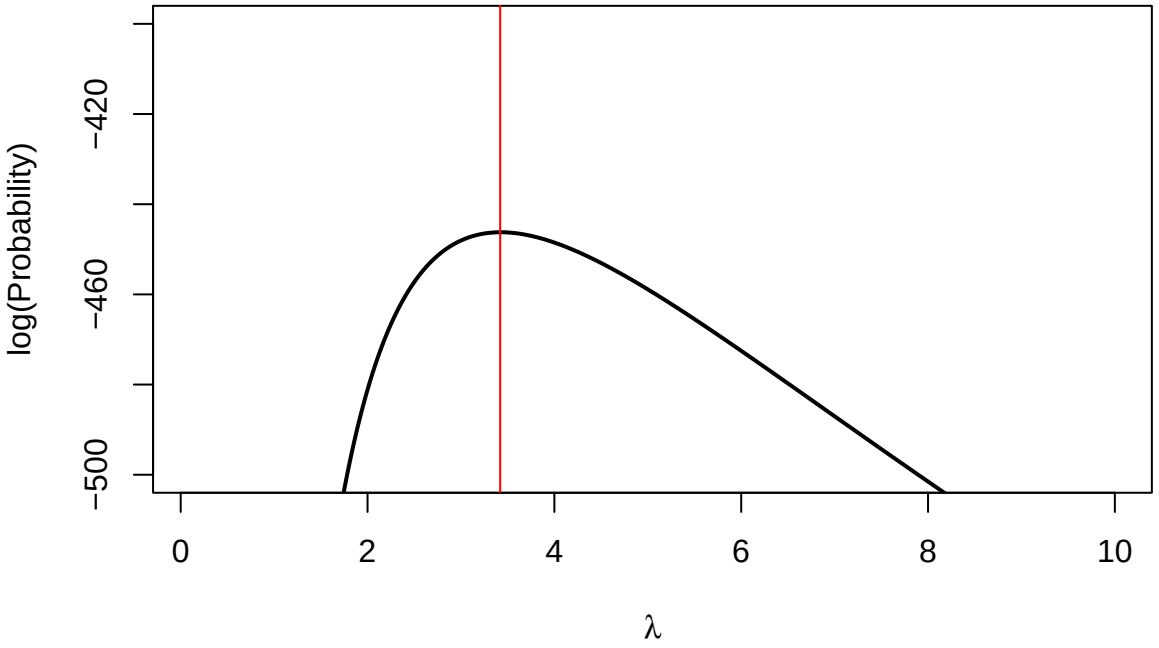


Figure 6.3: Log-likelihood function associated with the 200 i.i.d. observations. The vertical red line indicates the maximum of the function.

In that case, the value of λ that maxmimizes the probability is 3.42.

6.2.2 Definition and properties

$f(y;\theta)$ denotes the probability density function (p.d.f.) of a random variable Y which depends on a set of parameters θ. The density of n independent and identically distributed (i.i.d.) observations of Y is given by:

$$f(\mathbf{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta),$$

where $\mathbf{y}$ denotes the vector of observations; $\mathbf{y} = \{y_1, \dots, y_n\}$.

Definition 6.2 (Likelihood function)**.** The likelihood function is:

$$\mathcal{L} : \theta \to \mathcal{L}(\theta;\mathbf{y}) = f(\mathbf{y};\theta) = f(y_1, \dots, y_n;\theta).$$

We often work with $\log \mathcal{L}$, the **log-likelihood function**.

Example 6.1 (Gaussian distribution)**.** If $y_i \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\log \mathcal{L}(\theta; \mathbf{y}) = -\frac{1}{2} \sum_{i=1}^{n} \left(\log \sigma^2 + \log 2\pi + \frac{(y_i - \mu)^2}{\sigma^2} \right).$$

Definition 6.3 (Score)**.** The score $S(y; \theta)$ is given by $\frac{\partial \log f(y;\theta)}{\partial \theta}$.

If $y_i \sim \mathcal{N}(\mu, \sigma^2)$ (Example 6.1), then

$$\frac{\partial \log f(y;\theta)}{\partial \theta} = \begin{bmatrix} \dfrac{\partial \log f(y;\theta)}{\partial \mu} \\ \dfrac{\partial \log f(y;\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{y-\mu}{\sigma^2} \\ \frac{1}{2\sigma^2}\left(\frac{(y-\mu)^2}{\sigma^2} - 1 \right) \end{bmatrix}.$$

Proposition 6.1 (Score expectation)**.** *The expectation of the score is zero.*

Proof. We have:

$$\begin{aligned} \mathbb{E}\left(\frac{\partial \log f(Y;\theta)}{\partial \theta} \right) &= \int \frac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta) dy \\ &= \int \frac{\partial f(y;\theta)/\partial \theta}{f(y;\theta)} f(y;\theta) dy = \frac{\partial}{\partial \theta} \int f(y;\theta) dy \\ &= \partial 1 / \partial \theta = 0, \end{aligned}$$

which gives the result. □

Definition 6.4 (Fisher information matrix)**.** The information matrix is (minus) the the expectation of the second derivatives of the log-likelihood function:

$$\mathcal{I}_Y(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(Y;\theta)}{\partial \theta \partial \theta'} \right).$$

Proposition 6.2. *We have*

$$\mathcal{I}_Y(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(Y;\theta)}{\partial \theta} \right) \left(\frac{\partial \log f(Y;\theta)}{\partial \theta} \right)' \right] = \mathbb{V}ar[S(Y;\theta)].$$

Proof. We have $\frac{\partial^2 \log f(Y;\theta)}{\partial \theta \partial \theta'} = \frac{\partial^2 f(Y;\theta)}{\partial \theta \partial \theta'} \frac{1}{f(Y;\theta)} - \frac{\partial \log f(Y;\theta)}{\partial \theta} \frac{\partial \log f(Y;\theta)}{\partial \theta'}$. The expectation of the first right-hand side term is $\partial^2 1/(\partial \theta \partial \theta') = \mathbf{0}$, which gives the result. □

Example 6.2. If $y_i \sim$ *i.i.d.* $\mathcal{N}(\mu, \sigma^2)$, let $\theta = [\mu, \sigma^2]'$ then

$$\frac{\partial \log f(y;\theta)}{\partial \theta} = \left[\frac{y-\mu}{\sigma^2} \quad \frac{1}{2\sigma^2}\left(\frac{(y-\mu)^2}{\sigma^2} - 1 \right) \right]',$$

and

$$\mathcal{I}_Y(\theta) = \mathbb{E}\left(\frac{1}{\sigma^4} \begin{bmatrix} \sigma^2 & y-\mu \\ y-\mu & \frac{(y-\mu)^2}{\sigma^2} - \frac{1}{2} \end{bmatrix} \right) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

Proposition 6.3 (Additive property of the Information matrix)**.** *The information matrix resulting from two independent experiments is the sum of the information matrices:*

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta).$$

Proof. Directly deduced from the definition of the information matrix (Def. 6.4), using that the epxectation of a product of independent variables is the product of the expectations. □

Theorem 6.1 (Frechet-Darmois-Cramer-Rao bound)**.** *Consider an unbiased estimator of θ denoted by $\widehat{\theta}(Y)$. The variance of the random variable $\omega'\widehat{\theta}$ (which is a linear combination of the components of $\widehat{\theta}$) is larger than:*

$$(\omega'\omega)^2/(\omega'\mathcal{I}_Y(\theta)\omega).$$

Proof. The Cauchy-Schwarz inequality implies that $\sqrt{\mathbb{V}ar(\omega'\widehat{\theta}(Y))\mathbb{V}ar(\omega'S(Y;\theta))} \geq |\omega'\mathbb{C}ov[\widehat{\theta}(Y), S(Y;\theta)]\omega|$. Now, $\mathbb{C}ov[\widehat{\theta}(Y), S(Y;\theta)] = \int_y \widehat{\theta}(y)\frac{\partial \log f(y;\theta)}{\partial\theta} f(y;\theta)dy = \frac{\partial}{\partial\theta}\int_y \widehat{\theta}(y)f(y;\theta)dy = \mathbf{I}$ because $\widehat{\theta}$ is unbiased. Therefore $\mathbb{V}ar(\omega'\widehat{\theta}(Y)) \geq \mathbb{V}ar(\omega'S(Y;\theta))^{-1}(\omega'\omega)^2$. Prop. 6.2 leads to the result. □

Definition 6.5 (Identifiability)**.** The vector of parameters θ is identifiable if, for any other vector θ^*:

$$\theta^* \neq \theta \Rightarrow \mathcal{L}(\theta^*;\mathbf{y}) \neq \mathcal{L}(\theta;\mathbf{y}).$$

Definition 6.6 (Maximum Likelihood Estimator (MLE))**.** The maximum likelihood estimator (MLE) is the vector θ that maximizes the likelihood function. Formally:

$$\theta_{MLE} = \arg\max_\theta \mathcal{L}(\theta;\mathbf{y}) = \arg\max_\theta \log\mathcal{L}(\theta;\mathbf{y}). \tag{6.8}$$

Definition 6.7 (Likelihood equation)**.** A necessary condition for maximizing the likelihood function (under regularity assumption, see Hypotheses 6.1) is:

$$\frac{\partial \log\mathcal{L}(\theta;\mathbf{y})}{\partial\theta} = \mathbf{0}. \tag{6.9}$$

Hypothesis 6.1 (Regularity assumptions)**.** We have:

i. $\theta \in \Theta$ where Θ is compact.
ii. θ_0 is identified.
iii. The log-likelihood function is continuous in θ.
iv. $\mathbb{E}_{\theta_0}(\log f(Y;\theta))$ exists.
v. The log-likelihood function is such that $(1/n)\log\mathcal{L}(\theta;\mathbf{y})$ converges almost surely to $\mathbb{E}_{\theta_0}(\log f(Y;\theta))$, uniformly in $\theta \in \Theta$.
vi. The log-likelihood function is twice continuously differentiable in an open neighborood of θ_0.
vii. The matrix $\mathbf{I}(\theta_0) = -\mathbb{E}_0\left(\frac{\partial^2 \log\mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'}\right)$ —the Fisher Information matrix— exists and is nonsingular.

Proposition 6.4 (Properties of MLE)**.** *Under regularity conditions (Assumptions 6.1), the MLE is:*

a. ***Consistent****: plim* $\theta_{MLE} = \theta_0$ *(*θ_0 *is the true vector of parameters).*
b. ***Asymptotically normal****:*

$$\boxed{\sqrt{n}(\theta_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y(\theta_0)^{-1}).} \tag{6.10}$$

c. ***Asymptotically efficient****:* θ_{MLE} *is asymptotically efficient and achieves the Freechet-Darmois-Cramer-Rao lower bound for consistent estimators.*
d. ***Invariant****: The MLE of* $g(\theta_0)$ *is* $g(\theta_{MLE})$ *if* g *is a continuous and continuously differentiable function.*

Proof. See Appendix 9.5. □

Since $\mathcal{I}_Y(\theta_0) = \frac{1}{n}\mathbf{I}(\theta_0)$, the asymptotic covariance matrix of the MLE is $[\mathbf{I}(\theta_0)]^{-1}$, that is:

$$[\mathbf{I}(\theta_0)]^{-1} = \left[-\mathbb{E}_0\left(\frac{\partial^2 \log \mathcal{L}(\theta; \mathbf{y})}{\partial\theta\partial\theta'}\right)\right]^{-1}.$$

A direct (analytical) evaluation of this expectation is often out of reach. It can however be estimated by, either:

$$\hat{\mathbf{I}}_1^{-1} = \left(-\frac{\partial^2 \log \mathcal{L}(\theta_{MLE}; \mathbf{y})}{\partial\theta\partial\theta'}\right)^{-1}, \tag{6.11}$$

$$\hat{\mathbf{I}}_2^{-1} = \left(\sum_{i=1}^{n} \frac{\partial \log \mathcal{L}(\theta_{MLE}; y_i)}{\partial\theta} \frac{\partial \log \mathcal{L}(\theta_{MLE}; y_i)}{\partial\theta'}\right)^{-1}. \tag{6.12}$$

Asymptotically, we have $(\hat{\mathbf{I}}_1^{-1})\hat{\mathbf{I}}_2 = Id$, that is, the two formulas provide the same result.

In case of (suspected) misspecification, one can use the so-called *sandwich estimator* of the covariance matrix.[1] This covariance matrix is given by:

$$\hat{\mathbf{I}}_3^{-1} = \hat{\mathbf{I}}_2^{-1}\hat{\mathbf{I}}_1\hat{\mathbf{I}}_2^{-1}. \tag{6.13}$$

6.2.3 To sum up – MLE in practice

To implement MLE, we need:

- A parametric model (depending on the vector of parameters θ whose "true" value is θ_0) is specified.

[1] For more details, see, e.g., Charles Geyer's lectures notes.

- i.i.d. sources of randomness are identified.
- The density associated to one observation y_i is computed analytically (as a function of θ): $f(y;\theta)$.
- The log-likelihood is $\log \mathcal{L}(\theta;\mathbf{y}) = \sum_i \log f(y_i;\theta)$.
- The MLE estimator results from the optimization problem (this is Eq. (6.8)):

$$\theta_{MLE} = \arg\max_{\theta} \log \mathcal{L}(\theta;\mathbf{y}). \tag{6.14}$$

- We have: $\theta_{MLE} \sim \mathcal{N}(\theta_0, \mathbf{I}(\theta_0)^{-1})$, where $\mathbf{I}(\theta_0)^{-1}$ is estimated by means of Eq. (6.11), Eq. (6.12), or Eq. (6.13). Most of the time, this computation is numerical.

6.2.4 Example: MLE estimation of a mixture of Gaussian distribution

Consider the returns of the Swiss Market Index (SMI). Assume that these returns are independently drawn from a mixture of Gaussian distributions. The p.d.f. $f(x;\theta)$, with $\theta = [\mu_1, \sigma_1, \mu_2, \sigma_2, p]'$, is given by:

$$p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

(See p.d.f. of mixtures of Gaussian distributions.)

```
library(AEC);data(smi)
T <- dim(smi)[1]
h <- 5 # holding period (one week)
smi$r <- c(rep(NaN,h),
           100*c(log(smi$Close[(1+h):T]/smi$Close[1:(T-h)])))
indic.dates <- seq(1,T,by=5)  # weekly returns
smi <- smi[indic.dates,]
smi <- smi[complete.cases(smi),]
par(mfrow=c(1,1));par(plt=c(.15,.95,.1,.95))
plot(smi$Date,smi$r,type="l",xlab="",ylab="in percent")
abline(h=0,col="blue")
abline(h=mean(smi$r,na.rm = TRUE)+2*sd(smi$r,na.rm = TRUE),lty=3,col="blue")
abline(h=mean(smi$r,na.rm = TRUE)-2*sd(smi$r,na.rm = TRUE),lty=3,col="blue")
```

Build the log-likelihood function (fucntion `log.f`), and use the numerical BFGS algorithm to maximize it (using the `optim` wrapper):

```
f <- function(theta,y){ # Likelihood function
  mu.1 <- theta[1]; mu.2 <- theta[2]
  sigma.1 <- theta[3]; sigma.2 <- theta[4]
```

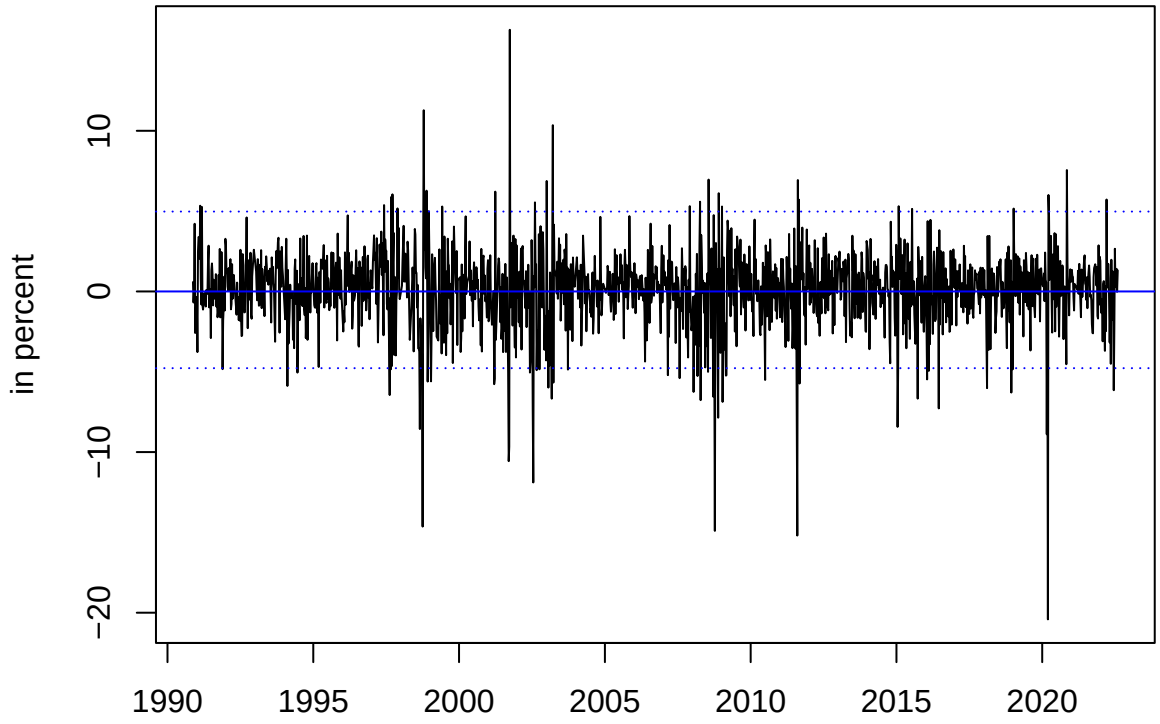


Figure 6.4: Time series of SMI weekly returns (source: Yahoo Finance).

```
  p <- exp(theta[5])/(1+exp(theta[5]))
  res <- p*1/sqrt(2*pi*sigma.1^2)*exp(-(y-mu.1)^2/(2*sigma.1^2)) +
    (1-p)*1/sqrt(2*pi*sigma.2^2)*exp(-(y-mu.2)^2/(2*sigma.2^2))
  return(res)
}
log.f <- function(theta,y){ #log-Likelihood function
  return(-sum(log(f(theta,y))))
}
res.optim <- optim(c(0,0,0.5,1.5,.5),
                   log.f,
                   y=smi$r,
                   method="BFGS", # could be "Nelder-Mead"
                   control=list(trace=FALSE,maxit=100),hessian=TRUE)
theta <- res.optim$par
theta
```

```
## [1]  0.3012379 -1.3167476  1.7715072  4.8197596  1.9454889
```

Next, compute estimates of the covariance matrix of the MLE (using Eqs. (6.11), (6.12), and (6.13)), and compare the three sets of resulting standard deviations for the five estimated paramters:

```
# Hessian approach:
I.1 <- solve(res.optim$hessian)
# Outer-product of gradient approach:
log.f.0 <- log(f(theta,smi$r))
epsilon <- .00000001
d.log.f <- NULL
for(i in 1:length(theta)){
  theta.i <- theta
  theta.i[i] <- theta.i[i] + epsilon
  log.f.i <- log(f(theta.i,smi$r))
  d.log.f <- cbind(d.log.f,
                   (log.f.i - log.f.0)/epsilon)
}
I.2 <- solve(t(d.log.f) %*% d.log.f)
# Misspecification-robust approach (sandwich formula):
I.3 <- I.1 %*% solve(I.2) %*% I.1
cbind(diag(I.1),diag(I.2),diag(I.3))
```

```
##              [,1]        [,2]       [,3]
## [1,] 0.003683422 0.003199481 0.00586160
## [2,] 0.226892824 0.194283391 0.38653389
## [3,] 0.005764271 0.002769579 0.01712255
## [4,] 0.194081311 0.047466419 0.83130838
## [5,] 0.092114437 0.040366005 0.31347858
```

According to the first (respectively third) type of estimate for the covariance matrix, a 95% confidence interval for μ_1 is [0.182, 0.42] (resp. [0.151, 0.451]).

Note that we have not directly estimated parameter p but $\nu = \log(p/(1-p))$ (in such a way that $p = \exp(\nu)/(1+\exp(\nu))$). In order to get an estimate of the standard deviation of our esitmate of p, we can implement the **Delta method**. This method is based on the fact that, for a function g that is continuous in the neighborhood of θ_0 and for large n, we have:

$$\mathbb{V}ar(g(\hat{\theta}_n)) \approx \frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathbb{V}ar(\hat{\theta}_n) \frac{\partial g(\hat{\theta}_n)'}{\partial \theta}. \tag{6.15}$$

```
g <- function(theta){
  mu.1 <- theta[1]; mu.2 <- theta[2]
  sigma.1 <- theta[3]; sigma.2 <- theta[4]
  p <- exp(theta[5])/(1+exp(theta[5]))
  return(c(mu.1,mu.2,sigma.1,sigma.2,p))
}
# Computation of g's gradient around estimated theta:
eps <- .00001
```

```
g.theta <- g(theta)
g.gradient <- NULL
for(i in 1:5){
  theta.perturb <- theta
  theta.perturb[i] <- theta[i] + eps
  g.gradient <- cbind(g.gradient,(g(theta.perturb)-g.theta)/eps)
}
Var <- g.gradient %*% I.3 %*% t(g.gradient)
stdv.g.theta <- sqrt(diag(Var))
stdv.theta <- sqrt(diag(I.3))
cbind(theta,stdv.theta,g.theta,stdv.g.theta)
```

```
##           theta stdv.theta    g.theta stdv.g.theta
## [1,]  0.3012379 0.07656108  0.3012379   0.07656108
## [2,] -1.3167476 0.62171850 -1.3167476   0.62171850
## [3,]  1.7715072 0.13085316  1.7715072   0.13085316
## [4,]  4.8197596 0.91176114  4.8197596   0.91176114
## [5,]  1.9454889 0.55989158  0.8749539   0.06125726
```

The previous results show that the MLE estimate of p is 0.8749539, and its standard deviation is approximately equal to 0.0612573.

To finish with, let us draw the estimated parametric p.d.f. (the mixture of Gaussian distribution), and compare it to a non-parametric (kernel-based) estimate of this p.d.f. (using function `density`):

```
x <- seq(-5,5,by=.01)
par(plt=c(.1,.95,.1,.95))
plot(x,f(theta,x),type="l",lwd=2,xlab="returns, in percent",ylab="",
     ylim=c(0,1.4*max(f(theta,x))))
lines(density(smi$r),type="l",lwd=2,lty=3)
lines(x,dnorm(x,mean=mean(smi$r),sd = sd(smi$r)),col="red",lty=2,lwd=2)
rug(smi$r,col="blue")
legend("topleft",
       c("Kernel estimate (non-parametric)",
         "Estimated mixture of Gaussian distr. (MLE, parametric)",
         "Normal distribution"),
       lty=c(3,1,2),lwd=c(2), # line width
       col=c("black","black","red"),pt.bg=c(1),pt.cex = c(1),
       bg="white",seg.len = 4)
```

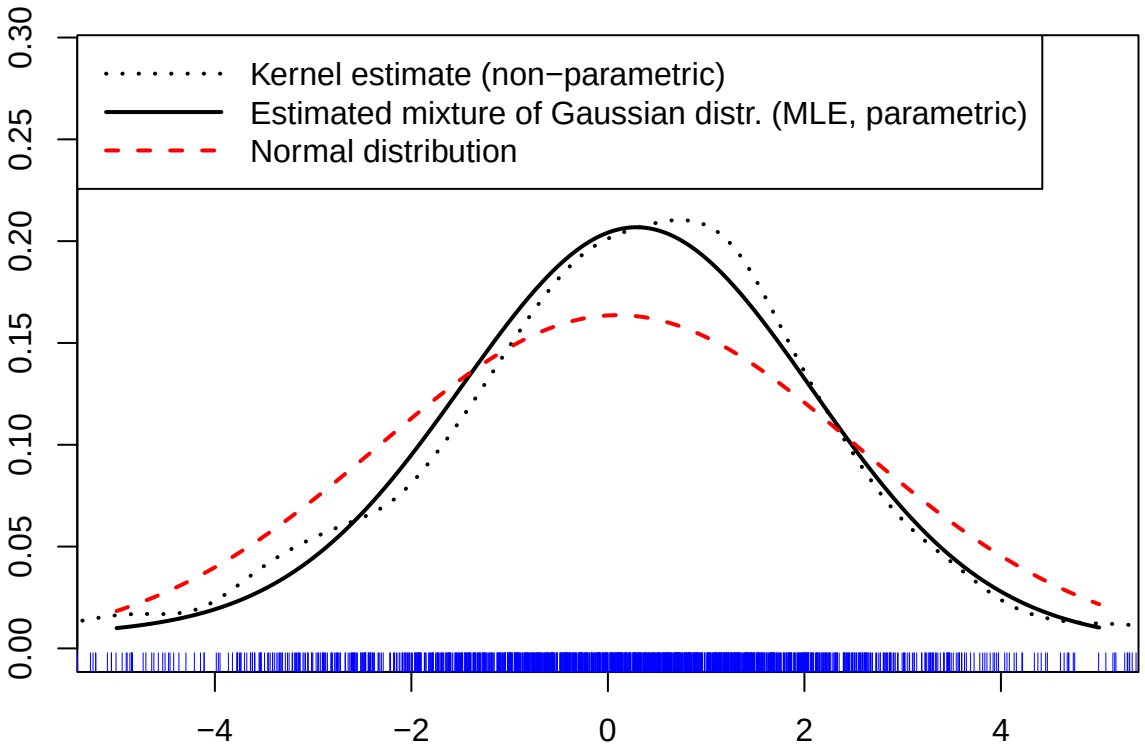


Figure 6.5: Comparison of different estimates of the distribution of returns.

6.2.5 Test procedures

Suppose we want to test the following parameter restrictions:

$$\boxed{H_0 : \underbrace{h(\theta)}_{r\times 1} = 0.} \tag{6.16}$$

In the context of MLE, three tests are largely used:

- Likelihood Ratio (LR) test,
- Wald (W) test,
- Lagrange Multiplier (LM) test.

Here is the rationale behind these three tests:[2]

- LR: If $h(\theta) = 0$, then imposing this restriction during the estimation (restricted estimator) should not result in a large decrease in the likelihood function (w.r.t the unrestricted estimation).
- Wald: If $h(\theta) = 0$, then $h(\widehat{\theta})$ should not be far from 0 (even if these restrictions are not imposed during the MLE).

[2] An interesting graphical presentation of the tests is proposed in Buse (1982).

- LM: If $h(\theta) = 0$, then the gradient of the likelihood function should be small when evaluated at the restricted estimator.

In terms of implementation, while the LR necessitates to estimate both restricted and unrestricted models, the Wald test requires the estimation of the unrestricted model only, and the LM tests requires the estimation of the restricted model only.

As shown below, the three test statistics associated with these three tests coincide asymptotically. (Therefore, they naturally have the same asymptotic distribution, that are χ^2.)

Proposition 6.5 (Asymptotic distribution of the Wald statistic)**.** *Under regularity conditions (Assumptions 6.1) and under $H_0 : h(\theta) = 0$, the Wald statistic, defined by:*

$$\boxed{\xi^W = h(\hat{\theta})' \mathbb{V}ar[h(\hat{\theta})]^{-1} h(\hat{\theta}),}$$

where

$$\mathbb{V}ar[h(\hat{\theta})] = \left(\frac{\partial h(\hat{\theta})}{\partial \theta'} \right) \mathbb{V}ar[\hat{\theta}] \left(\frac{\partial h(\hat{\theta})'}{\partial \theta} \right), \tag{6.17}$$

is asymptotically $\chi^2(r)$, where the number of degrees of freedom r corresponds to the dimension of $h(\theta)$. (Note that Eq. (6.17) *is the same as the one used in the Delta method, see Eq.* (6.15)*.)*

The Wald test, defined by the critical region

$$\{\xi^W \geq \chi^2_{1-\alpha}(r)\},$$

where $\chi^2_{1-\alpha}(r)$ denotes the quantile of level $1-\alpha$ of the $\chi^2(r)$ distribution, has asymptotic level α and is consistent.[3]

Proof. See Appendix 9.5. □

In practice, in Eq. (6.17), $\mathbb{V}ar[\hat{\theta}]$ is replaced by an estimate given, e.g., by Eq. (6.11), Eq. (6.12), or Eq. (6.13).

Proposition 6.6 (Asymptotic distribution of the LM test statistic)**.** *Under regularity conditions (Assumptions 6.1) and under $H_0 : h(\theta) = 0$, the LM statistic*

$$\boxed{\xi^{LM} = \left(\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta'} \bigg|_{\theta=\hat{\theta}^0} \right) [\mathbf{I}(\hat{\theta}^0)]^{-1} \left(\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} \bigg|_{\theta=\hat{\theta}^0} \right),} \tag{6.18}$$

(where $\hat{\theta}^0$ is the restricted MLE estimator) is $\chi^2(r)$.

The test defined by the critical region:

$$\{\xi^{LM} \geq \chi^2_{1-\alpha}(r)\}$$

has asymptotic level α and is consistent (see Defs. 9.7 and 9.8). This test is called Score *or* Lagrange Multiplier (LM) *test.*

[3] See Defs. 9.7 and 9.8 for definitions of the asymptotic levels and consistency of tests.

Proof. See Appendix 9.5. □

Definition 6.8 (Likelihood Ratio test statistics)**.** The likelihood ratio associated to a restriction of the form $H_0 : h(\theta) = 0$ is given by:

$$LR = \frac{\mathcal{L}_R(\theta; \mathbf{y})}{\mathcal{L}_U(\theta; \mathbf{y})} \quad (\in [0, 1]),$$

where $\mathcal{L}_R$ (respectively $\mathcal{L}_U$) is the likelihood function that imposes (resp. that does not impose) the restriction. The likelihood ratio test statistic is given by $-2\log(LR)$, that is:

$$\boxed{\xi^{LR} = 2(\log \mathcal{L}_U(\theta; \mathbf{y}) - \log \mathcal{L}_R(\theta; \mathbf{y})).}$$

Proposition 6.7 (Asymptotic equivalence of LR, LM, and Wald tests)**.** *Under the null hypothesis H_0, we have, asymptotically:*

$$\xi^{LM} = \xi^{LR} = \xi^{W}.$$

Proof. See Appendix 9.5. □

6.3 Bayesian approach

6.3.1 Introduction

An excellent introduction to Bayesian methods is proposed by Martin Haugh, 2017.

As suggested by the name of this approach, the starting point is the Bayes formula:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\&B)}{\mathbb{P}(B)},$$

where A and B are two "events". For instance, A may be: parameter α (conceived as something stochastic) lies in interval $[a, b]$. Assume that you are interested in the probability of occurrence of A. Without any specific information (or "unconditionally"), this probability if $\mathbb{P}(A)$. Your evaluation of this probability can only be better if you are provided with any additional form of information. Typically, if the event B tends to occur simultaneously with A, then knowledge of B can be useful. The Bayes formula says how this additional information (on B) can be used to "update" the probability of event A.

In our case, this intuition will work as follows: assume that you know the form of the data-generating process (DGP). That is, you know the structure of the model used to draw some stochastic data; you also know the type of distributions used to generate these data. However, you do not know the numerical values of all the parameters characterizing the DGP. Let us denote by θ the vector of unknown parameters. While these parameters are not known exactly, assume that we have –even without having observed any data– some **priors** on their distribution. Then, as was the case in the example above (with A and B), the observation

of data generated by the model can only reduce the uncertainty associated with θ. Loosely speaking, combining the priors and the observations of data generated by the model should result in "thinner" distributions for the components of θ. The latter distributions are called the **posterior distributions**.[4]

Let us formalize this intuition. Define the prior by $f_\theta(\theta)$ and the model realizations (the "data") by vector $\mathbf{y}$. The joint distribution of $(\mathbf{y}, \theta)$ is given by:

$$f_{Y,\theta}(\mathbf{y}, \theta) = f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta),$$

and, symmetrically, by

$$f_{Y,\theta}(\mathbf{y}, \theta) = f_{\theta|Y}(\theta, \mathbf{y}) f_Y(\mathbf{y}),$$

where $f_{\theta|Y}(\cdot, \mathbf{y})$, the distribution of the parameters conditional on the observations, is the **posterior** distribution.

The last two equations imply that:

$$f_{\theta|Y}(\theta, \mathbf{y}) = \frac{f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta)}{f_Y(\mathbf{y})}. \tag{6.19}$$

Note that f_Y is the marginal (or unconditional) distribution of $\mathbf{y}$, that can be written:

$$f_Y(\mathbf{y}) = \int f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta) d\theta. \tag{6.20}$$

Eq. (6.19) is sometimes rewritten as follows:

$$f_{\theta|Y}(\theta, \mathbf{y}) \propto f_{\theta,Y}(\theta, \mathbf{y}) := f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta), \tag{6.21}$$

where $\propto$ means, loosely speaking, "*proportional to*". In rare instances, starting from given priors, one can analytically compute the posterior distribution $f_\theta(\theta, \mathbf{y})$. However, in most cases, this is out of reach. One then has to resort to numerical approaches to compute the posterior distribution. Monte Carlo Markov Chains (MCMC) is one of them.

According to the Bernstein-von Mises Theorem, Bayesian and MLE estimators have the same large sample properties. (In particular, the Bayesian approach also achieve the FDCR bound, see Theorem 6.1.) The intuition behind this result is that the influence of the prior diminishes with increasing sample sizes.

6.3.2 Monte-Carlo Markov Chains

MCMC techniques aim at using simulations to approach a distribution whose distribution is difficult to obtain analytically. Indeed, in some circumstances, one can draw in a distribution even if we do not know its analytical expression.

[4]The output of the Bayesian approach will be the (posterior) distribution of the vector of parameters (θ). When we speak about the *distributions of the components of* θ, we mean the marginal distributions of each component of the vector.

Definition 6.9 (Markov Chain)**.** The sequence $\{z_i\}$ is said to be a (first-order) Markovian process is it satisfies:

$$f(z_i|z_{i-1}, z_{i-2}, \dots) = f(z_i|z_{i-1}).$$

The Metropolis-Hastings (MH) algorithm is a specific MCMC approach that allows to generate samples of θ's whose distribution approximately corresponds to the posterior distribution of Eq. (6.19).

The MH algorithm is a recursive algorithm. That is, one can draw the i^{th} value of θ, denoted by θ_i, if one has already drawn θ_{i-1}. Assume we have θ_{i-1}. We obtain a value for θ_i by implementing the following steps:

1. Draw $\tilde{\theta}_i$ from the conditional distribution $Q_{\tilde{\theta}|\theta}(\cdot, \theta_{i-1})$, called **proposal distribution**.
2. Draw u in a uniform distribution on $[0, 1]$.
3. Compute
$$\alpha(\tilde{\theta}_i, \theta_{i-1}) := \min\left(\frac{f_{\theta,Y}(\tilde{\theta}_i, \mathbf{y})}{f_{\theta,Y}(\theta_{i-1}, \mathbf{y})} \times \frac{Q_{\tilde{\theta}|\theta}(\theta_{i-1}, \tilde{\theta}_i)}{Q_{\tilde{\theta}|\theta}(\tilde{\theta}_i, \theta_{i-1})}, 1\right), \tag{6.22}$$
where $f_{\theta,Y}$ is given in Eq. (6.21).
4. If $u < \alpha(\tilde{\theta}_i, \theta_{i-1})$, then take $\theta_i = \tilde{\theta}_i$, otherwise we leave θ_i equal to θ_{i-1}.

It can be shown that, the distribution of the draws converges to the posterior distribution. That is, after a sufficiently large number of iterations, the draws can be considered to be drawn from the posterior distribution.[5]

To get some insights into the algorithm, consider the case of a **symmetric proposal distribution**, that is:

$$Q_{\tilde{\theta}|\theta}(\tilde{\theta}_i, \theta_{i-1}) = Q_{\tilde{\theta}|\theta}(\theta_{i-1}, \tilde{\theta}_i). \tag{6.23}$$

We then have:

$$\alpha(\tilde{\theta}, \theta_{i-1}) = \min\left(\frac{q(\tilde{\theta}, y)}{q(\theta_{i-1}, y)}, 1\right). \tag{6.24}$$

Remember that, up to the marginal distribution of the data ($f_Y(\mathbf{y})$), $f_{\theta,Y}(\tilde{\theta}, \mathbf{y})$ is the probability of observing $\mathbf{y}$ conditional on having a model parameterized by $\tilde{\theta}$. Then, under Eq. (6.24), it appears that if this probability is larger for $\tilde{\theta}$ than for θ_{i-1} (in which case $\tilde{\theta}$ seems "more consistent with the observations $\mathbf{y}$" than θ_{i-1}), we accept θ_i. By contrast, if $f_{\theta,Y}(\tilde{\theta}, \mathbf{y}) < f_{\theta,Y}(\theta_{i-1}, \mathbf{y})$, then we do not necessarily accept the proposed value $\tilde{\theta}$, especially if $f_{\theta,Y}(\tilde{\theta}, \mathbf{y}) \ll f_{\theta,Y}(\theta_{i-1}, \mathbf{y})$ (in which case $\tilde{\theta}$ seems far less consistent with the observations $\mathbf{y}$ than θ_{i-1}, and, accordingly, the acceptance probability, namely $\alpha(\tilde{\theta}, \theta_{i-1})$, is small).

The choice of the **proposal distribution** $Q_{\tilde{\theta}|\theta}$ is crucial to get a rapid convergence of the algorithm. Looking at Eq. (6.22), it is easily seen that the optimal choice would be

[5]The proof of this claim is based on the fact that, if θ_{i-1} is drawn from the posterior distribution, then it is also the case for θ_i.

$Q_{\tilde{\theta}|\theta}(\cdot, \theta_i) = f_{\theta|Y}(\cdot, \mathbf{y})$. In that case, we would have $\alpha(\tilde{\theta}_i, \theta_{i-1}) \equiv 1$ (see Eq. (6.22)). We would then accept all draws from the proposal distribution, as this distribution would directly be the posterior distribution. Of course, this situation is not realistoc as the objective of the algorithm is precisely to approximate the posterior distribution.

A common choice for Q is a multivariate normal distribution. If θ is of dimension K, we can for instance use:

$$Q(\tilde{\theta}, \theta) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^K} \exp\left(-\frac{1}{2}\sum_{j=1}^{K}\frac{(\tilde{\theta}_j - \theta_j)^2}{\sigma^2}\right),$$

which is an example of symmetric proposal distribution (see Eq. (6.23)). Equivalently, we then have:

$$\tilde{\theta} = \theta + \varepsilon,$$

where ε is a K-dimensional vector of independent zero-mean normal disturbances of variance σ^2.[6] One then has to determine an appropriate value for σ. If it is too low, then α will be close to 1 (as $\tilde{\theta}_i$ will be close to θ_{i-1}), and we will accept very often the proposed value $(\tilde{\theta}_i)$. This seems to be a favourable situation. But it may not be. Indeed, it means that it will take a large number of iterations to explore the whole distribution of θ. What if σ is very large? In this case, it is likely that the porposed values $(\tilde{\theta}_i)$ will often result in poor likelihoods; The probability of acceptance will then be low and the Markov chain may be blocked at its initial value. Therefore, intermediate values of σ^2 have to be determined. The acceptance rate (i.e., the average value of $\alpha(\tilde{\theta}, \theta_{i-1})$) can be used as a guide for that. Indeed, a literature explores the optimal values for such acceptance rate (in order to obtain the best possible fit of the posterior for a minimum number of algorithm iterations). In particular, following Roberts et al. (1997), people often target acceptance rate of the order of magnitude of 20%.

It is important to note that, to implement this approach, one only has to be able to compute the joint p.d.f. $q(\theta, \mathbf{y}) = f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta)$ (Eq. (6.21)). That is, as soon as one can evaluate the likelihood $(f_{Y|\theta}(\mathbf{y}, \theta))$ and the prior $(f_\theta(\theta))$, we can employ this methodology.

6.3.3 Example: AR(1) specification

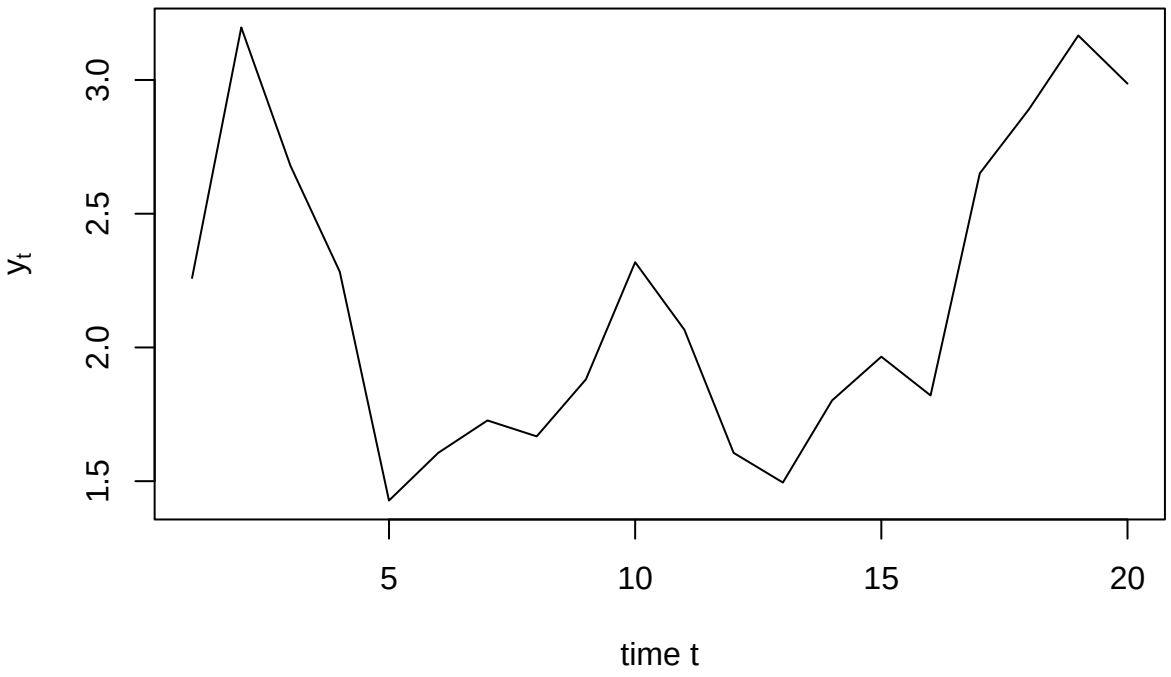
In the following example, we employ MCMC in order to estimate the posterior distributions of the three parameters defining an AR(1) model (see Section 8.2.2). The specification is as follows:

$$y_t = \mu + \rho y_{t-1} + \sigma\varepsilon_t, \quad \varepsilon_t \sim i.i.d.\, \mathcal{N}(0, 1).$$

Hence, we have $\theta = [\mu, \rho, \sigma]$. Let us first simulate the process on T periods:

[6] We could also have different variances for the different components of θ. However, this may lead to complicated settings. A useful practice consists in looking for model (re)parametrization –based, e.g., on exponential and/or logistic functions– that are such that the components of θ are all expected to be of the order of magnitude of the unity.

```
mu <- .6; rho <- .8; sigma <- .5 # true model specification
T <- 20 # number of observations
y0 <- mu/(1-rho)
Y <- NULL
for(t in 1:T){
  if(t==1){y <- y0}
  y <- mu + rho*y + sigma * rnorm(1)
  Y <- c(Y,y)}
plot(Y,type="l",xlab="time t",ylab=expression(y[t]))
```



Next, let us write the likelihood function, i.e. $f_{Y|\theta}(\mathbf{y}, \theta)$. For ρ, which is expected to be between 0 and 1, we use a logistic transformation. For σ, that is expected to be positive, we use an exponential transformation.

```
likelihood <- function(param,Y){
  mu  <- param[1]
  rho <- exp(param[2])/(1+exp(param[2]))
  sigma <- exp(param[3])
  MU <- mu/(1-rho)
  SIGMA2 <- sigma^2/(1-rho^2)
  L <- 1/sqrt(2*pi*SIGMA2)*exp(-(Y[1]-MU)^2/(2*SIGMA2))
  Y1 <- Y[2:length(Y)]
  Y0 <- Y[1:(length(Y)-1)]
  aux <- 1/sqrt(2*pi*sigma^2)*exp(-(Y1-mu-rho*Y0)^2/(2*sigma^2))
  L <- L * prod(aux)
  return(L)
}
```

Next define function `rQ` that draws from the (Gaussian) proposal distribution, as well as function `Q`, that computes $Q_{\tilde{\theta}|\theta}(\tilde{\theta}, \theta)$:

```
rQ <- function(x,a){
  n <- length(x)
  y <- x + a * rnorm(n)
  return(y)}
Q <- function(y,x,a){
  q <- 1/sqrt(2*pi*a^2)*exp(-(y - x)^2/(2*a^2))
  return(prod(q))}
```

We consider Gaussian priors:

```
prior <- function(param,means_prior,stdv_prior){
  f <- 1/sqrt(2*pi*stdv_prior^2)*exp(-(param -
                                       means_prior)^2/(2*stdv_prior^2))
  return(prod(f))}
```

Function `p_tilde` corresponds to $f_{\theta,Y}$:

```
p_tilde <- function(param,Y,means_prior,stdv_prior){
  p <- likelihood(param,Y) * prior(param,means_prior,stdv_prior)
  return(p)}
```

We can now define function α (Eq. (6.22)):

```
alpha <- function(y,x,means_prior,stdv_prior,a){
  aux <- p_tilde(y,Y,means_prior,stdv_prior)/
    p_tilde(x,Y,means_prior,stdv_prior) * Q(y,x,a)/Q(x,y,a)
  alpha_proba <- min(aux,1)
  return(alpha_proba)}
```

Now, all is set for us to write the MCMC function:

```
MCMC <- function(Y,means_prior,stdv_prior,a,N){
  x <- means_prior
  all_theta <- NULL
  count_accept <- 0
  for(i in 1:N){
    y <- rQ(x,a)
    alph <- alpha(y,x,means_prior,stdv_prior,a)
    #print(alph)
    u <- runif(1)
```

```
    if(u < alph){
      count_accept <- count_accept + 1
      x <- y}
    all_theta <- rbind(all_theta,x)}
  print(paste("Acceptance rate:",toString(round(count_accept/N,3))))
  return(all_theta)}
```

Specify the Gaussian priors:

```
true_values <- c(mu,log(rho/(1-rho)),log(sigma))
means_prior <- c(1,0,0) # as if we did not know the true values
stdv_prior <- rep(2,3)
resultMCMC <- MCMC(Y,means_prior,stdv_prior,a=.45,N=20000)
```

```
## [1] "Acceptance rate: 0.098"
```

```
par(mfrow=c(2,3))
for(i in 1:length(means_prior)){
  m <- means_prior[i]
  s <- stdv_prior[i]
  x <- seq(m-3*s,m+3*s,length.out = 100)
  par(mfg=c(1,i))
  aux <- density(resultMCMC[,i])
  par(plt=c(.15,.95,.15,.85))
  plot(x,dnorm(x,m,s),type="l",xlab="",ylab="",main=paste("Parameter",i),
       ylim=c(0,max(aux$y)))
  lines(aux$x,aux$y,col="red",lwd=2)
  abline(v=true_values[i],lty=2,col="blue")
  par(mfg=c(2,i))
  plot(resultMCMC[,i],1:length(resultMCMC[,i]),xlim=c(min(x),max(x)),
       type="l",xlab="",ylab="")}
```

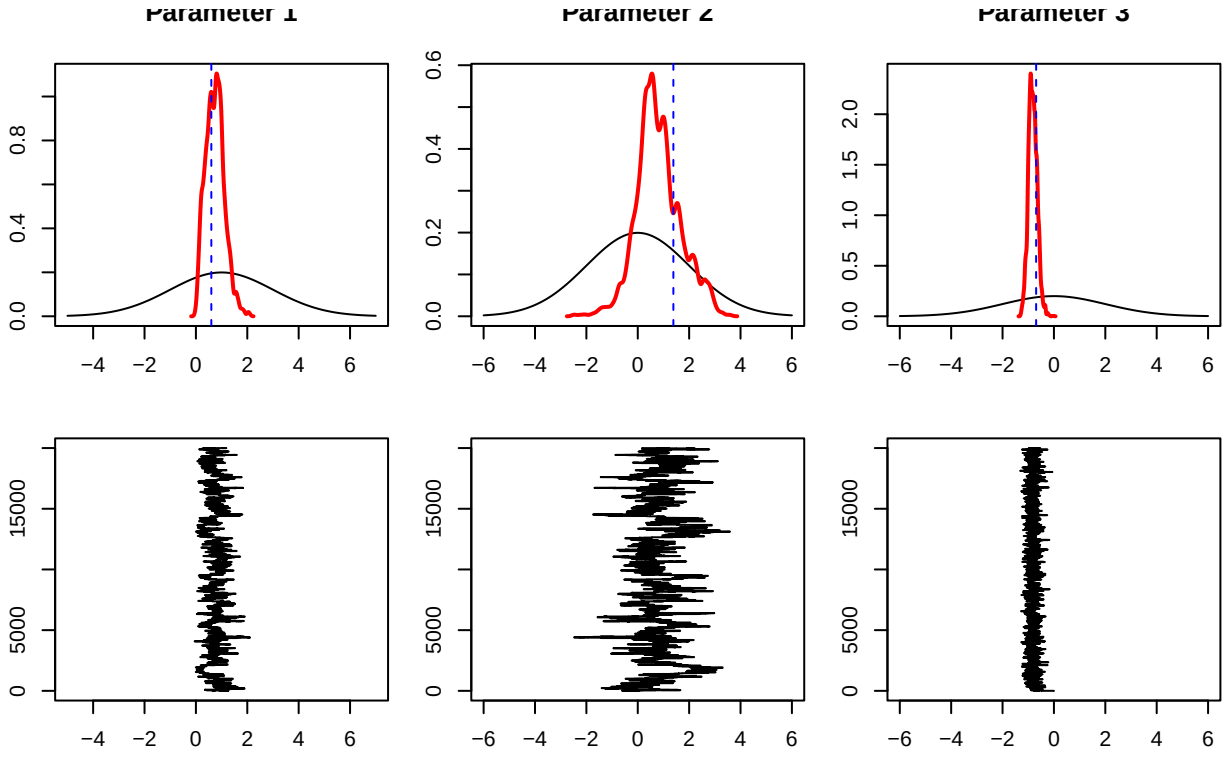


Figure 6.6: The upper line of plot compares prior (black) and posterior (red) distributions. The vertical dashed blue lines indicate the true values of the parameters. The second row of plots show the sequence of θ_i's generated by the MCMC algorithm. These sequences are the ones used to produce the posterior distributions (red lines) in the upper plots.

Chapter 7

Binary-choice models

In microeconometric models, the variables of interest often feature restricted distributions —for instance with discontinuous support—, which necessitates specific models. Typically, in many instances, the variables to be explained (the y_i's) have only two possible values (0 and 1, say). That is, they are binary variables. The probability for them to be equal to either 0 or 1 may depend on independent variables, gathered in vectors $\mathbf{x}_i$ ($K \times 1$).

The spectrum of applications is wide:

- Binary decisions (e.g. in referendums, being owner or renter, living in the city or in the countryside, in/out of the labour force,...),
- Contamination (disease or default),
- Success/failure (exams).

Without loss of generality, the model reads:

$$y_i|\mathbf{X} \sim \mathcal{B}(g(\mathbf{x}_i;\theta)), \tag{7.1}$$

where $g(\mathbf{x}_i;\theta)$ is the parameter of the Bernoulli distribution. In other words, conditionally on $\mathbf{X}$:

$$y_i = \begin{cases} 1 & \text{with probability } g(\mathbf{x}_i;\theta) \\ 0 & \text{with probability } 1 - g(\mathbf{x}_i;\theta), \end{cases} \tag{7.2}$$

where θ is a vector of parameters to be estimated.

An estimation strategy is to assume that $g(\mathbf{x}_i;\theta)$ can be proxied by $\tilde{\theta}'\mathbf{x}_i$ and to run a linear regression to estimate $\tilde{\theta}$ (a situation called **Linear Probability Model, LPM**):

$$y_i = \tilde{\theta}'\mathbf{x}_i + \varepsilon_i.$$

Notwithstanding the fact that this specification does not exclude negative probabilities or probabilities greater than one, it could be compatible with the *assumption of zero conditional mean* (Hypothesis 4.2) and with the *assumption of non-correlated residuals* (Hypothesis 4.4), but more difficultly with the *homoskedasticity assumption* (Hypothesis 4.3). Moreover, the

ε_i's cannot be Gaussian (because $y_i \in \{0, 1\}$). Hence, using a linear regression to study the relationship between $\mathbf{x}_i$ and y_i can be consistent but it is inefficient.

Figure 7.1 illustrates the fit resulting from an application of the LPM model to binary (dependent) variables.

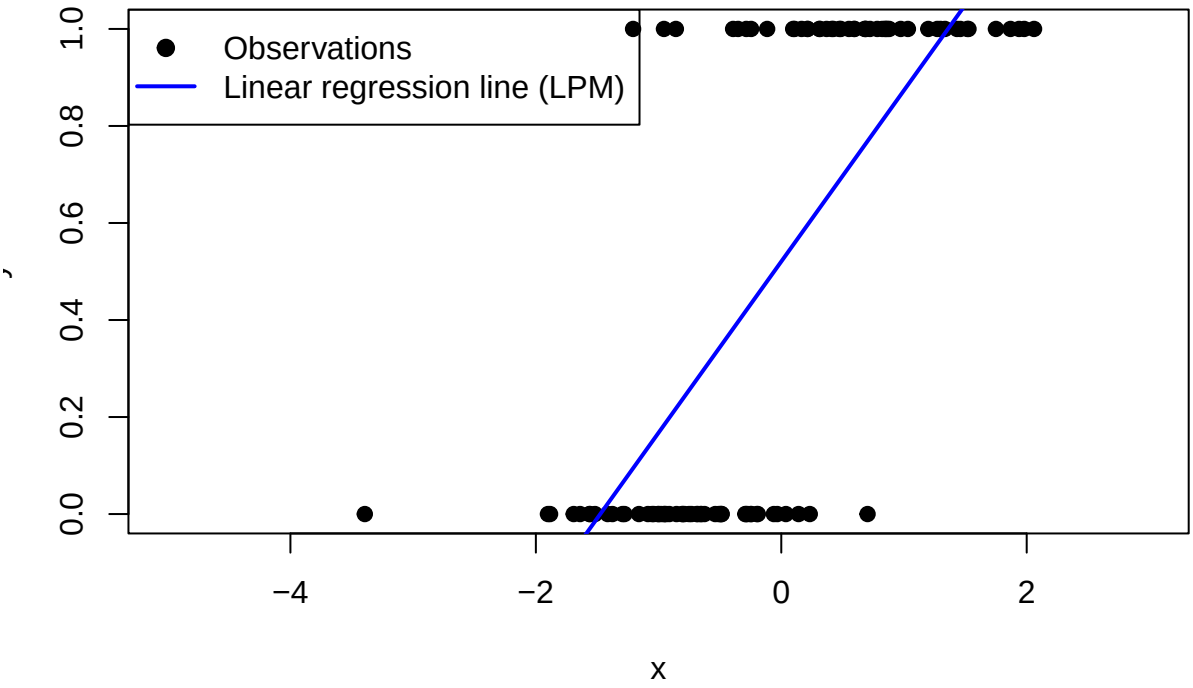


Figure 7.1: Fitting a binary variable with a linear model (Linear Probability Model, LPM). The model is $\mathbb{P}(y_i = 1|x_i) = \Phi(0.5 + 2x_i)$, where Φ is the c.d.f. of the normal distribution and where $x_i \sim i.i.d.\, \mathcal{N}(0, 1)$.

Except for its last row (LPM case), Table 7.1 provides examples of functions g valued in $[0, 1]$, and that can therefore used in models of the type: $\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = g(\theta'\mathbf{x}_i)$ (see Eq. (7.2)). The "linear" case is given for comparison, but note that it does not satisfy $g(\theta'\mathbf{x}_i) \in [0, 1]$ for any value of $\theta'\mathbf{x}_i$.

Table 7.1: This table provides examples of function g, s.t. $\mathbb{P}(y_i = 1|\mathbf{x}_i; heta) = g(\theta'\mathbf{x}_i)$. The LPM case (last row) is given for comparison but, again, it does not satisfy $g(\theta'\mathbf{x}_i) \in [0, 1]$ for any value of $\theta'\mathbf{x}_i$.

Model	Function g	Derivative
Probit	Φ	ϕ
Logit	$\dfrac{\exp(x)}{1+\exp(x)}$	$\dfrac{\exp(x)}{(1+\exp(x))^2}$
log-log	$1 - \exp(-\exp(x))$	$\exp(-\exp(x))\exp(x)$
linear (LPM)	x	1

Figure 7.2 displays the first three g functions appearing in Table 7.1.

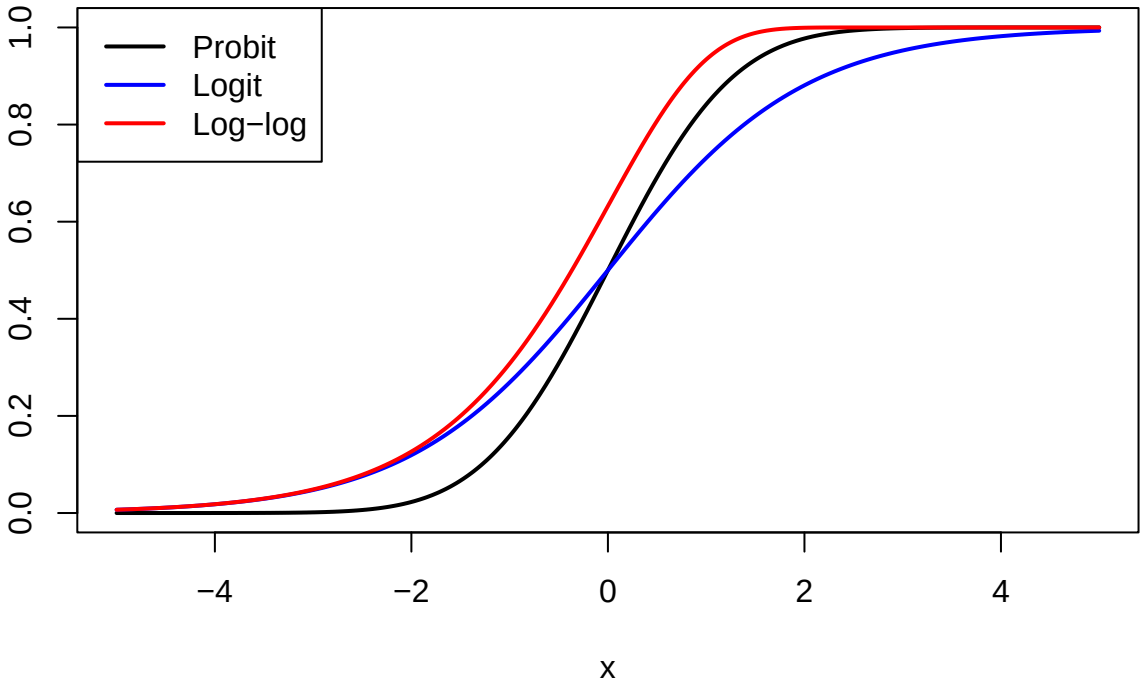


Figure 7.2: Probit, Logit, and Log-log functions.

The **probit** and the **logit** models are popular binary-choice models. In the probit model, we have:

$$g(z) = \Phi(z), \tag{7.3}$$

where Φ is the c.d.f. of the normal distribution. And for the logit model:

$$g(z) = \frac{1}{1 + \exp(-z)}. \tag{7.4}$$

Figure 7.3 shows the conditional probabilities associated with the (probit) model that had been used to generate the data of Figure 7.1.

7.1 Interpretation in terms of latent variable, and utility-based models

The probit model has an interpretation in terms of latent variables, which, in turn, is often exploited in structural models, called **Random Utility Models (RUM)**. In such structural models, it is assumed that the agents that have to take the decision do so by selecting the outcome that provides them with the larger utility (for agent i, two possible outcomes: $y_i = 0$ or $y_i = 1$). Part of this utility is observed by the econometrician —it depends on the covariates $\mathbf{x}_i$— and part of it is latent.

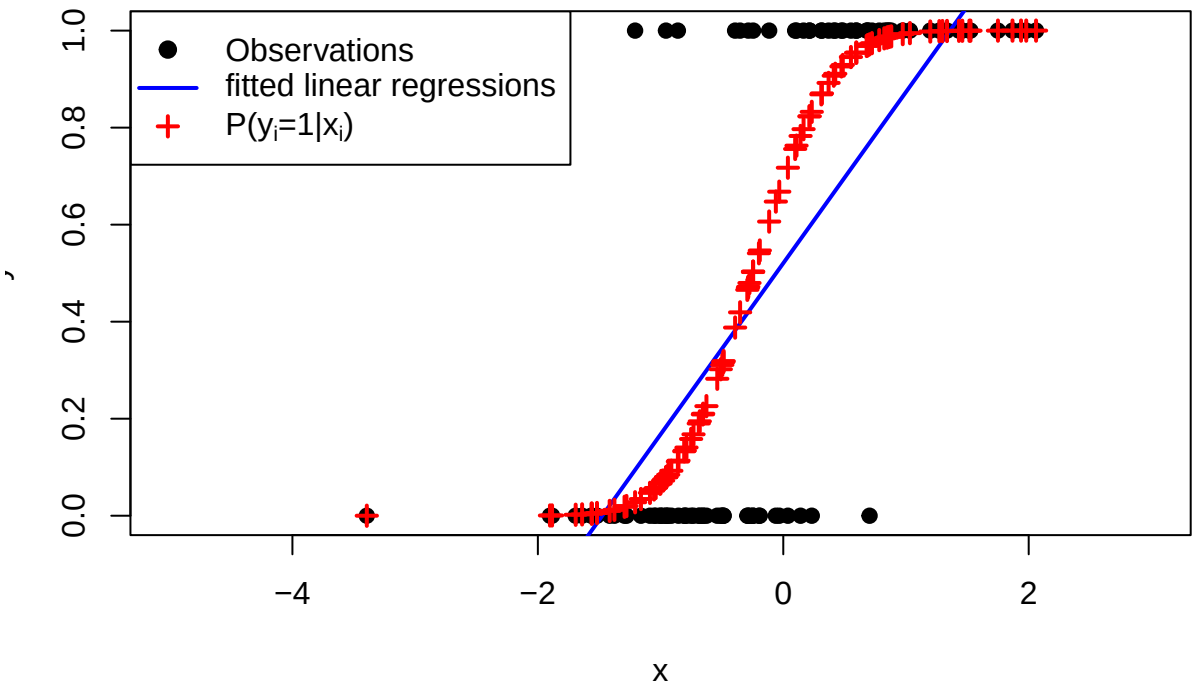


Figure 7.3: The model is $\mathbb{P}(y_i = 1|x_i) = \Phi(0.5 + 2x_i)$, where Φ is the c.d.f. of the normal distribution and where $x_i \sim i.i.d.\, \mathcal{N}(0, 1)$. Crosses give the model-implied probabilities of having $y_i = 1$ (conditional on x_i).

In the probit model, we have:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = \Phi(\theta'\mathbf{x}_i) = \mathbb{P}(-\varepsilon_i < \theta'\mathbf{x}_i),$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$. That is:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = \mathbb{P}(0 < y_i^*),$$

where $y_i^* = \theta'\mathbf{x}_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, 1)$. Variable y_i^* can be interpreted as a (latent) variable that determines y_i (since $y_i = \mathbb{I}_{\{y_i^* > 0\}}$).

Figure 7.4 illustrates this situation.

Assume that agent (i) chooses $y_i = 1$ if the utility associated with this choice $(U_{i,1})$ is higher than the one associated with $y_i = 0$ (that is $U_{i,0}$). Assume further that the utility of agent i, if she chooses outcome j $(\in \{0, 1\})$, is given by

$$U_{i,j} = V_{i,j} + \varepsilon_{i,j},$$

where $V_{i,j}$ is the deterministic component of the utility associated with choice and where $\varepsilon_{i,j}$ is a random (agent-specific) component. Moreover, posit $V_{i,j} = \theta_j'\mathbf{x}_i$. We then have:

$$\begin{aligned} \mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) &= \mathbb{P}(\theta_1'\mathbf{x}_i + \varepsilon_{i,1} > \theta_0'\mathbf{x}_i + \varepsilon_{i,0}) \\ &= F(\theta_1'\mathbf{x}_i - \theta_0'\mathbf{x}_i) = F([\theta_1 - \theta_0]'\mathbf{x}_i), \end{aligned} \tag{7.5}$$

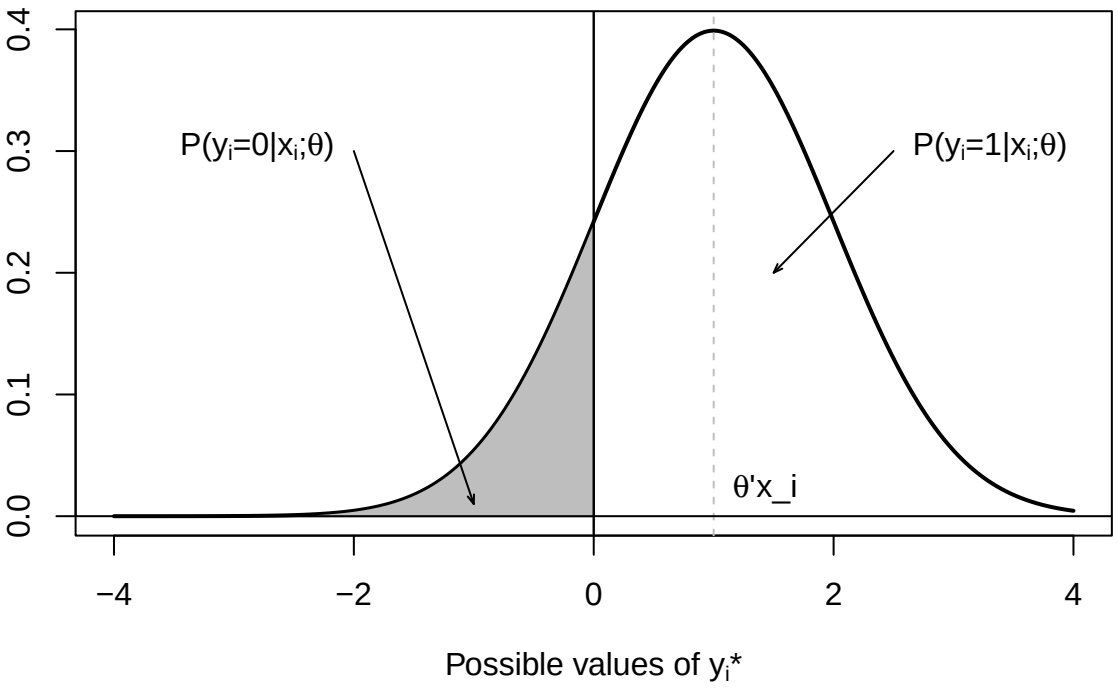


Figure 7.4: Distribution of y_i^* conditional on $\mathbf{x}_i$.

where F is the c.d.f. of $\varepsilon_{i,0} - \varepsilon_{i,1}$.

Note that only the difference $\theta_1 - \theta_0$ is identifiable (as opposed to θ_1 *and* θ_0). Indeed, replacing U with aU ($a > 0$) gives the same model. This *scaling* issue can be solved by fixing the variance of $\varepsilon_{i,0} - \varepsilon_{i,1}$.

Example 7.1 (Migration and income)**.** The RUM approach has been used by Nakosteen and Zimmer (1980) to study migration choices. Their model is based on the comparison of marginal costs and benefits associated with migration. The main ingredients of their approach are as follows:

- Wage that can be earned at the present location: $y_p^* = \theta_p' \mathbf{x}_p + \varepsilon_p$.
- Migration cost: $C^* = \theta_c' \mathbf{x}_c + \varepsilon_c$.
- Wage earned elsewhere: $y_m^* = \theta_m' \mathbf{x}_m + \varepsilon_m$.

In this context, agents decision to migrate if $y_m^* > y_p^* + C^*$, i.e. if

$$y^* = y_m^* - y_p^* - C^* = \theta' \mathbf{x} + \underbrace{\varepsilon}_{=\varepsilon_m - \varepsilon_c - \varepsilon_p} > 0,$$

where $\mathbf{x}$ is the union of the $\mathbf{x}_i$s, for $i \in \{p, m, c\}$.

7.2 Alternative-Varying Regressors

In some cases, regressors may depend on the considered alternative (0 or 1). For instance:

- When modeling the decision to participate in the labour force (or not), the wage depends on the alternative. Typically, it is zero if the considered agent has decided not to work (and strictly positive otherwise).
- In the context of the choice of transportation mode, "time cost" depends on the considered transportation mode.

In terms of utility, we then have:

$$V_{i,j} = \theta_j^{(u)'} \mathbf{u}_{i,j} + \theta_j^{(v)'} \mathbf{v}_i,$$

where the $\mathbf{u}_{i,j}$'s are regressors associated with agent i, but taking different values for the different choices ($j = 0$ or $j = 1$). In that case, Eq. (7.5) becomes:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = F\left(\theta_1^{(u)'} \mathbf{u}_{i,1} - \theta_0^{(u)'} \mathbf{u}_{i,0} + [\theta_1^{(v)} - \theta_0^{(v)}]' \mathbf{v}_i\right), \tag{7.6}$$

and, if $\theta_1^{(u)} = \theta_0^{(u)} = \theta^{(u)}$ —as is customary— we get:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = F\left(\theta_1^{(u)'} (\mathbf{u}_{i,1} - \mathbf{u}_{i,0}) + [\theta_1^{(v)} - \theta_0^{(v)}]' \mathbf{v}_i\right). \tag{7.7}$$

Example 7.2 (Fishing-mode dataset)**.** The fishing-mode dataset used in Cameron and Trivedi (2005) (Chapters 14 and 15) contains alternative-specific variables. Specifically, for each individual, the price and catch rate depend on the fishing model. In the table reported below, lines `price` and `catch` correspond to the prices and catch rates associated with the chosen alternative.

```
library(mlogit)
data("Fishing",package="mlogit")
stargazer::stargazer(Fishing,type="text")
```

```
##
## ===========================================================
## Statistic       N     Mean    St. Dev.    Min      Max
## -----------------------------------------------------------
## price.beach   1,182  103.422   103.641   1.290   843.186
## price.pier    1,182  103.422   103.641   1.290   843.186
## price.boat    1,182  55.257    62.713    2.290   666.110
## price.charter 1,182  84.379    63.545   27.290   691.110
## catch.beach   1,182   0.241     0.191    0.068    0.533
## catch.pier    1,182   0.162     0.160    0.001    0.452
## catch.boat    1,182   0.171     0.210   0.0002    0.737
## catch.charter 1,182   0.629     0.706    0.002    2.310
## income        1,182 4,099.337 2,461.964 416.667 12,500.000
## -----------------------------------------------------------
```

7.3 Estimation

These models can be estimated by Maximum Likelihood approaches (see Section 6.2).

To simplify the exposition, we consider the $\mathbf{x}_i$ vectors of covariates to be deterministic. Moreover, we assume that the r.v. are independent across entities i. How to write the likelihood in that case? It is easily checked that:

$$f(y_i|\mathbf{x}_i;\theta) = g(\theta'\mathbf{x}_i)^{y_i}(1-g(\theta'\mathbf{x}_i))^{1-y_i}.$$

Therefore, if the observations $(\mathbf{x}_i, y_i)$ are independent across entities i, we obtain:

$$\log \mathcal{L}(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} y_i \log[g(\theta'\mathbf{x}_i)] + (1-y_i)\log[1-g(\theta'\mathbf{x}_i)].$$

The likelihood equation reads (FOC of the optimization program, see Def. 6.7):

$$\frac{\partial \log \mathcal{L}(\theta;\mathbf{y},\mathbf{X})}{\partial \theta} = \mathbf{0},$$

that is:

$$\sum_{i=1}^{n} y_i \mathbf{x}_i \frac{g'(\theta'\mathbf{x}_i)}{g(\theta'\mathbf{x}_i)} - (1-y_i)\mathbf{x}_i \frac{g'(\theta'\mathbf{x}_i)}{1-g(\theta'\mathbf{x}_i)} = \mathbf{0}.$$

This is a nonlinear (multivariate) equation that can be solved numerically. Under regularity conditions (Hypotheses 6.1), we approximately have (Prop. 6.4):

$$\theta_{MLE} \sim \mathcal{N}(\theta_0, \mathbf{I}(\theta_0)^{-1}),$$

where

$$\mathbf{I}(\theta_0) = -\mathbb{E}_0\left(\frac{\partial^2 \log \mathcal{L}(\theta;\mathbf{y},\mathbf{X})}{\partial\theta\partial\theta'}\right) = n\mathcal{I}_Y(\theta_0).$$

For finite samples, we can e.g. approximate $\mathbf{I}(\theta_0)^{-1}$ by Eq. (6.11):

$$\mathbf{I}(\theta_0)^{-1} \approx -\left(\frac{\partial^2 \log \mathcal{L}(\theta_{MLE};\mathbf{y},\mathbf{X})}{\partial\theta\partial\theta'}\right)^{-1}.$$

In the Probit case (see Table 7.1), it can be shown that we have:

$$\frac{\partial^2 \log \mathcal{L}(\theta;\mathbf{y},\mathbf{X})}{\partial\theta\partial\theta'} = -\sum_{i=1}^{n} g'(\theta'\mathbf{x}_i)[\mathbf{x}_i\mathbf{x}_i'] \times$$
$$\left[y_i \frac{g'(\theta'\mathbf{x}_i) + \theta'\mathbf{x}_i g(\theta'\mathbf{x}_i)}{g(\theta'\mathbf{x}_i)^2} + (1-y_i)\frac{g'(\theta'\mathbf{x}_i) - \theta'\mathbf{x}_i(1-g(\theta'\mathbf{x}_i))}{(1-g(\theta'\mathbf{x}_i))^2}\right].$$

In the Logit case (see Table 7.1), it can be shown that we have:

$$\frac{\partial^2 \log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X})}{\partial \theta \partial \theta'} = -\sum_{i=1}^{n} g'(\theta' \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i',$$

where $g'(x) = \dfrac{\exp(-x)}{(1+\exp(-x))^2}$.

Remark that, since $g'(x) > 0$, $-\partial^2 \log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X})/\partial \theta \partial \theta'$ is positive definite.

7.4 Marginal effects

How to measure marginal effects, i.e. the effect on the probability that $y_i = 1$ of a marginal increase of $x_{i,k}$? This object is given by:

$$\frac{\partial \mathbb{P}(y_i = 1 | \mathbf{x}_i; \theta)}{\partial x_{i,k}} = \underbrace{g'(\theta' \mathbf{x}_i)}_{>0} \theta_k,$$

which is of the same sign as θ_k if function g is monotonously increasing.

For agent i, this marginal effect is consistently estimated by $g'(\theta'_{MLE} \mathbf{x}_i) \theta_{MLE,k}$. It is important to see that the marginal effect depends on $\mathbf{x}_i$: respective increases by 1 unit of $x_{i,k}$ (entity i) and of $x_{j,k}$ (entity j) do not necessarily have the same effect on $\mathbb{P}(y_i = 1 | \mathbf{x}_i; \theta)$ as on $\mathbb{P}(y_j = 1 | \mathbf{x}_j; \theta)$. To address this issue, one can compute some measures of "average" marginal effect. There are two main solutions. For each explanatory variable k:

i. Denoting by $\hat{\mathbf{x}}$ the sample average of the $\mathbf{x}_i$s, compute $g'(\theta'_{MLE} \hat{\mathbf{x}}) \theta_{MLE,k}$.
ii. Compute the average (across i) of $g'(\theta'_{MLE} \mathbf{x}_i) \theta_{MLE,k}$.

7.5 Goodness of fit

There is no obvious version of "R^2" for binary-choice models. Existing measures are called **pseudo-R^2 measures**.

Denoting by $\log \mathcal{L}_0(\mathbf{y})$ the (maximum) log-likelihood that would be obtained for a model containing only a constant term (i.e. with $\mathbf{x}_i = 1$ for all i), the McFadden's pseudo-R^2 is given by:

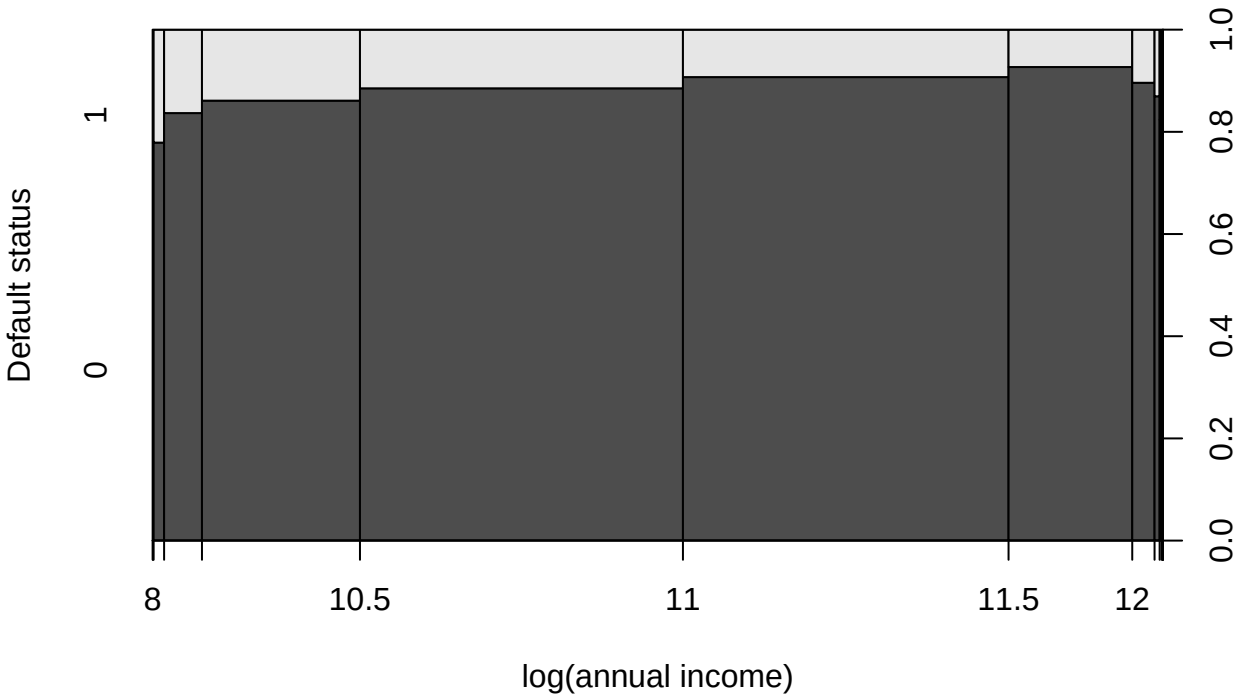
$$R^2_{MF} = 1 - \frac{\log \mathcal{L}(\theta; \mathbf{y})}{\log \mathcal{L}_0(\mathbf{y})}.$$

Intuitively, $R^2_{MF} = 0$ if the explanatory variables do not convey any information on the outcome y. Indeed, in this case, the model is not better than the reference model, that simply captures the fraction of y_i's that are equal to 1.

Example 7.3 (Credit and defaults (Lending-club dataset))**.** This example makes use of the `credit` data of package `AEC`. The objective is to model the default probabilities of borrowers.

Let us first represent the relationship between the fraction of households that have defaulted on their loan and their annual income:

```
library(AEC)
credit$Default <- 0
credit$Default[credit$loan_status == "Charged Off"] <- 1
credit$Default[credit$loan_status ==
                 "Does not meet the credit policy. Status:Charged Off"] <- 1
credit$amt2income <- credit$loan_amnt/credit$annual_inc
plot(as.factor(credit$Default)~log(credit$annual_inc),
     ylevels=2:1,ylab="Default status",xlab="log(annual income)")
```



The previous figure suggests that the effect of annual income on the probability of default is non-monotonous. We will therefore include a quadratic term in one of our specification (namely `eq1` below).

We consider three specifications. The first one (`eq0`), with no explanatory variables, is trivial. It will just be used to compute the pseudo-R^2. In the second (`eq1`), we consider a few covariates (loan amount, the ratio between the amount and annual income, The number of more-than-30 days past-due incidences of delinquency in the borrower's credit file for the past 2 years, and a quadratic function of annual income). In the third model (`eq2`), we add a credit rating.

```
eq0 <- glm(Default ~ 1,data=credit,family=binomial(link="probit"))
eq1 <- glm(Default ~ log(loan_amnt) + amt2income + delinq_2yrs +
             log(annual_inc)+ I(log(annual_inc)^2),
           data=credit,family=binomial(link="probit"))
eq2 <- glm(Default ~ grade + log(loan_amnt) + amt2income + delinq_2yrs +
             log(annual_inc)+ I(log(annual_inc)^2),
           data=credit,family=binomial(link="probit"))
stargazer::stargazer(eq0,eq1,eq2,type="text",no.space = TRUE)
```

```
##
## ================================================
##                          Dependent variable:
##                     ----------------------------
##                                 Default
##                        (1)        (2)        (3)
## ------------------------------------------------
## gradeB                                    0.400***
##                                           (0.055)
## gradeC                                    0.587***
##                                           (0.057)
## gradeD                                    0.820***
##                                           (0.061)
## gradeE                                    0.874***
##                                           (0.091)
## gradeF                                    1.230***
##                                           (0.147)
## gradeG                                    1.439***
##                                           (0.227)
## log(loan_amnt)                 -0.149**  -0.194***
##                                (0.060)    (0.061)
## amt2income                     1.266***   1.222***
##                                (0.383)    (0.393)
## delinq_2yrs                    0.096***    0.009
##                                (0.034)    (0.035)
## log(annual_inc)                -1.444**    -0.874
##                                (0.569)    (0.586)
## I(log(annual_inc)2)            0.064**     0.038
##                                (0.025)    (0.026)
## Constant           -1.231***   7.937***    4.749
##                     (0.017)    (3.060)    (3.154)
## ------------------------------------------------
## Observations         9,156      9,156      9,156
## Log Likelihood     -3,157.696 -3,120.625 -2,981.343
## Akaike Inf. Crit.  6,317.392  6,253.250  5,986.686
```

```
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Let us compute the pseudo R2 for the last two models:

```
logL0 <- logLik(eq0);logL1 <- logLik(eq1);logL2 <- logLik(eq2)
pseudoR2_eq1 <- 1 - logL1/logL0 # pseudo R2
pseudoR2_eq2 <- 1 - logL2/logL0 # pseudo R2
c(pseudoR2_eq1,pseudoR2_eq2)
```

```
## [1] 0.01173993 0.05584870
```

Let us now compute the (average) marginal effects, using method ii of Section 7.4:

```
mean(dnorm(predict(eq2)),na.rm=TRUE)*eq2$coefficients
```

```
##          (Intercept)                gradeB                gradeC
##          0.840731198          0.070747353          0.103944305
##                gradeD                gradeE                gradeF
##          0.145089219          0.154773742          0.217702041
##                gradeG       log(loan_amnt)            amt2income
##          0.254722161         -0.034289921          0.216251992
##          delinq_2yrs      log(annual_inc) I(log(annual_inc)^2)
##          0.001574178         -0.154701321          0.006813694
```

There is an issue for the `annual_inc` variable. Indeed, the previous computation does not realize that this variable appears twice among the explanatory variables (through `log(annual_inc)` and `I(log(annual_inc)^2)`). To address this, one can proceed as follows: (1) we construct a new counterfactual dataset where annual incomes are increased by 1%, (2) we use the model to compute model-implied probabilities of default on this new dataset and (3), we subtract the probabilities resulting from the original dataset from these counterfactual probabilities:

```
new_credit <- credit
new_credit$annual_inc <- 1.01 * new_credit$annual_inc
bas_predict_eq2  <- predict(eq2, newdata = credit, type = "response")
# This is equivalent to pnorm(predict(eq2, newdata = credit))
new_predict_eq2  <- predict(eq2, newdata = new_credit, type = "response")
mean(new_predict_eq2 - bas_predict_eq2)
```

```
## [1] -6.562126e-05
```

The negative sign means that, on average across the entities considered in the analysis, a 1% increase in annual income results in a decrease in the default probability. This average effect is however pretty low. To get an economic sense of the size of this effect, let us compute the average effect associated with a unit increase in the number of delinquencies:

```
new_credit <- credit
new_credit$delinq_2yrs <- credit$delinq_2yrs + 1
new_predict_eq2  <- predict(eq2, newdata = new_credit, type = "response")
mean(new_predict_eq2 - bas_predict_eq2)
```

```
## [1] 0.001582332
```

We can employ a likelihood ratio test (see Def. 6.8) to see if the two variables associated with annual income are jointly statistically significant (in the context of `eq1`):

```
eq1restr <- glm(Default ~ log(loan_amnt) + amt2income + delinq_2yrs,
                data=credit,family=binomial(link="probit"))
LRstat <- 2*(logL1 - logLik(eq1restr))
pvalue <- 1 - c(pchisq(LRstat,df=2))
```

The computation gives a p-value of 0.0436.

Example 7.4 (Replicating Table 14.2 of Cameron and Trivedi (2005))**.** The following lines of codes replicate Table 14.2 of Cameron and Trivedi (2005) (see Example 7.2).

```
data.reduced <- subset(Fishing,mode %in% c("charter","pier"))
data.reduced$lnrelp <- log(data.reduced$price.charter/data.reduced$price.pier)
data.reduced$y <- 1*(data.reduced$mode=="charter")
# check first line of Table 14.1:
price.charter.y0 <- mean(data.reduced$pcharter[data.reduced$y==0])
price.charter.y1 <- mean(data.reduced$pcharter[data.reduced$y==1])
price.charter    <- mean(data.reduced$pcharter)
# Run probit regression:
reg.probit <- glm(y ~ lnrelp,
                  data=data.reduced,
                  family=binomial(link="probit"))
# Run Logit regression:
reg.logit <- glm(y ~ lnrelp,
                 data=data.reduced,
                 family=binomial(link="logit"))
# Run OLS regression:
reg.OLS <- lm(y ~ lnrelp,
              data=data.reduced)
# Replicates Table 14.2 of Cameron and Trivedi:
stargazer::stargazer(reg.logit, reg.probit, reg.OLS,no.space = TRUE,
                     type="text")
```

```
##
## ================================================================
##                                 Dependent variable:
##                     ---------------------------------------------
##                                          y
##                     logistic   probit             OLS
##                        (1)       (2)              (3)
## ----------------------------------------------------------------
## lnrelp              -1.823*** -1.056***        -0.243***
##                      (0.145)   (0.075)          (0.010)
## Constant            2.053***  1.194***          0.784***
##                      (0.169)   (0.088)          (0.013)
## ----------------------------------------------------------------
## Observations           630       630              630
## R2                                                0.463
## Adjusted R2                                       0.462
## Log Likelihood      -206.827  -204.411
## Akaike Inf. Crit.    417.654   412.822
## Residual Std. Error                          0.330 (df = 628)
## F Statistic                                542.123*** (df = 1; 628)
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

7.6 Predictions and ROC curves

How to compute model-implied predicted outcomes? As is the case for y_i, predicted outcomes $\widehat{y}_i$ need to be valued in $\{0, 1\}$. A natural choice consists in considering that $\widehat{y}_i = 1$ if $\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) > 0.5$, i.e., in taking a cutoff of $c = 0.5$. There exist, though, situations where doing so is not relevant. For instance, we may have some models where all predicted probabilities are small, but some less than others. In this context, a model-implied probability of 10% (say) could characterize a "high-risk" entity. However, using a cutoff of 50% would not identify this level of riskiness.

The **receiver operating characteristics (ROC)** curve consitutes a more general approach. The idea is to remain agnostic and to consider all possible values of the cutoff c. It works as follows. For each potential cutoff $c \in [0, 1]$, compute (and plot):

- The fraction of $y = 1$ values correctly classified (*True Positive Rate*) against
- The fraction of $y = 0$ values incorrectly specified (*False Positive Rate*).

Such a curve mechanically starts at (0,0) —which corresponds to $c = 1$— and terminates at (1,1) –situation when $c = 0$.

In the case of no predictive ability (worst situation), the ROC curve is a straight line between (0,0) and (1,1).

Example 7.5 (ROC with the fishing-mode dataset)**.** Figure 7.5 shows the ROC curve associated with the probit model estimated in Example 7.4.

```
library(pROC)
predict_model <- predict.glm(reg.probit,type = "response")
roc(data.reduced$y, predict_model, percent=T,
    boot.n=1000, ci.alpha=0.9, stratified=T, plot=TRUE, grid=TRUE,
    show.thres=TRUE, legacy.axes = TRUE, reuse.auc = TRUE,
    print.auc = TRUE, print.thres.col = "blue", ci=TRUE,
    ci.type="bars", print.thres.cex = 0.7, col = 'red',
    main = paste("ROC curve using","(N = ",nrow(data.reduced),")") )
```

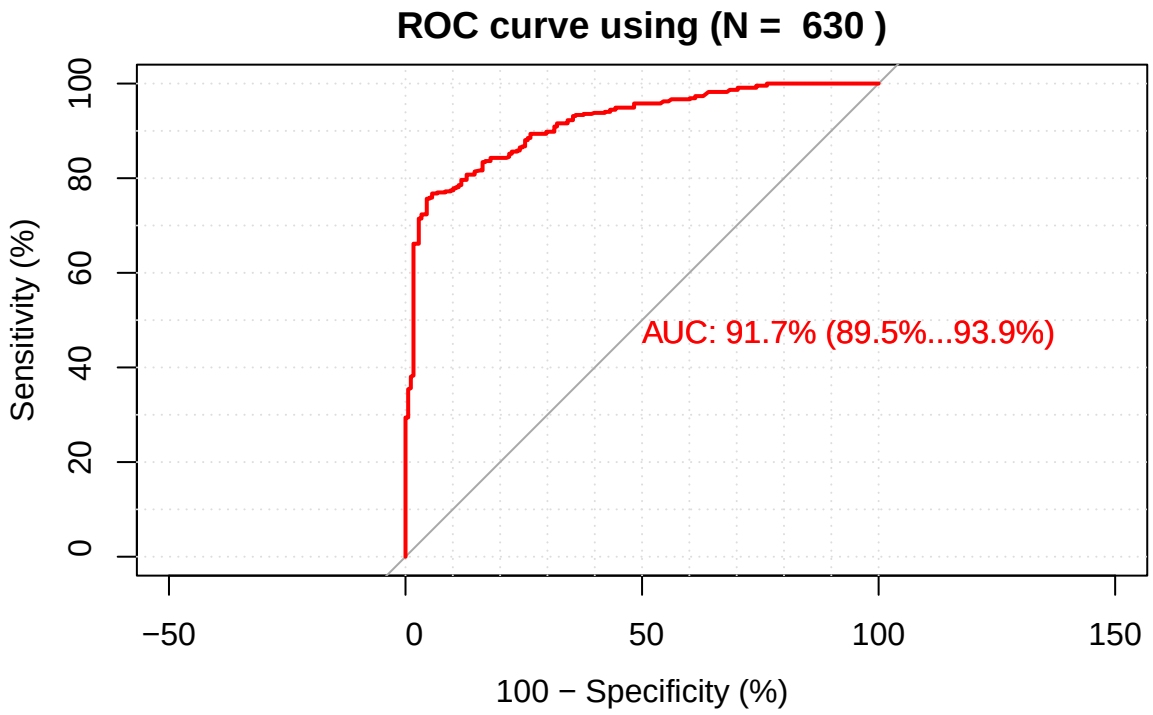


Figure 7.5: Application of the ROC methodology on the fishing-mode dataset.

Chapter 8

Time Series

8.1 Introduction to time series

A time series is an infinite sequence of random variables indexed by time: $\{y_t\}_{t=-\infty}^{+\infty} = \{..., y_{-2}, y_{-1}, y_0, y_1, ..., y_t, ...\}$, $y_i \in \mathbb{R}^k$. In practice, we only observe samples, typically: $\{y_1, ..., y_T\}$.

Standard time series models are built using **shocks** that we will often denote by ε_t. Typically, $\mathbb{E}(\varepsilon_t) = 0$. In many models, the shocks are supposed to be i.i.d., but there exist other (less restrictive) notions of shocks. In particular, the definition of many processes is based on whote noises:

Definition 8.1 (White noise)**.** The process $\{\varepsilon_t\}_{t\in]-\infty,+\infty[}$ is a white noise if, for all t:

a. $\mathbb{E}(\varepsilon_t) = 0$,
b. $\mathbb{E}(\varepsilon_t^2) = \sigma^2 < \infty$ and
c. for all $s \neq t$, $\mathbb{E}(\varepsilon_t \varepsilon_s) = 0$.

Another type of shocks that are commonly used are Martingale Difference Sequences:

Definition 8.2 (Martingale Difference Sequence)**.** The process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a martingale difference sequence (MDS) if $\mathbb{E}(|\varepsilon_t|) < \infty$ and if, for all t,

$$\underbrace{\mathbb{E}_{t-1}(\varepsilon_t)}_{\text{Expectation conditional on the past}} = 0.$$

By definition, if y_t is a martingale, then $y_t - y_{t-1}$ is a MDS.

Example 8.1 (ARCH process)**.** The Autoregressive conditional heteroskedasticity (ARCH) process is an example of shock that satisfies the MDS definition but that is not i.i.d.:

$$\varepsilon_t = \sigma_t \times z_t,$$

where $z_t \sim i.i.d.\, \mathcal{N}(0,1)$ and $\sigma_t^2 = w + \alpha \varepsilon_{t-1}^2$.

Example 8.2. A white noise process is not necessarily a MDS. This is for instance the following process:

$$\varepsilon_t = z_t + z_{t-1}z_{t-2},$$

where $z_t \sim i.i.d.\mathcal{N}(0,1)$.

Let us now introduce the lag operator. The lag operator, denoted by L, is defined on the time series space and is defined by:

$$L : \{y_t\}_{t=-\infty}^{+\infty} \to \{w_t\}_{t=-\infty}^{+\infty} \quad \text{with} \quad w_t = y_{t-1}. \tag{8.1}$$

We have: $L^2 y_t = y_{t-2}$ and, more generally, $L^k y_t = y_{t-k}$.

Consider a time series y_t defined by $y_t = \mu + \phi y_{t-1} + \varepsilon_t$, where the ε_t's are i.i.d. $\mathcal{N}(0,\sigma^2)$. Using the lag operator, the dynamics of y_t can be expressed as follows:

$$(1 - \phi L)y_t = \mu + \varepsilon_t.$$

It is easily checked that we have $L^2 y_t = y_{t-2}$ and, generally, $L^k y_t = y_{t-k}$.

If it exists, the **unconditional (or marginal) mean** of the random variable y_t is given by:

$$\mu_t := \mathbb{E}(y_t) = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t) dy_t,$$

where f_{Y_t} is the unconditional (or marginal) density of y_t. Similarly, if it exists, the **unconditional (or marginal) variance** of the random variable y_t is:

$$\mathbb{V}ar(y_t) = \int_{-\infty}^{\infty} (y_t - \mathbb{E}(y_t))^2 f_{Y_t}(y_t) dy_t.$$

Definition 8.3 (Autocovariance)**.** The j^{th} autocovariance of y_t is given by:

$$\begin{aligned} \gamma_{j,t} &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} [y_t - \mathbb{E}(y_t)][y_{t-j} - \mathbb{E}(y_{t-j})] \times \\ &\quad f_{Y_t, Y_{t-1},\dots,Y_{t-j}}(y_t, y_{t-1}, \dots, y_{t-j}) dy_t dy_{t-1} \dots dy_{t-j} \\ &= \mathbb{E}([y_t - \mathbb{E}(y_t)][y_{t-j} - \mathbb{E}(y_{t-j})]), \end{aligned}$$

where $f_{Y_t, Y_{t-1},\dots,Y_{t-j}}(y_t, y_{t-1}, \dots, y_{t-j})$ is the joint distribution of $y_t, y_{t-1}, \dots, y_{t-j}$.

In particular, $\gamma_{0,t} = \mathbb{V}ar(y_t)$.

Definition 8.4 (Covariance stationarity)**.** The process y_t is covariance stationary —or weakly stationary— if, for all t and j,

$$\mathbb{E}(y_t) = \mu \quad \text{and} \quad \mathbb{E}\{(y_t - \mu)(y_{t-j} - \mu)\} = \gamma_j.$$

Figure 8.1 displays the simulation of a process that is not covariance stationary. This process follows $y_t = 0.1t + \varepsilon_t$, where $\varepsilon_t \sim i.i.d.\,\mathcal{N}(0,1)$. Indeed, for such a process, we have: $\mathbb{E}(y_t) = 0.1t$, which depends on t.

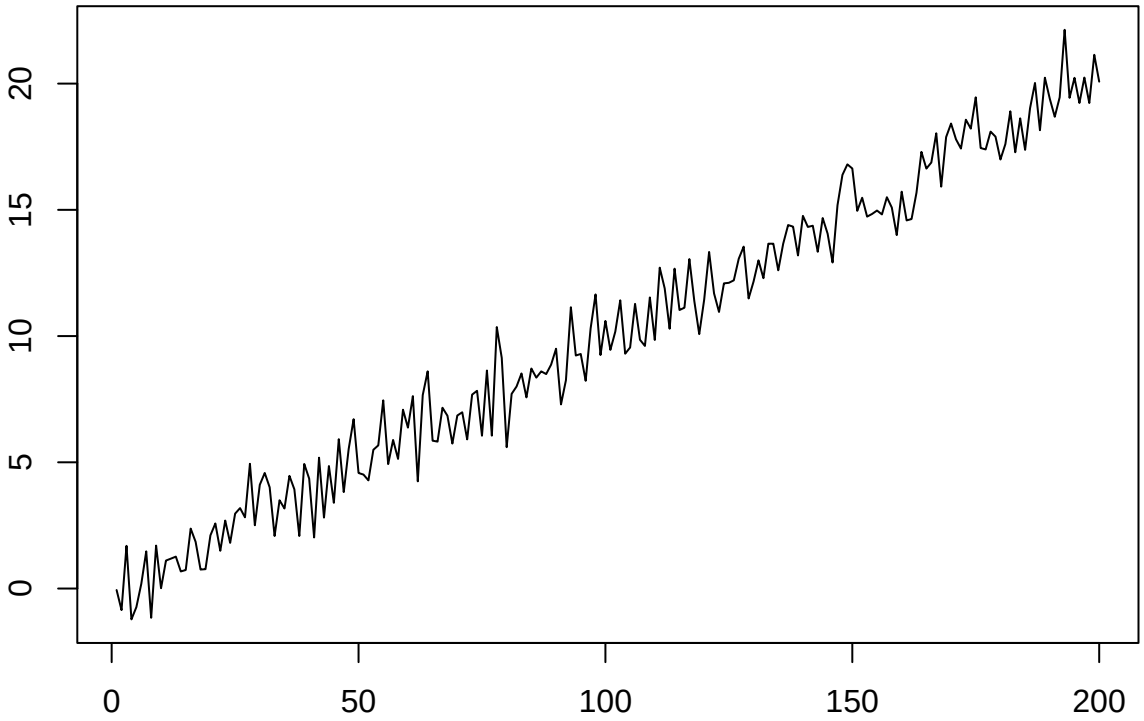


Figure 8.1: Example of a process that is not covariance stationary ($y_t = 0.1t + \varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0,1)$).

Definition 8.5 (Strict stationarity)**.** The process y_t is strictly stationary if, for all t and all sets of integers $J = \{j_1, ..., j_n\}$, the distribution of $(y_t, y_{t+j_1}, ..., y_{t+j_n})$ depends on J but not on t.

The following process is covariance stationary but not strictly stationary:

$$y_t = \mathbb{I}_{\{t<1000\}}\varepsilon_{1,t} + \mathbb{I}_{\{t\geq 1000\}}\varepsilon_{2,t},$$

where $\varepsilon_{1,t} \sim \mathcal{N}(0,1)$ and $\varepsilon_{2,t} \sim \sqrt{\frac{\nu-2}{\nu}}t(\nu)$ and $\nu = 4$.

Proposition 8.1. *If y_t is covariance stationary, then $\gamma_j = \gamma_{-j}$.*

Proof. Since y_t is covariance stationary, the covariance between y_t and y_{t-j} (i.e γ_j) is the same as that between y_{t+j} and y_{t+j-j} (i.e. γ_{-j}). □

Definition 8.6 (Auto-correlation)**.** The j^{th} auto-correlation of a covariance-stationary process is:

$$\rho_j = \frac{\gamma_j}{\gamma_0}.$$

Consider a long historical time series of the Swiss GDP growth, taken from the Jordà et al. (2017) dataset.[1]

[1]Version 6 of the dataset, available on this website.

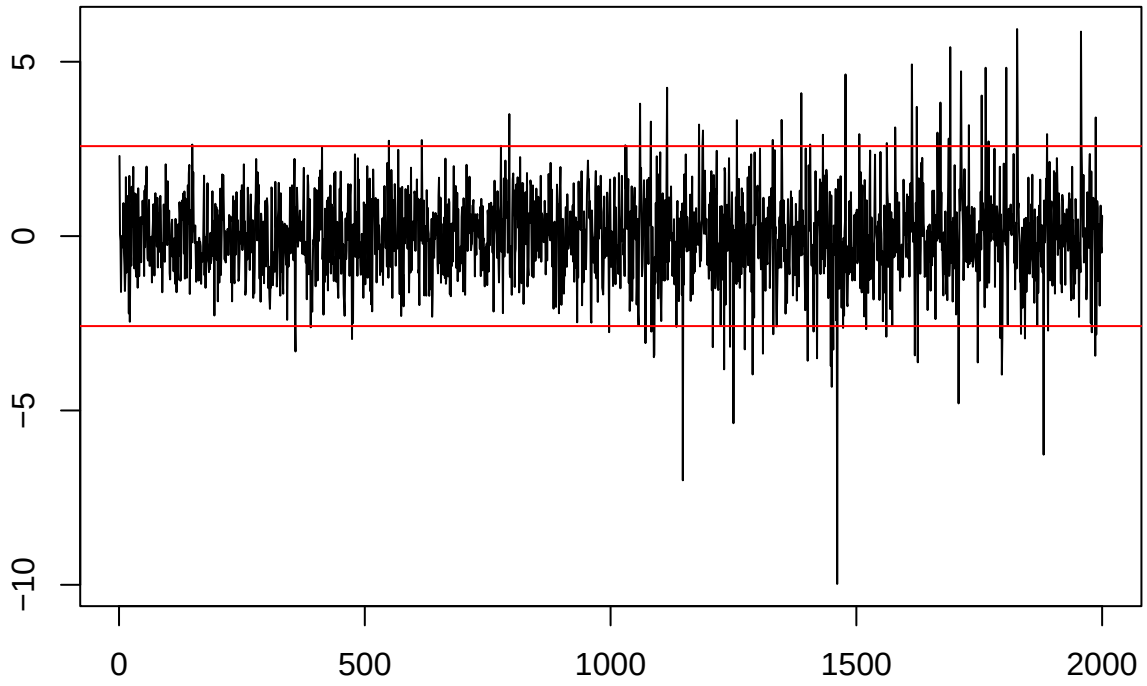


Figure 8.2: Example of a process that is covariance stationary but not strictly stationary. The red lines delineate the 99% confidence interval of the standard normal distribution (± 2.58).

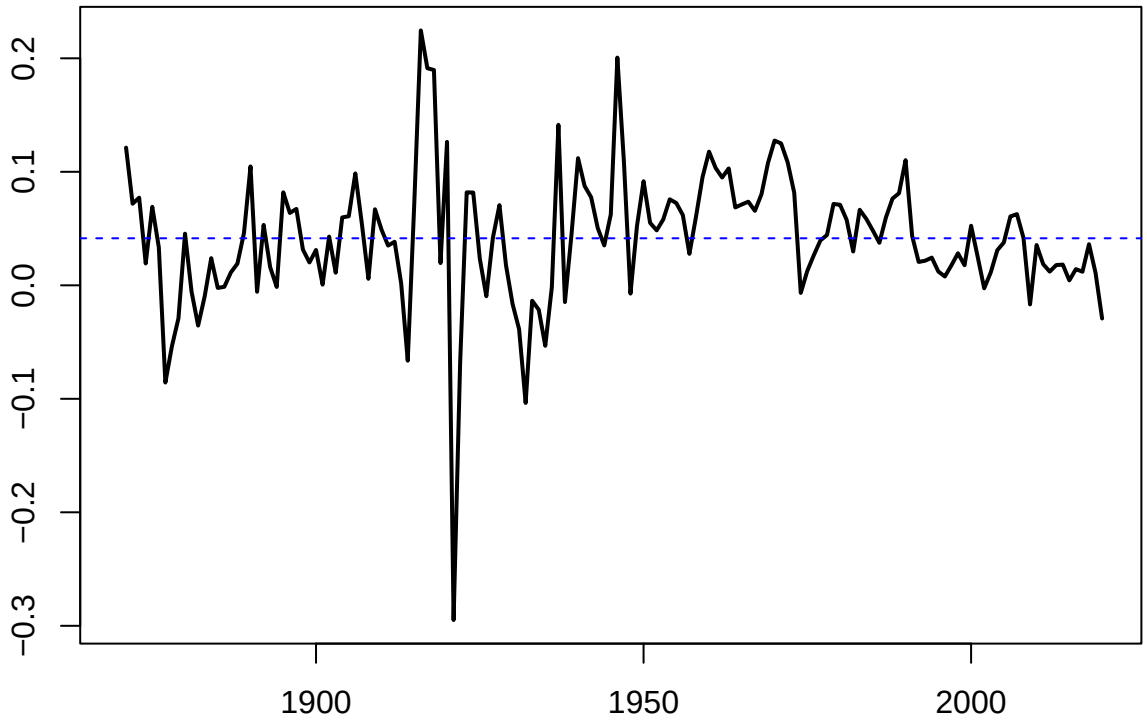


Figure 8.3: Annual growth rate of Swiss GDP, based on the Jorda-Schularick-Taylor Macrohistory Database.

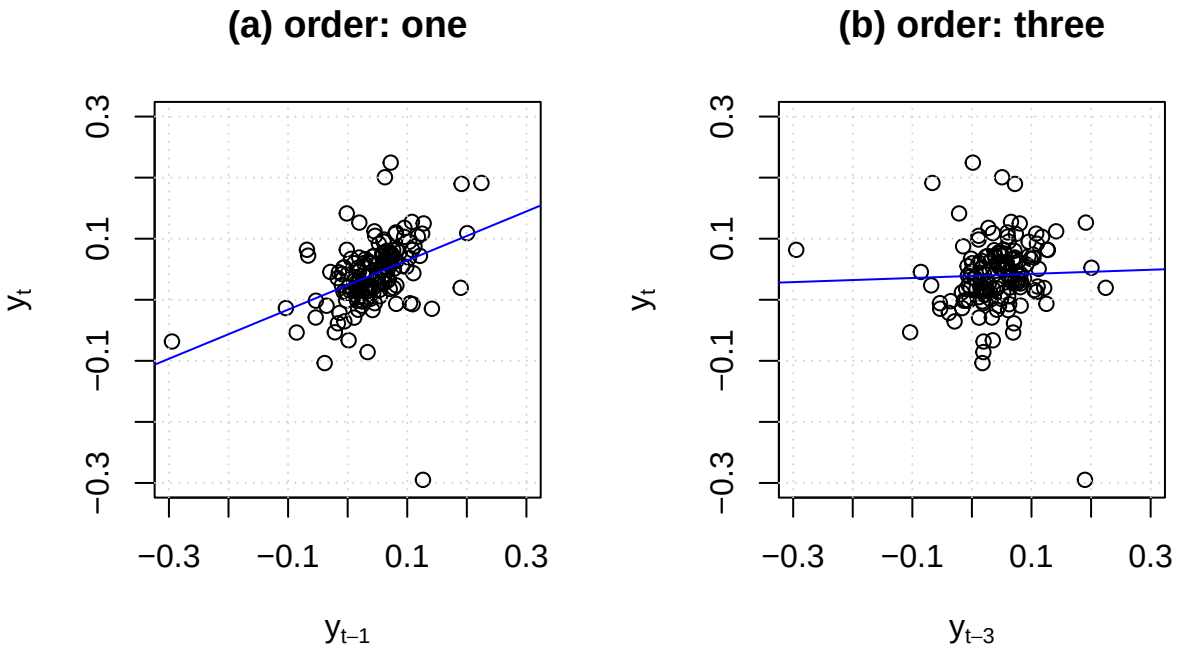


Figure 8.4: For order j, the slope of the blue line is, approximately, $\hat{\gamma}_j/\widehat{\mathbb{V}ar}(y_t)$, where hats indicate sample moments.

Definition 8.7 (Mean ergodicity)**.** The covariance-stationary process y_t is ergodic for the mean if:

$$\text{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T} y_t = \mathbb{E}(y_t).$$

Definition 8.8 (Second-moment ergodicity)**.** The covariance-stationary process y_t is ergodic for second moments if, for all j:

$$\text{plim}_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T}(y_t-\mu)(y_{t-j}-\mu) = \gamma_j.$$

It should be noted that ergodicity and stationarity are different properties. Typically if the process $\{x_t\}$ is such that, $\forall t$, $x_t \equiv y$, where $y \sim \mathcal{N}(0,1)$ (say), then $\{x_t\}$ is stationary but not ergodic.

Theorem 8.1 (Central Limit Theorem for covariance-stationary processes)**.** *If process y_t is covariance stationary and if the series of autocovariances is absolutely summable ($\sum_{j=-\infty}^{+\infty}|\gamma_j| < \infty$), then:*

$$\bar{y}_T \overset{m.s.}{\to} \mu = \mathbb{E}(y_t) \tag{8.2}$$

$$lim_{T\to+\infty} T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right] = \sum_{j=-\infty}^{+\infty}\gamma_j \tag{8.3}$$

$$\sqrt{T}(\bar{y}_T-\mu) \overset{d}{\to} \mathcal{N}\left(0, \sum_{j=-\infty}^{+\infty}\gamma_j\right). \tag{8.4}$$

[Mean square (m.s.) and distribution (d.) convergences: see Definitions 9.19 and 9.17.]

Proof. By Proposition 9.8, Eq. (8.3) implies Eq. (8.2). For Eq. (8.3), see Appendix 9.5. For Eq. (8.4), see Anderson (1971), p. 429. □

Definition 8.9 (Long-run variance)**.** Under the assumptions of Theorem 8.1, the limit appearing in Eq. (8.3) exists and is called **long-run variance**. It is denoted by S, i.e.:

$$S = \Sigma_{j=-\infty}^{+\infty} \gamma_j = \lim_{T \to +\infty} T\mathbb{E}[(\bar{y}_T - \mu)^2].$$

If y_t is ergodic for second moments (see Def. 8.8), a natural estimator of S is:

$$\hat{\gamma}_0 + 2\sum_{\nu=1}^{q} \hat{\gamma}_\nu, \tag{8.5}$$

where $\hat{\gamma}_\nu = \frac{1}{T}\sum_{\nu+1}^{T}(y_t - \bar{y})(y_{t-\nu} - \bar{y})$.

However, for small samples, Eq. (8.5) does not necessarily result in a positive definite matrix. Newey and West (1987) have proposed an estimator that does not have this defect. Their estimator is given by:

$$S^{NW} = \hat{\gamma}_0 + 2\sum_{\nu=1}^{q} \left(1 - \frac{\nu}{q+1}\right) \hat{\gamma}_\nu. \tag{8.6}$$

Loosely speaking, Theorem 8.1 says that, for a given sample size, the higher the "persistency" of a process, the lower the accuracy of the sample mean as an estimate of the population mean. To illustrate, consider three processes that feature the same marginal variance (equal to one, say), but different autocorrelations: 0%, 70%, and 99.9%. Figure 8.5 displays simulated paths of such three processes. It indeed appears that, the larger the autocorrelation of the process, the further the sample mean (dashed red line) from the population mean (red solid line).

The same type of simulations can be performed using this ShinyApp (use panel "AR(1) Simulation").

8.2 Univariate processes

8.2.1 Moving Average (MA) processes

Definition 8.10. Consider a white noise process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ (Def. 8.1). Then y_t is a first-order moving average process if, for all t:

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}.$$

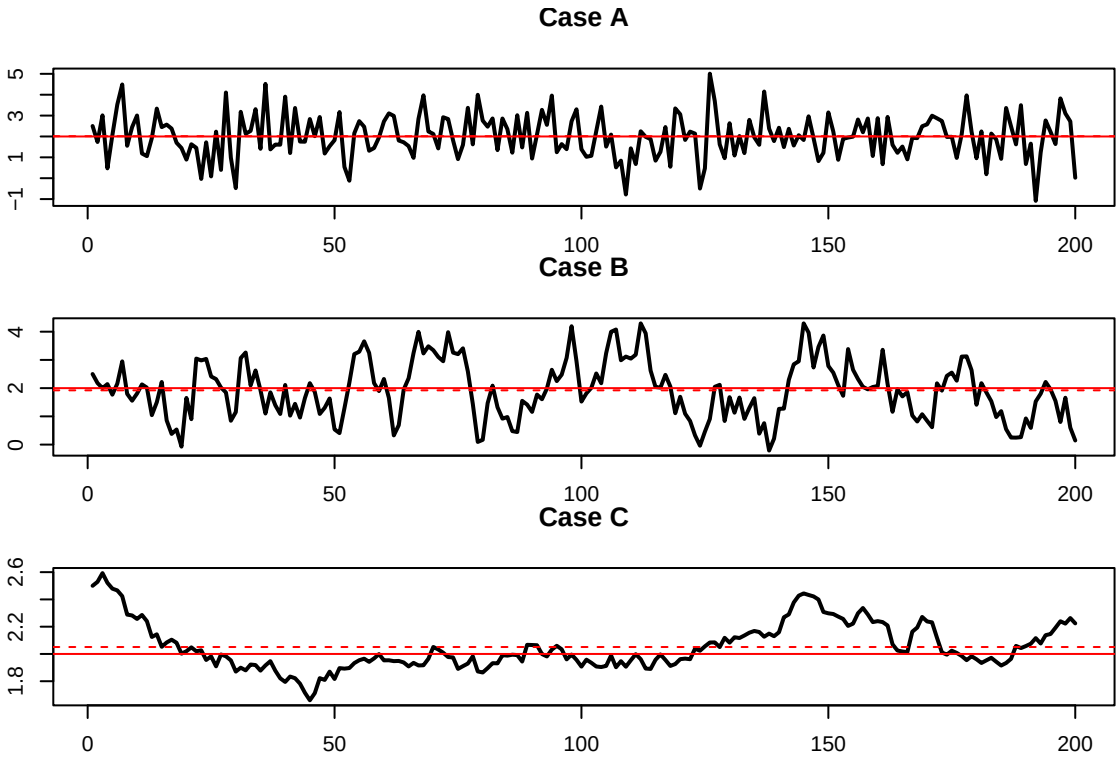


Figure 8.5: The three samples have been simulated using the following data generating process: $x_t = \mu + \rho(x_{t-1} - \mu) + \sqrt{1-\rho^2}\varepsilon_t$, where $\varepsilon_t \sim \mathcal{N}(0,1)$. Case A: $\rho = 0$; Case B: $\rho = 0.7$; Case C: $\rho = 0.999$. In the three cases, $\mathbb{E}(x_t) = \mu = 2$ and $\mathbb{V}ar(x_t) = 1$.

If $\mathbb{E}(\varepsilon_t^2) = \sigma^2$, it is easily obtained that the unconditional mean and variances of y_t are:

$$\mathbb{E}(y_t) = \mu, \quad \mathbb{V}ar(y_t) = (1 + \theta^2)\sigma^2.$$

The first auto-covariance is:

$$\gamma_1 = \mathbb{E}\{(y_t - \mu)(y_{t-1} - \mu)\} = \theta\sigma^2.$$

Higher-order auto-covariances are zero ($\gamma_j = 0$ for $j > 1$). Therefore: An MA(1) process is covariance-stationary (Def. 8.4).

For a MA(1) process, the autocorrelation of order j (see Def. 8.6) is given by:

$$\rho_j = \begin{cases} 1 & \text{if} \quad j = 0, \\ \theta/(1 + \theta^2) & \text{if} \quad j = 1 \\ 0 & \text{if} \quad j > 1. \end{cases}$$

Notice that process y_t defined through:

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},$$

where $\mathbb{V}ar(\varepsilon_t) = \sigma^2$, has the same mean and autocovariances as

$$y_t = \mu + \varepsilon_t^* + \frac{1}{\theta}\varepsilon_{t-1}^*,$$

where $\mathbb{V}ar(\varepsilon_t^*) = \theta^2\sigma^2$. That is, even if we perfectly know the mean and auto-covariances of this process, it is not possible to identify which specification is the one that has been used to generate the data. Only one of these two specifications is said to be *fundamental*, that is the one that satisfies $|\theta_1| < 1$.

Definition 8.11 (MA(q) process)**.** A q^{th} order Moving Average process is defined through:

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}.$$

where $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a white noise process (Def. 8.1).

Proposition 8.2 (Covariance-stationarity of an MA(q) process)**.** *Finite-order Moving Average processes are covariance-stationary.*

Moreover, the autocovariances of an MA(q) process (as defined in Def. 8.11) are given by:

$$\gamma_j = \begin{cases} \sigma^2(\theta_j\theta_0 + \theta_{j+1}\theta_1 + \cdots + \theta_q\theta_{q-j}) & \textit{for} \quad j \in \{0, \ldots, q\} \\ 0 & \textit{for} \quad j > q, \end{cases} \tag{8.7}$$

where we use the notation $\theta_0 = 1$, and $\mathbb{V}ar(\varepsilon_t) = \sigma^2$.

Proof. The unconditional expectation of y_t does not depend on time, since $\mathbb{E}(y_t) = \mu$. Let's turn to autocovariances. We can extend the series of the θ_j's by setting $\theta_j = 0$ for $j > q$. We then have:

$$\begin{aligned}\mathbb{E}((y_t - \mu)(y_{t-j} - \mu)) &= \mathbb{E}\left[(\theta_0\varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_j\varepsilon_{t-j} + \theta_{j+1}\varepsilon_{t-j-1} + \dots) \times \right. \\ &\quad \left. (\theta_0\varepsilon_{t-j} + \theta_1\varepsilon_{t-j-1} + \dots)\right].\end{aligned}$$

Then use the fact that $\mathbb{E}(\varepsilon_t\varepsilon_s) = 0$ if $t \neq s$ (because $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a white noise process). □

Figure 8.6 displays simulated paths of two MA processes (an MA(1) and an MA(4)). Such simulations can be produced by using panel "ARMA(p,q)" of this web interface.

```
library(AEC)
T <- 100;nb.sim <- 1
y.0 <- c(0)
c <- 1;phi <- c(0);sigma <- 1
theta <- c(1,1) # MA(1) specification
y.sim <- sim.arma(c,phi,theta,sigma,T,y.0,nb.sim)
par(mfrow=c(1,2))
par(plt=c(.2,.9,.2,.85))
plot(y.sim[,1],xlab="",ylab="",type="l",lwd=2,
     main=expression(paste(theta[0],"=1, ",theta[1],"=1",sep="")))
abline(h=c)
theta <- c(1,1,1,1,1) # MA(4) specification
y.sim <- sim.arma(c,phi,theta,sigma,T,y.0,nb.sim)
plot(y.sim[,1],xlab="",ylab="",type="l",lwd=2,
     main=expression(paste(theta[0],"=...=",theta[4],"=1",sep="")))
abline(h=c)
```

What if the order q of an MA(q) process gets infinite? The notion of **infinite-order Moving Average process** exists and is important in time series analysis. The (infinite) sequence of θ_j has to satisfy some conditions for such a process to be well-defined (see Theorem 8.2 below). These conditions relate to the "summability" of $\{\theta_i\}_{i\in\mathbb{N}}$ (see Definition 8.12).

Definition 8.12 (Absolute and square summability)**.** The sequence $\{\theta_i\}_{i\in\mathbb{N}}$ is absolutely summable if $\sum_{i=0}^{\infty} |\theta_i| < +\infty$, and it is square summable if $\sum_{i=0}^{\infty} \theta_i^2 < +\infty$.

According to Prop. 9.8, absolute summability implies square summability.

Theorem 8.2 (Existence condition for an infinite MA process)**.** *If $\{\theta_i\}_{i\in\mathbb{N}}$ is square summable (see Def. 8.12) and if $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ is a white noise process (see Def. 8.1), then*

$$\mu + \sum_{i=0}^{+\infty} \theta_i \varepsilon_{t-i}$$

defines a well-behaved [covariance-stationary] process, called infinite-order MA process (MA(∞)).

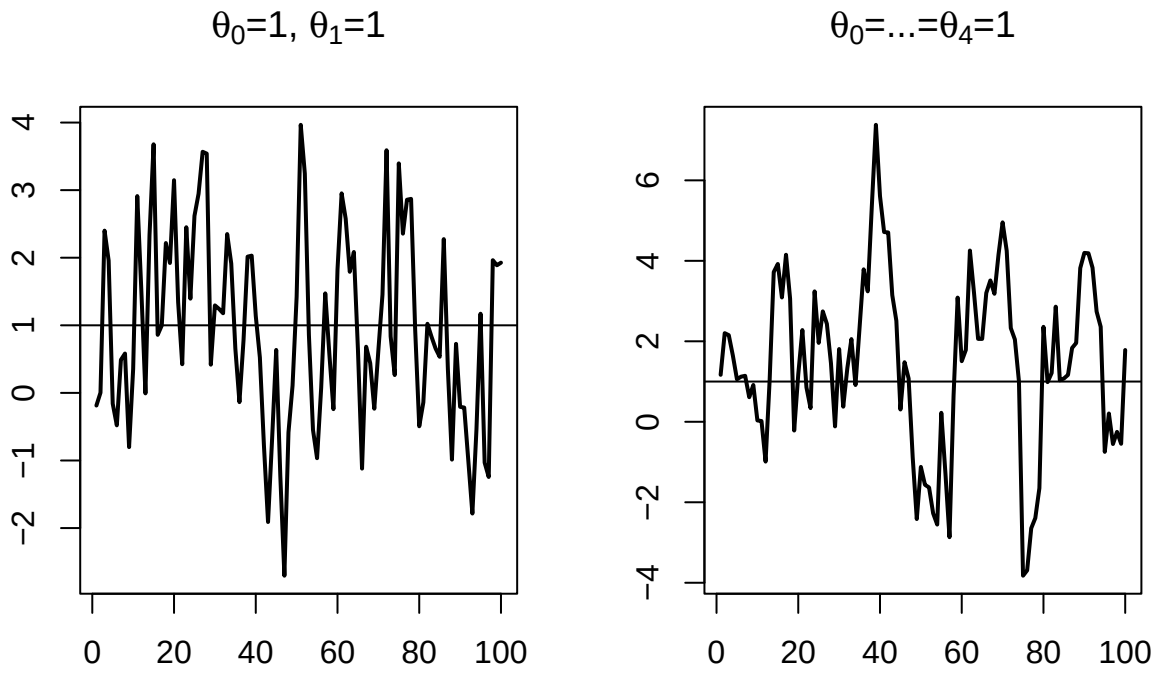


Figure 8.6: Simulation of MA processes.

Proof. See Appendix 3.A in Hamilton. "Well behaved" means that $\Sigma_{i=0}^{T}\theta_{t-i}\varepsilon_{t-i}$ converges in mean square (Def. 9.17) to some random variable Z_t. The proof makes use of the fact that:

$$\mathbb{E}\left[\left(\sum_{i=N}^{M}\theta_i\varepsilon_{t-i}\right)^2\right] = \sum_{i=N}^{M}|\theta_i|^2\sigma^2,$$

and that, when $\{\theta_i\}$ is square summable, $\forall\eta > 0$, $\exists N$ s.t. the right-hand-side term in the last equation is lower than η for all $M \geq N$ (static Cauchy criterion, Theorem 9.2). This implies that $\Sigma_{i=0}^{T}\theta_i\varepsilon_{t-i}$ converges in mean square (stochastic Cauchy criterion, see Theorem 9.3). □

Proposition 8.3 (First two moments of an infinite MA process)**.** *If $\{\theta_i\}_{i\in\mathbb{N}}$ is absolutely summable, i.e. if $\sum_{i=0}^{\infty}|\theta_i| < +\infty$, then*

i. $y_t = \mu + \sum_{i=0}^{+\infty}\theta_i\varepsilon_{t-i}$ *exists (Theorem 8.2) and is such that:*

$$\begin{aligned}\mathbb{E}(y_t) &= \mu\\ \gamma_0 = \mathbb{E}([y_t-\mu]^2) &= \sigma^2(\theta_0^2+\theta_1^2+\dots)\\ \gamma_j = \mathbb{E}([y_t-\mu][y_{t-j}-\mu]) &= \sigma^2(\theta_0\theta_j+\theta_1\theta_{j+1}+\dots).\end{aligned}$$

ii. Process y_t has absolutely summable auto-covariances, which implies that the results of Theorem 8.1 (Central Limit) apply.

Proof. The absolute summability of $\{\theta_i\}$ and the fact that $\mathbb{E}(\varepsilon^2) < \infty$ imply that the order of integration and summation is interchangeable (see Hamilton, 1994, Footnote p. 52), which proves (i). For (ii), see end of Appendix 3.A in Hamilton (1994). □

8.2.2 Auto-Regressive (AR) processes

Definition 8.13 (First-order AR process (AR(1)))**.** Consider a white noise process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ (see Def. 8.1). Process y_t is an AR(1) process if it is defined by the following difference equation:

$$y_t = c + \phi y_{t-1} + \varepsilon_t.$$

That kind of process can be simulated by using panel "AR(1) Simulation" of this web interface.

If $|\phi| \geq 1$, y_t is not stationary. Indeed, we have:

$$y_{t+k} = c + \varepsilon_{t+k} + \phi(c + \varepsilon_{t+k-1}) + \phi^2(c + \varepsilon_{t+k-2}) + \cdots + \phi^{k-1}(c + \varepsilon_{t+1}) + \phi^k y_t.$$

Therefore, the conditional variance

$$\mathbb{V}ar_t(y_{t+k}) = \sigma^2(1 + \phi^2 + \phi^4 + \cdots + \phi^{2(k-1)})$$

does not converge for large k's. This implies that $\mathbb{V}ar(y_t)$ does not exist.

By contrast, if $|\phi| < 1$, one can see that:

$$y_t = c + \varepsilon_t + \phi(c + \varepsilon_{t-1}) + \phi^2(c + \varepsilon_{t-2}) + \cdots + \phi^k(c + \varepsilon_{t-k}) + ...$$

Hence, if $|\phi| < 1$, the unconditional mean and variance of y_t are:

$$\mathbb{E}(y_t) = \frac{c}{1-\phi} =: \mu \quad \text{and} \quad \mathbb{V}ar(y_t) = \frac{\sigma^2}{1-\phi^2}.$$

Let us compute the j^{th} autocovariance of the AR(1) process:

$$\begin{aligned}
\mathbb{E}([y_t - \mu][y_{t-j} - \mu]) &= \mathbb{E}([\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots + \phi^j\varepsilon_{t-j} + \phi^{j+1}\varepsilon_{t-j-1} \ldots] \times \\
&\quad [\varepsilon_{t-j} + \phi\varepsilon_{t-j-1} + \phi^2\varepsilon_{t-j-2} + \cdots + \phi^k\varepsilon_{t-j-k} + \ldots]) \\
&= \mathbb{E}(\phi^j\varepsilon_{t-j}^2 + \phi^{j+2}\varepsilon_{t-j-1}^2 + \phi^{j+4}\varepsilon_{t-j-2}^2 + \ldots) \\
&= \frac{\phi^j\sigma^2}{1-\phi^2}.
\end{aligned}$$

Therefore $\rho_j = \phi^j$.

By what precedes, we have:

Proposition 8.4 (Covariance-stationarity of an AR(1) process)**.** *The AR(1) process, as defined in Def. 8.13, is covariance-stationary iff* $|\phi| < 1$.

Definition 8.14 (AR(p) process)**.** Consider a white noise process $\{\varepsilon_t\}_{t=-\infty}^{+\infty}$ (see Def. 8.1). Process y_t is a p^{th}-order autoregressive process (AR(p)) if its dynamics is defined by the following difference equation (with $\phi_p \neq 0$):

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t. \tag{8.8}$$

As we will see, the covariance-stationarity of process y_t hinges on matrix F defined as:

$$F = \begin{bmatrix} \phi_1 & \phi_2 & \dots & & \phi_p \\ 1 & 0 & \dots & & 0 \\ 0 & 1 & \dots & & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \tag{8.9}$$

Note that this matrix F is such that if y_t follows Eq. (8.8), then process $\mathbf{y}_t$ follows:

$$\mathbf{y}_t = \mathbf{c} + F\mathbf{y}_{t-1} + \xi_t$$

with

$$\mathbf{c} = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \xi_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+1} \end{bmatrix}.$$

Proposition 8.5 (The eigenvalues of matrix F)**.** *The eigenvalues of F (defined by Eq.* (8.9)*) are the solutions of:*

$$\lambda^p - \phi_1\lambda^{p-1} - \dots - \phi_{p-1}\lambda - \phi_p = 0. \tag{8.10}$$

Proposition 8.6 (Covariance-stationarity of an AR(p) process)**.** *These four statements are equivalent:*

i. Process $\{y_t\}$, defined in Def. 8.14, is covariance-stationary.
ii. The eigenvalues of F (as defined Eq. (8.9)*) lie strictly within the unit circle.*
iii. The roots of Eq. (8.11) *(below) lie strictly outside the unit circle.*

$$1 - \phi_1 z - \dots - \phi_{p-1}z^{p-1} - \phi_p z^p = 0. \tag{8.11}$$

iv. The roots of Eq. (8.12) *(below) lie strictly inside the unit circle.*

$$\lambda^p - \phi_1\lambda^{p-1} - \dots - \phi_{p-1}\lambda - \phi_p = 0. \tag{8.12}$$

Proof. We consider the case where the eigenvalues of F are distinct; Jordan decomposition can be used in the general case. When the eigenvalues of F are distinct, F admits the following spectral decomposition: $F = PDP^{-1}$, where D is diagonal. Using the notations introduced in Eq. (8.9), we have:

$$\mathbf{y}_t = \mathbf{c} + F\mathbf{y}_{t-1} + \xi_t.$$

Let's introduce $\mathbf{d} = P^{-1}\mathbf{c}$, $\mathbf{z}_t = P^{-1}\mathbf{y}_t$ and $\eta_t = P^{-1}\xi_t$. We have:

$$\mathbf{z}_t = \mathbf{d} + D\mathbf{z}_{t-1} + \eta_t.$$

Because D is diagonal, the different component of $\mathbf{z}_t$, denoted by $z_{i,t}$, follow AR(1) processes. The (scalar) autoregressive parameters of these AR(1) processes are the diagonal entries of D –which also are the eigenvalues of F– that we denote by λ_i.

Process y_t is covariance-stationary iff $\mathbf{y}_t$ also is covariance-stationary, which is the case iff all $z_{i,t}$, $i \in [1, p]$, are covariance-stationary. By Prop. 8.4, process $z_{i,t}$ is covariance-stationary iff $|\lambda_i| < 1$. This proves that (i) is equivalent to (ii). Prop. 8.5 further proves that (ii) is equivalent to (iv). Finally, it is easily seen that (iii) is equivalent to (iv) (as long as $\phi_p \neq 0$). □

Using the lag operator (see Eq (8.1)), if y_t is a covariance-stationary AR(p) process (Def. 8.14), we can write:

$$y_t = \mu + \psi(L)\varepsilon_t,$$

where

$$\psi(L) = (1 - \phi_1 L - \cdots - \phi_p L^p)^{-1}, \tag{8.13}$$

and

$$\mu = \mathbb{E}(y_t) = \frac{c}{1 - \phi_1 - \cdots - \phi_p}. \tag{8.14}$$

In the following lines of codes, we compute the eigenvalues of the F matrices associated with the following processes (where ε_t is a white noise):

$$\begin{aligned} x_t &= 0.9x_{t-1} - 0.2x_{t-2} + \varepsilon_t \\ y_t &= 1.1y_{t-1} - 0.3y_{t-2} + \varepsilon_t \\ w_t &= 1.4w_{t-1} - 0.7w_{t-2} + \varepsilon_t \\ z_t &= 0.9z_{t-1} + 0.2z_{t-2} + \varepsilon_t \end{aligned}$$

```
F <- matrix(c(.9,1,-.2,0),2,2)
lambda_x <- eigen(F)$values
F[1,] <- c(1.1,-.3)
lambda_y <- eigen(F)$values
F[1,] <- c(1.4,-.7)
lambda_w <- eigen(F)$values
F[1,] <- c(.9,.2)
lambda_z <- eigen(F)$values
rbind(lambda_x,lambda_y,lambda_w,lambda_z)
```

```
##                          [,1]                  [,2]
## lambda_x 0.500000+0.0000000i  0.4000000+0.0000000i
## lambda_y 0.600000+0.0000000i  0.5000000+0.0000000i
## lambda_w 0.700000+0.4582576i  0.7000000-0.4582576i
## lambda_z 1.084429+0.0000000i -0.1844289+0.0000000i
```

The absolute values of the eigenvalues associated with process w_t are both equal to 0.837. Therefore, according to Proposition 8.6, processes x_t, y_t, and w_t are covariance-stationary, but not z_t (because the absolute value of one of the eigenvalues of the F matrix associated with this process is larger than 1).

The computation of the autocovariances of y_t is based on the so-called **Yule-Walker equations** (Eq. (8.15)). Let's rewrite Eq. (8.8):

$$(y_t - \mu) = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \varepsilon_t.$$

Multiplying both sides by $y_{t-j} - \mu$ and taking expectations leads to the (Yule-Walker) equations:

$$\gamma_j = \begin{cases} \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p} & if \quad j > 0 \\ \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma^2 & for \quad j = 0. \end{cases} \tag{8.15}$$

Using $\gamma_j = \gamma_{-j}$ (Prop. 8.1), one can express $(\gamma_0, \gamma_1, \ldots, \gamma_p)$ as functions of $(\sigma^2, \phi_1, \ldots, \phi_p)$.

8.2.3 PACF approach to identify AR/MA processes

We have seen that the k^{th}-order auto-correlation of a MA(q) process is null if $k > q$. This is exploited, in practice, to determine the order of a MA process. Moreover, since this is not the case for an AR process, this can be used to distinguish an AR from an MA process.

There exists an equivalent approach to determine whether a process can be modeled as an AR process; it is based on partial auto-correlations:

Definition 8.15 (Partial auto-correlation)**.** In a time series context, the partial auto-correlation $(\phi_{h,h})$ of process $\{y_t\}$ is defined as the partial correlation of y_{t+h} and y_t given $y_{t+h-1}, \cdots, y_{t+1}$. (see Def. 9.5 for the definition of partial correlation.)

If $h > p$, the regression of y_{t+h} on $y_{t+h-1}, \ldots, y_{t+1}$ is:

$$y_{t+h} = c + \phi_1 y_{t+h-1} + \cdots + \phi_p y_{t+h-p} + \varepsilon_{t+h}.$$

The residuals of the latter regressions (ε_{t+h}) are uncorrelated to y_t. Then the partial auto-correlation is zero for $h > p$.

Besides, it can be shown that $\phi_{p,p} = \phi_p$. Hence $\phi_{p,p} = \phi_p$ but $\phi_{h,h} = 0$ for $h > p$. This can be used to determine the order of an AR process. By contrast (importantly) if y_t follows an MA(q) process, then $\phi_{k,k}$ asymptotically approaches zero instead of cutting off abruptly.

As illustrated below, functions `acf` and `pacf` can be conveniently used to employ the (P)ACF approach. (Note also the use of function `sim.arma` to simulate ARMA processes.)

```
library(AEC)
par(mfrow=c(3,2))
par(plt=c(.2,.9,.2,.95))
theta <- c(1,2,1);phi=0
y.sim <- sim.arma(c=0,phi,theta,sigma=1,T=1000,y.0=0,nb.sim=1)
par(mfg=c(1,1));plot(y.sim,type="l",lwd=2)
par(mfg=c(2,1));acf(y.sim)
par(mfg=c(3,1));pacf(y.sim)
theta <- c(1);phi=0.9
```

```
y.sim <- sim.arma(c=0,phi,theta,sigma=1,T=1000,y.0=0,nb.sim=1)
par(mfg=c(1,2));plot(y.sim,type="l",lwd=2)
par(mfg=c(2,2));acf(y.sim)
par(mfg=c(3,2));pacf(y.sim)
```

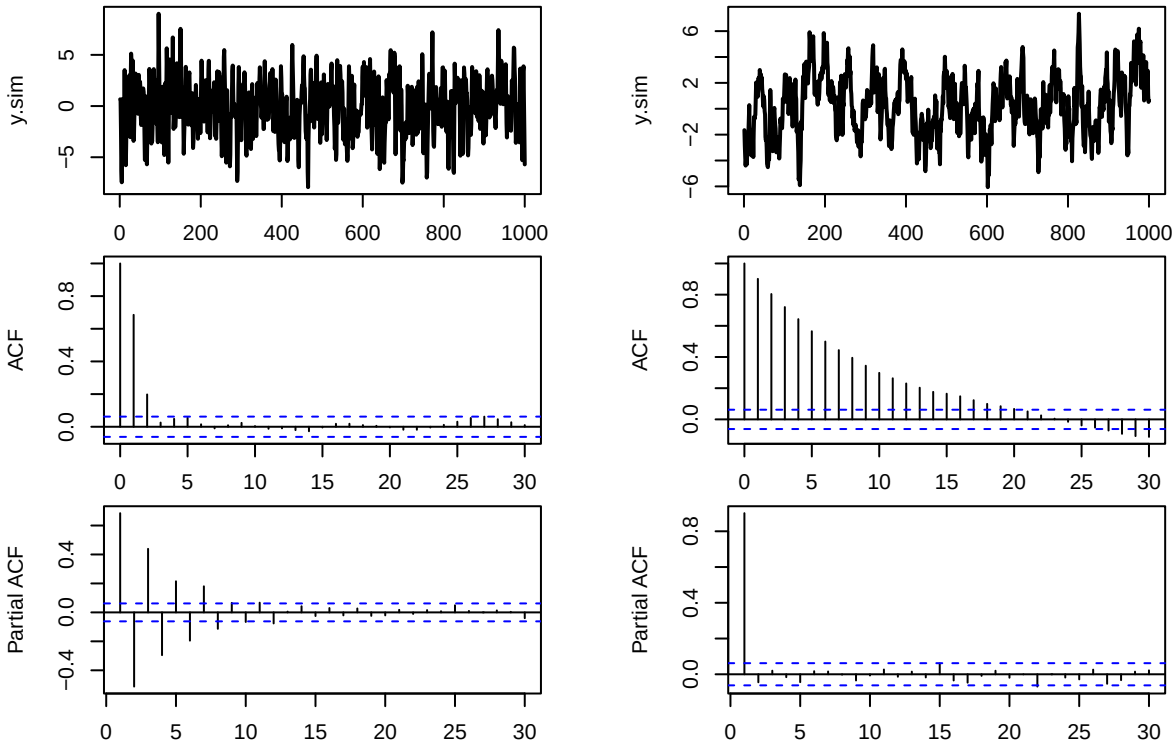


Figure 8.7: ACF/PACF analysis of two processes (MA process on the left, AR on the right).

8.3 Forecasting

Forecasting has always been an important part of the time series field (De Gooijer and Hyndman (2006)). Macroeconomic forecasts are done in many places: Public Administration (notably Treasuries), Central Banks, International Institutions (e.g. IMF, OECD), banks, big firms. These institutions are interested in the **point estimates** ($\sim$ most likely value) of the variable of interest. They also sometimes need to measure the **uncertainty** ($\sim$ dispersion of likely outcomes) associated to the point estimates.[2]

Forecasts produced by professional forecasters are available on these web pages:

- Philly Fed Survey of Professional Forecasters.

[2]In its inflation report, the Bank of England displays charts showing the conditional distribution of future inflation, called fan charts. This fan charts show the uncertainty associated with future inflation. See this page.

- ECB Survey of Professional Forecasters.
- IMF World Economic Outlook.
- OECD Global Economic Outlook.
- European Commission Economic Forecasts.

How to formalize the forecasting problem? Assume the current date is t. We want to forecast the value that variable y_t will take on date $t+1$ (i.e., y_{t+1}) based on the observation of a set of variables gathered in vector x_t (x_t may contain lagged values of y_t).

The forecaster aims at minimizing (a function of) the forecast error. It is usal to consider the following (quadratic) loss function:

$$\underbrace{\mathbb{E}([y_{t+1} - y_{t+1}^*]^2)}_{\text{Mean square error (MSE)}}$$

where y_{t+1}^* is the forecast of y_{t+1} (function of x_t).

Proposition 8.7 (Smallest MSE)**.** *The smallest MSE is obtained with MSEthe expectation of y_{t+1} conditional on x_t.*

Proof. See Appendix 9.5. □

Proposition 8.8. *Among the class of linear forecasts, the smallest MSE is obtained with the linear projection of y_{t+1} on x_t. This projection, denoted by $\widehat{P}(y_{t+1}|x_t) := \alpha' x_t$, satisfies:*

$$\mathbb{E}\left([y_{t+1} - \alpha' x_t]x_t\right) = \mathbf{0}. \tag{8.16}$$

Proof. Consider the function $f : \alpha \to \mathbb{E}\left([y_{t+1} - \alpha' x_t]^2\right)$. We have:

$$f(\alpha) = \mathbb{E}\left(y_{t+1}^2 - 2y_t x_t'\alpha + \alpha' x_t x_t'\alpha]\right).$$

We have $\partial f(\alpha)/\partial\alpha = \mathbb{E}(-2y_{t+1}x_t + 2x_t x_t'\alpha)$. The function is minimised for $\partial f(\alpha)/\partial\alpha = 0$. □

Eq. (8.16) implies that $\mathbb{E}\left(y_{t+1}x_t\right) = \mathbb{E}\left(x_t x_t'\right)\alpha$. (Note that $x_t x_t'\alpha = x_t(x_t'\alpha) = (\alpha' x_t)x_t'$.)
Hence, if $\mathbb{E}\left(x_t x_t'\right)$ is nonsingular,

$$\alpha = \left[\mathbb{E}\left(x_t x_t'\right)\right]^{-1}\mathbb{E}\left(y_{t+1}x_t\right). \tag{8.17}$$

The MSE then is:

$$\mathbb{E}([y_{t+1} - \alpha' x_t]^2) = \mathbb{E}(y_{t+1}^2) - \mathbb{E}\left(y_{t+1}x_t'\right)\left[\mathbb{E}\left(x_t x_t'\right)\right]^{-1}\mathbb{E}\left(x_t y_{t+1}\right).$$

Consider the regression $y_{t+1} = \beta'\mathbf{x}_t + \varepsilon_{t+1}$. The OLS estimate is:

$$\mathbf{b} = \left[\underbrace{\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_t\mathbf{x}_t'}_{\mathbf{m}_1}\right]^{-1}\left[\underbrace{\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_t' y_{t+1}}_{\mathbf{m}_2}\right].$$

If $\{x_t, y_t\}$ is covariance-stationary and ergodic for the second moments then the sample moments ($\mathbf{m}_1$ and $\mathbf{m}_2$) converges in probability to the associated population moments and $\mathbf{b} \xrightarrow{p} \alpha$ (where α is defined in Eq. (8.17)).

Example 8.3 (Forecasting an MA(q) process)**.** Consider the MA(q) process:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\{\varepsilon_t\}$ is a white noise sequence (Def. 8.1).

We have:[3]

$$\begin{array}{l} \mathbb{E}(y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \dots) = \\ \left\{ \begin{array}{lll} \mu + \theta_h \varepsilon_t + \cdots + \theta_q \varepsilon_{t-q+h} & for & h \in [1, q] \\ \mu & for & h > q \end{array} \right. \end{array}$$

and

$$\begin{array}{l} \mathbb{V}ar(y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \dots) = \mathbb{E}\left([y_{t+h} - \mathbb{E}(y_{t+h}|\varepsilon_t, \varepsilon_{t-1}, \dots)]^2\right) = \\ \left\{ \begin{array}{lll} \sigma^2(1 + \theta_1^2 + \cdots + \theta_{h-1}^2) & for & h \in [1, q] \\ \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2) & for & h > q. \end{array} \right. \end{array}$$

Example 8.4 (Forecasting an AR(p) process)**.** (See this web interface.) Consider the AR(p) process:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a white noise sequence (Def. 8.1).

Using the notation of Eq. (8.9), we have:

$$\mathbf{y}_t - \mu = F(\mathbf{y}_{t-1} - \mu) + \xi_t,$$

with $\mu = [\mu, \dots, \mu]'$ (μ is defined in Eq. (8.14)). Hence:

$$\mathbf{y}_{t+h} - \mu = \xi_{t+h} + F\xi_{t+h-1} + \cdots + F^{h-1}\xi_{t+1} + F^h(\mathbf{y}_t - \mu).$$

Therefore:

$$\begin{array}{rcl} \mathbb{E}(\mathbf{y}_{t+h}|y_t, y_{t-1}, \dots) & = & \mu + F^h(\mathbf{y}_t - \mu) \\ \mathbb{V}ar\left([\mathbf{y}_{t+h} - \mathbb{E}(\mathbf{y}_{t+h}|y_t, y_{t-1}, \dots)]\right) & = & \Sigma + F\Sigma F' + \cdots + F^{h-1}\Sigma(F^{h-1})', \end{array}$$

where:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots \\ 0 & 0 & \\ \vdots & & \ddots \end{bmatrix}.$$

[3]The present reasoning relies on the assumption that the ε_t's are observed. But this is generally not the case in practice, where the ε_t's generally have to be estimated.

Alternative approach: Taking the (conditional) expectations of both sides of

$$y_{t+h} - \mu = \phi_1(y_{t+h-1} - \mu) + \phi_2(y_{t+h-2} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \varepsilon_{t+h},$$

we obtain:

$$\begin{aligned} \mathbb{E}(y_{t+h}|y_t, y_{t-1}, \dots) &= \mu + \phi_1 \left(\mathbb{E}[y_{t+h-1}|y_t, y_{t-1}, \dots] - \mu\right) + \\ &\quad \phi_2 \left(\mathbb{E}[y_{t+h-2}|y_t, y_{t-1}, \dots] - \mu\right) + \cdots + \\ &\quad \phi_p \left(\mathbb{E}[y_{t+h-p}|y_t, y_{t-1}, \dots] - \mu\right), \end{aligned}$$

which can be exploited recursively.

The recursion begins with $\mathbb{E}(y_{t-k}|y_t, y_{t-1}, \dots) = y_{t-k}$ (for any $k \geq 0$).

Assessing the performances of a forecasting model

Once one has fitted a model on a given dataset (of length T, say), one compute MSE (mean square errors) to evaluate the performance of the model. But this MSE is the **in-sample** one. It is easy to reduce in-sample MSE. Typically, if the model is estimated by OLS, adding covariates mechanically reduces the MSE (see Props. 4.4 and 4.5). That is, even if additional data are irrelevant, the R^2 of the regression increases. Adding irrelevant variables increases the (in-sample) R^2 but is bound to increase the **out-of-sample** MSE.

Therefore, it is important to analyse **out-of-sample** performances of the forecasting model:

a. Estimate a model on a sample of reduced size $(1, \dots, T^*$, with $T^* < T)$
b. Use the remaining available periods $(T^* + 1, \dots, T)$ to compute **out-of-sample** forecasting errors (and compute their MSE). In an out-of-sample exercise, it is important to make sure that the data used to produce a forecasts (as of date T^*) where indeed available on date T^*.

Diebold-Mariano test

How to compare different forecasting approaches? Diebold and Mariano (1995) have proposed a simple test to address this question.

Assume that you want to compare approaches A and B. You have historical data sets and you have implemented both approaches in the past, providing you with two sets of forecasting errors: $\{e_t^A\}_{t=1,\dots,T}$ and $\{e_t^B\}_{t=1,\dots,T}$.

It may be the case that your forecasts serve a specific purpose and that, for instance, you dislike positive forecasting errors and you care less about negative errors. We assume you are able to formalise this by means of a **loss function** $L(e)$. For instance:

- If you dislike large positive errors, you may set $L(e) = \exp(e)$.
- If you are concerned about both positive and negative errors (indifferently), you may set $L(e) = e^2$ (standard approach).

Let us define the sequence $\{d_t\}_{t=1,...,T} \equiv \{L(e_t^A) - L(e_t^B)\}_{t=1,...,T}$ and assume that this sequence is covariance stationary. We consider the following null hypothesis: $H_0 : \bar{d} = 0$, where $\bar{d}$ denotes the population mean of the d_ts. Under H_0 and under the assumption of covariance-stationarity of d_t, we have (Theorem @ref{(hm:CLTcovstat)):

$$\sqrt{T}\bar{d}_T \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=-\infty}^{+\infty} \gamma_j\right),$$

where the γ_js are the autocovariances of d_t.

Hence, assuming that $\hat{\sigma}^2$ is a consistent estimate of $\sum_{j=-\infty}^{+\infty} \gamma_j$ (for instance the one given by Newey and West (1987)), we have, under H_0:

$$DM_T := \sqrt{T}\frac{\bar{d}_T}{\sqrt{\hat{\sigma}^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

DM_T is the test statistics. For a test of size α, the critical region is:[4]

$$]-\infty, -\Phi^{-1}(1-\alpha/2)] \cup [\Phi^{-1}(1-\alpha/2), +\infty[,$$

where Φ is the c.d.f. of the standard normal distribution.

Example 8.5 (Forecasting Swiss GDP growth)**.** We use a long historical time series of the Swiss GDP growth taken from the Jordà et al. (2017) dataset.[5]

We want to forecast this GDP growth. We envision two specifications : an AR(1) specification (the one advocated by the AIC criteria), and an ARMA(2,2) specification. We are interested in 2-year-ahead forecasts (i.e., $h = 2$ since the data are yearly).

```
library(AEC)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
data <- subset(JST,iso=="CHE")
T <- dim(data)[1]
y <- c(NaN,log(data$gdp[2:T]/data$gdp[1:(T-1)]))
first.date <- T-50
e1 <- NULL; e2 <- NULL;h<-2
for(T.star in first.date:(T-h)){
```

[4]This ShinyApp application illustrates the notion of statistical test (illustrating the p-value and the cirtical region, in particular).

[5]See this website.

```
  estim.model.1 <- arima(y[1:T.star],order=c(1,0,0))
  estim.model.2 <- arima(y[1:T.star],order=c(2,0,2))
  e1 <- c(e1,y[T.star+h] - predict(estim.model.1,n.ahead=h)$pred[h])
  e2 <- c(e2,y[T.star+h] - predict(estim.model.2,n.ahead=h)$pred[h])
}
res.DM <- dm.test(e1,e2,h = h,alternative = "greater")
res.DM
```

```
##
##  Diebold-Mariano Test
##
## data:  e1e2
## DM = -0.82989, Forecast horizon = 2, Loss function power = 2, p-value =
## 0.7946
## alternative hypothesis: greater
```

With `alternative = "greater"` The alternative hypothesis is that method 2 is more accurate than method 1. Since we do not reject the null (the p-value being of 0.795), we are not led to use the more sophisticated model (ARMA(2,2)) and we keep the simple AR(1) model.

Assume now that we want to compare the AR(1) process to a multivariate (VAR) model. We consider a bivariate VAR, where GDP growth is complemented with CPI-based inflation rate.

```
library(vars)
infl <- c(NaN,log(data$cpi[2:T]/data$cpi[1:(T-1)]))
y_var <- cbind(y,infl)
e3 <- NULL
for(T.star in first.date:(T-h)){
  estim.model.3 <- VAR(y_var[2:T.star,],p=1)
  e3 <- c(e3,y[T.star+h] - predict(estim.model.3,n.ahead=h)$fcst$y[h,1])
}
res.DM <- dm.test(e1,e2,h = h,alternative = "greater")
res.DM
```

```
##
##  Diebold-Mariano Test
##
## data:  e1e2
## DM = -0.82989, Forecast horizon = 2, Loss function power = 2, p-value =
## 0.7946
## alternative hypothesis: greater
```

Again, we do not find that the alternative model (here the VAR(1) model) is better than the AR(1) model to forecast GDP growth.

Chapter 9

Appendix

9.1 Principal component analysis (PCA)

Principal component analysis (PCA) is a classical and easy-to-use statistical method to reduce the dimension of large datasets containing variables that are linearly driven by a relatively small number of factors. This approach is widely used in data analysis and image compression.

Suppose that we have T observations of a n-dimensional random vector x, denoted by $x_1, x_2, ..., x_T$. We suppose that each component of x is of mean zero.

Let denote with X the matrix given by $\begin{bmatrix} x_1 & x_2 & ... & x_T \end{bmatrix}'$. Denote the j^{th} column of X by X_j.

We want to find the linear combination of the x_i's $(x.u)$, with $\|u\| = 1$, with "maximum variance." That is, we want to solve:

$$\begin{array}{ll} \arg\max\limits_{u} & u'X'Xu. \\ \text{s.t.} & \|u\| = 1 \end{array} \tag{9.1}$$

Since $X'X$ is a positive definite matrix, it admits the following decomposition:

$$\begin{aligned} X'X &= PDP' \\ &= P \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} P', \end{aligned}$$

where P is an orthogonal matrix whose columns are the eigenvectors of $X'X$.

We can order the eigenvalues such that $\lambda_1 \geq ... \geq \lambda_n$. (Since $X'X$ is positive definite, all these eigenvalues are positive.)

Since P is orthogonal, we have $u'X'Xu = u'PDP'u = y'Dy$ where $\|y\| = 1$. Therefore, we have $y_i^2 \leq 1$ for any $i \leq n$.

As a consequence:

$$y'Dy = \sum_{i=1}^{n} y_i^2 \lambda_i \leq \lambda_1 \sum_{i=1}^{n} y_i^2 = \lambda_1.$$

It is easily seen that the maximum is reached for $y = [1, 0, \cdots, 0]'$. Therefore, the maximum of the optimization program (Eq. (9.1)) is obtained for $u = P[1, 0, \cdots, 0]'$. That is, u is the eigenvector of $X'X$ that is associated with its larger eigenvalue (first column of P).

Let us denote with F the vector that is given by the matrix product XP. The columns of F, denoted by F_j, are called **factors**. We have:

$$F'F = P'X'XP = D.$$

Therefore, in particular, the F_j's are orthogonal.

Since $X = FP'$, the X_j's are linear combinations of the factors. Let us then denote with $\hat{X}_{i,j}$ the part of X_i that is explained by factor F_j, we have:

$$\begin{aligned} \hat{X}_{i,j} &= p_{ij} F_j \\ X_i &= \sum_j \hat{X}_{i,j} = \sum_j p_{ij} F_j. \end{aligned}$$

Consider the share of variance that is explained—through the n variables $(X_1, ..., X_n)$—by the first factor F_1:

$$\frac{\sum_i \hat{X}'_{i,1} \hat{X}_{i,1}}{\sum_i X'_i X_i} = \frac{\sum_i p_{i1} F'_1 F_1 p_{i1}}{tr(X'X)} = \frac{\sum_i p_{i1}^2 \lambda_1}{tr(X'X)} = \frac{\lambda_1}{\sum_i \lambda_i}.$$

Intuitively, if the first eigenvalue is large, it means that the first factor captures a large share of the fluctutaions of the n X_i's.

By the same token, it is easily seen that the fraction of the variance of the n variables that is explained by factor j is given by:

$$\frac{\sum_i \hat{X}'_{i,j} \hat{X}_{i,j}}{\sum_i X'_i X_i} = \frac{\lambda_j}{\sum_i \lambda_i}.$$

Let us illustrate PCA on the term structure of yields. The term strucutre of yields (or yield curve) is know to be driven by only a small number of factors (e.g., Litterman and Scheinkman (1991)). One can typically employ PCA to recover such factors. The data used in the example below are taken from the Fred database (tickers: "DGS6MO","DGS1", ...). The second plot shows the factor loardings, that indicate that the first factor is a level factor (loadings = black line), the second factor is a slope factor (loadings = blue line), the third factor is a curvature factor (loadings = red line).

To run a PCA, one simply has to apply function `prcomp` to a matrix of data:

```
library(AEC)
USyields <- USyields[complete.cases(USyields),]
yds <- USyields[c("Y1","Y2","Y3","Y5","Y7","Y10","Y20","Y30")]
PCA.yds <- prcomp(yds,center=TRUE,scale. = TRUE)
```

Let us know visualize some results. The first plot of Figure 9.1 shows the share of total variance explained by the different principal components (PCs). The second plot shows the facotr loadings. The two bottom plots show how yields (in black) are fitted by linear combinations of the first two PCs only.

```
par(mfrow=c(2,2))
par(plt=c(.1,.95,.2,.8))
barplot(PCA.yds$sdev^2/sum(PCA.yds$sdev^2),
        main="Share of variance expl. by PC's")
axis(1, at=1:dim(yds)[2], labels=colnames(PCA.yds$x))
nb.PC <- 2
plot(-PCA.yds$rotation[,1],type="l",lwd=2,ylim=c(-1,1),
     main="Factor loadings (1st 3 PCs)",xaxt="n",xlab="")
axis(1, at=1:dim(yds)[2], labels=colnames(yds))
lines(PCA.yds$rotation[,2],type="l",lwd=2,col="blue")
lines(PCA.yds$rotation[,3],type="l",lwd=2,col="red")
Y1.hat <- PCA.yds$x[,1:nb.PC] %*% PCA.yds$rotation["Y1",1:2]
Y1.hat <- mean(USyields$Y1) + sd(USyields$Y1) * Y1.hat
plot(USyields$date,USyields$Y1,type="l",lwd=2,
     main="Fit of 1-year yields (2 PCs)",
     ylab="Obs (black) / Fitted by 2PCs (dashed blue)")
lines(USyields$date,Y1.hat,col="blue",lty=2,lwd=2)
Y10.hat <- PCA.yds$x[,1:nb.PC] %*% PCA.yds$rotation["Y10",1:2]
Y10.hat <- mean(USyields$Y10) + sd(USyields$Y10) * Y10.hat
plot(USyields$date,USyields$Y10,type="l",lwd=2,
     main="Fit of 10-year yields (2 PCs)",
     ylab="Obs (black) / Fitted by 2PCs (dashed blue)")
lines(USyields$date,Y10.hat,col="blue",lty=2,lwd=2)
```

9.2 Linear algebra: definitions and results

Definition 9.1 (Eigenvalues)**.** The eigenvalues of of a matrix M are the numbers λ for which:

$$|M - \lambda I| = 0,$$

where $|\bullet|$ is the determinant operator.

Proposition 9.1 (Properties of the determinant)**.** *We have:*

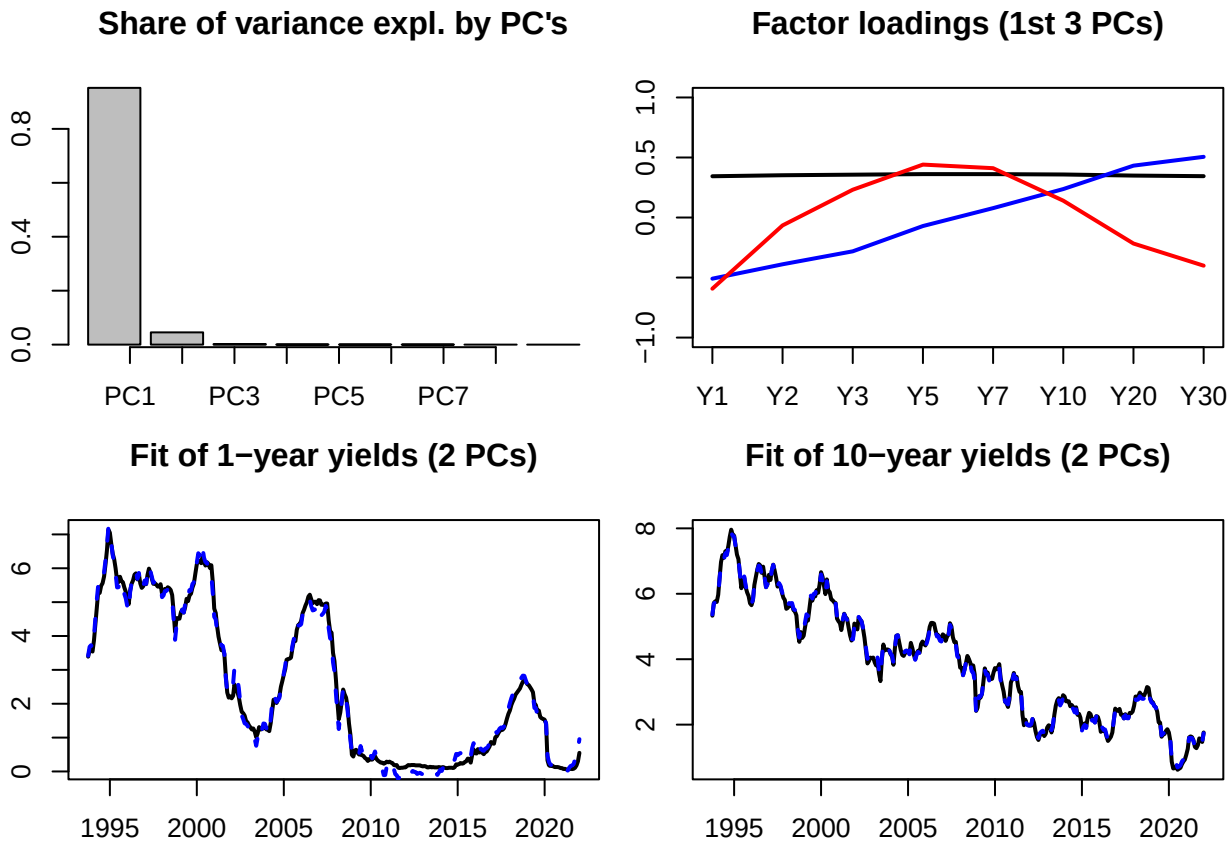


Figure 9.1: Some PCA results. The dataset contains 8 time series of U.S. interest rates of different maturities.

- $|MN| = |M| \times |N|$.
- $|M^{-1}| = |M|^{-1}$.
- *If M admits the diagonal representation $M = TDT^{-1}$, where D is a diagonal matrix whose diagonal entries are $\{\lambda_i\}_{i=1,\ldots,n}$, then:*

$$|M - \lambda I| = \prod_{i=1}^{n} (\lambda_i - \lambda).$$

Definition 9.2 (Moore-Penrose inverse)**.** If $M \in \mathbb{R}^{m \times n}$, then its Moore-Penrose pseudo inverse (exists and) is the unique matrix $M^* \in \mathbb{R}^{n \times m}$ that satisfies:

i. $MM^*M = M$
ii. $M^*MM^* = M^*$
iii. $(MM^*)' = MM^*$ .iv $(M^*M)' = M^*M$.

Proposition 9.2 (Properties of the Moore-Penrose inverse)**.**

- *If M is invertible then $M^* = M^{-1}$.*
- *The pseudo-inverse of a zero matrix is its transpose. **
- **
  The pseudo-inverse of the pseudo-inverse is the original matrix.

Definition 9.3 (Idempotent matrix)**.** Matrix M is idempotent if $M^2 = M$.

If M is a symmetric idempotent matrix, then $M'M = M$.

Proposition 9.3 (Roots of an idempotent matrix)**.** *The eigenvalues of an idempotent matrix are either 1 or 0.*

Proof. If λ is an eigenvalue of an idempotent matrix M then $\exists x \neq 0$ s.t. $Mx = \lambda x$. Hence $M^2x = \lambda Mx \Rightarrow (1 - \lambda)Mx = 0$. Either all element of Mx are zero, in which case $\lambda = 0$ or at least one element of Mx is nonzero, in which case $\lambda = 1$. □

Proposition 9.4 (Idempotent matrix and chi-square distribution)**.** *The rank of a symmetric idempotent matrix is equal to its trace.*

Proof. The result follows from Prop. 9.3, combined with the fact that the rank of a symmetric matrix is equal to the number of its nonzero eigenvalues. □

Proposition 9.5 (Constrained least squares)**.** *The solution of the following optimisation problem:*

$$\begin{aligned} \min_{\beta} \quad & ||\mathbf{y} - \mathbf{X}\beta||^2 \\ & \textit{subject to } \mathbf{R}\beta = \mathbf{q} \end{aligned}$$

is given by:

$$\boxed{\beta^r = \beta_0 - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\beta_0 - \mathbf{q}),}$$

where $\beta_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Proof. See for instance Jackman, 2007. □

Proposition 9.6 (Inverse of a partitioned matrix)**.** *We have:*

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix}.$$

Definition 9.4 (Matrix derivatives)**.** Consider a fonction $f : \mathbb{R}^K \to \mathbb{R}$. Its first-order derivative is:

$$\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = \begin{bmatrix} \frac{\partial f}{\partial b_1}(\mathbf{b}) \\ \vdots \\ \frac{\partial f}{\partial b_K}(\mathbf{b}) \end{bmatrix}.$$

We use the notation:

$$\frac{\partial f}{\partial \mathbf{b}'}(\mathbf{b}) = \left(\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b})\right)'.$$

Proposition 9.7. *We have:*

- *If $f(\mathbf{b}) = A'\mathbf{b}$ where A is a $K \times 1$ vector then $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = A$.*
- *If $f(\mathbf{b}) = \mathbf{b}'A\mathbf{b}$ where A is a $K \times K$ matrix, then $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = 2A\mathbf{b}$.*

Proposition 9.8 (Square and absolute summability)**.** *We have:*

$$\underbrace{\sum_{i=0}^{\infty} |\theta_i| < +\infty}_{\textit{Absolute summability}} \quad \Rightarrow \quad \underbrace{\sum_{i=0}^{\infty} \theta_i^2 < +\infty}_{\textit{Square summability}}.$$

Proof. See Appendix 3.A in Hamilton. Idea: Absolute summability implies that there exist N such that, for $j > N$, $|\theta_j| < 1$ (deduced from Cauchy criterion, Theorem 9.2 and therefore $\theta_j^2 < |\theta_j|$. □

9.3 Statistical analysis: definitions and results

9.3.1 Moments and statistics

Definition 9.5 (Partial correlation)**.** The **partial correlation** between y and z, controlling for some variables $\mathbf{X}$ is the sample correlation between y^* and z^*, where the latter two variables are the residuals in regressions of y on $\mathbf{X}$ and of z on $\mathbf{X}$, respectively.

This correlation is denoted by $r_{yz}^{\mathbf{X}}$. By definition, we have:

$$r_{yz}^{\mathbf{X}} = \frac{\mathbf{z}^{*\prime}\mathbf{y}^*}{\sqrt{(\mathbf{z}^{*\prime}\mathbf{z}^*)(\mathbf{y}^{*\prime}\mathbf{y}^*)}}. \tag{9.2}$$

Definition 9.6 (Skewness and kurtosis)**.** Let Y be a random variable whose fourth moment exists. The expectation of Y is denoted by μ.

- The skewness of Y is given by:
$$\frac{\mathbb{E}[(Y-\mu)^3]}{\{\mathbb{E}[(Y-\mu)^2]\}^{3/2}}.$$
- The kurtosis of Y is given by:
$$\frac{\mathbb{E}[(Y-\mu)^4]}{\{\mathbb{E}[(Y-\mu)^2]\}^{2}}.$$

Theorem 9.1 (Cauchy-Schwarz inequality)**.** *We have:*

$$|\mathbb{C}ov(X,Y)| \leq \sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}$$

and, if $X \neq= $ and $Y \neq 0$, the equality holds iff X and Y are the same up to an affine transformation.

Proof. If $\mathbb{V}ar(X) = 0$, this is trivial. If this is not the case, then let's define Z as $Z = Y - \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$. It is easily seen that $\mathbb{C}ov(X,Z) = 0$. Then, the variance of $Y = Z + \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$ is equal to the sum of the variance of Z and of the variance of $\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$, that is:

$$\mathbb{V}ar(Y) = \mathbb{V}ar(Z) + \left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}\right)^2 \mathbb{V}ar(X) \geq \left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}\right)^2 \mathbb{V}ar(X).$$

The equality holds iff $\mathbb{V}ar(Z) = 0$, i.e. iff $Y = \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X + cst$. □

Definition 9.7 (Asymptotic level)**.** An asymptotic test with critical region Ω_n has an asymptotic level equal to α if:

$$\sup_{\theta\in\Theta} \lim_{n\to\infty} \mathbb{P}_\theta(S_n \in \Omega_n) = \alpha,$$

where S_n is the test statistic and Θ is such that the null hypothesis H_0 is equivalent to $\theta \in \Theta$.

Definition 9.8 (Asymptotically consistent test)**.** An asymptotic test with critical region Ω_n is consistent if:

$$\forall \theta \in \Theta^c, \quad \mathbb{P}_\theta(S_n \in \Omega_n) \to 1,$$

where S_n is the test statistic and Θ^c is such that the null hypothesis H_0 is equivalent to $\theta \notin \Theta^c$.

Definition 9.9 (Kullback discrepancy)**.** Given two p.d.f. f and f^*, the Kullback discrepancy is defined by:

$$I(f,f^*) = \mathbb{E}^*\left(\log\frac{f^*(Y)}{f(Y)}\right) = \int \log\frac{f^*(y)}{f(y)} f^*(y)dy.$$

Proposition 9.9 (Properties of the Kullback discrepancy)**.** *We have:*

i. $I(f, f^*) \geq 0$
ii. $I(f, f^*) = 0$ *iff* $f \equiv f^*$.

Proof. $x \to -\log(x)$ is a convex function. Therefore $\mathbb{E}^*(-\log f(Y)/f^*(Y)) \geq -\log \mathbb{E}^*(f(Y)/f^*(Y)) = 0$ (proves (i)). Since $x \to -\log(x)$ is strictly convex, equality in (i) holds if and only if $f(Y)/f^*(Y)$ is constant (proves (ii)). □

Definition 9.10 (Characteristic function)**.** For any real-valued random variable X, the characteristic function is defined by:

$$\phi_X : u \to \mathbb{E}[\exp(iuX)].$$

9.3.2 Standard distributions

Definition 9.11 (F distribution)**.** Consider $n = n_1 + n_2$ i.i.d. $\mathcal{N}(0,1)$ r.v. X_i. If the r.v. F is defined by:

$$F = \frac{\sum_{i=1}^{n_1} X_i^2}{\sum_{j=n_1+1}^{n_1+n_2} X_j^2} \frac{n_2}{n_1}$$

then $F \sim \mathcal{F}(n_1, n_2)$. (See Table 9.4 for quantiles.)

Definition 9.12 (Student-t distribution)**.** Z follows a Student-t (or t) distribution with ν degrees of freedom (d.f.) if:

$$Z = X_0 \Big/ \sqrt{\frac{\sum_{i=1}^{\nu} X_i^2}{\nu}}, \quad X_i \sim i.i.d. \mathcal{N}(0,1).$$

We have $\mathbb{E}(Z) = 0$, and $\mathbb{V}ar(Z) = \frac{\nu}{\nu-2}$ if $\nu > 2$. (See Table 9.2 for quantiles.)

Definition 9.13 (Chi-square distribution)**.** Z follows a χ^2 distribution with ν d.f. if $Z = \sum_{i=1}^{\nu} X_i^2$ where $X_i \sim i.i.d. \mathcal{N}(0,1)$. We have $\mathbb{E}(Z) = \nu$. (See Table 9.3 for quantiles.)

Definition 9.14 (Cauchy distribution)**.** The probability distribution function of the Cauchy distribution defined by a location parameter μ and a scale parameter γ is:

$$f(x) = \frac{1}{\pi\gamma \left(1 + \left[\frac{x-\mu}{\gamma}\right]^2\right)}.$$

The mean and variance of this distribution are undefined.

Proposition 9.10 (Inner product of a multivariate Gaussian variable)**.** *Let X be a n-dimensional multivariate Gaussian variable: $X \sim \mathcal{N}(0, \Sigma)$. We have:*

$$X'\Sigma^{-1}X \sim \chi^2(n).$$

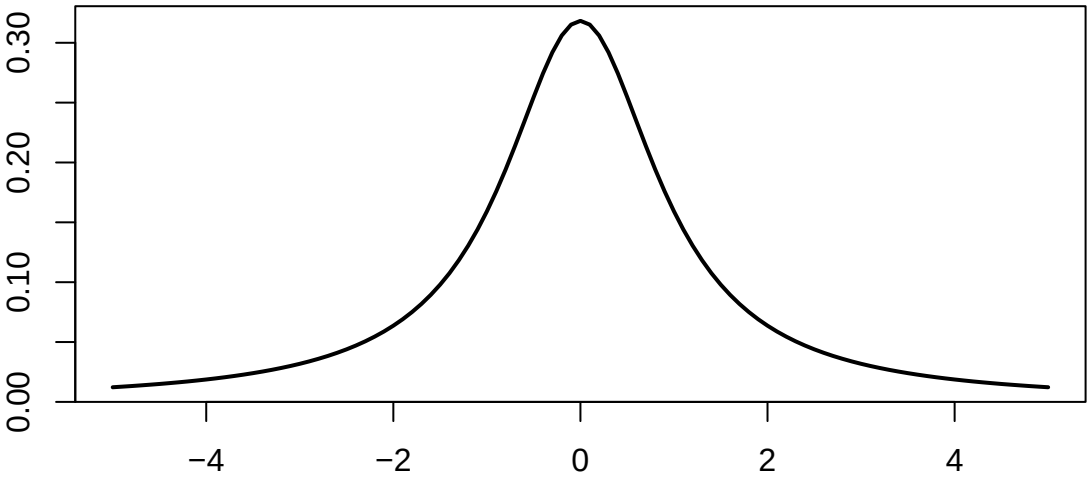


Figure 9.2: Pdf of the Cauchy distribution ($\mu = 0$, $\gamma = 1$).

Proof. Because Σ is a symmetrical definite positive matrix, it admits the spectral decomposition PDP' where P is an orthogonal matrix (i.e. $PP' = Id$) and D is a diagonal matrix with non-negative entries. Denoting by $\sqrt{D^{-1}}$ the diagonal matrix whose diagonal entries are the inverse of those of D, it is easily checked that the covariance matrix of $Y := \sqrt{D^{-1}}P'X$ is Id. Therefore Y is a vector of uncorrelated Gaussian variables. The properties of Gaussian variables imply that the components of Y are then also independent. Hence $Y'Y = \sum_i Y_i^2 \sim \chi^2(n)$.

It remains to note that $Y'Y = X'PD^{-1}P'X = X'\mathbb{V}ar(X)^{-1}X$ to conclude. □

Definition 9.15 (Generalized Extreme Value (GEV) distribution)**.** The vector of disturbances $\varepsilon = [\varepsilon_{1,1}, ..., \varepsilon_{1,K_1}, ..., \varepsilon_{J,1}, ..., \varepsilon_{J,K_J}]'$ follows the Generalized Extreme Value (GEV) distribution if its c.d.f. is:

$$F(\varepsilon, \rho) = \exp(-G(e^{-\varepsilon_{1,1}}, ..., e^{-\varepsilon_{J,K_J}}; \rho))$$

with

$$\begin{aligned} G(\mathbf{Y}; \rho) &\equiv G(Y_{1,1}, ..., Y_{1,K_1}, ..., Y_{J,1}, ..., Y_{J,K_J}; \rho) \\ &= \sum_{j=1}^{J} \left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j} \right)^{\rho_j} \end{aligned}$$

9.3.3 Stochastic convergences

Proposition 9.11 (Chebychev's inequality)**.** *If $\mathbb{E}(|X|^r)$ is finite for some $r > 0$ then:*

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[|X - c|^r]}{\varepsilon^r}.$$

In particular, for $r = 2$:

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[(X - c)^2]}{\varepsilon^2}.$$

Proof. Remark that $\varepsilon^r \mathbb{I}_{\{|X|\geq\varepsilon\}} \leq |X|^r$ and take the expectation of both sides. □

Definition 9.16 (Convergence in probability)**.** The random variable sequence x_n converges in probability to a constant c if $\forall\varepsilon$, $\lim_{n\to\infty} \mathbb{P}(|x_n - c| > \varepsilon) = 0$.

It is denoted as: $\operatorname{plim} x_n = c$.

Definition 9.17 (Convergence in the Lr norm)**.** x_n converges in the r-th mean (or in the L^r-norm) towards x, if $\mathbb{E}(|x_n|^r)$ and $\mathbb{E}(|x|^r)$ exist and if

$$\lim_{n\to\infty} \mathbb{E}(|x_n - x|^r) = 0.$$

It is denoted as: $x_n \xrightarrow{L^r} c$.

For $r = 2$, this convergence is called **mean square convergence**.

Definition 9.18 (Almost sure convergence)**.** The random variable sequence x_n converges almost surely to c if $\mathbb{P}(\lim_{n\to\infty} x_n = c) = 1$.

It is denoted as: $x_n \xrightarrow{a.s.} c$.

Definition 9.19 (Convergence in distribution)**.** x_n is said to converge in distribution (or in law) to x if

$$\lim_{n\to\infty} F_{x_n}(s) = F_x(s)$$

for all s at which F_X –the cumulative distribution of X– is continuous.

It is denoted as: $x_n \xrightarrow{d} x$.

Proposition 9.12 (Rules for limiting distributions (Slutsky))**.** *We have:*

i. ***Slutsky's theorem:*** *If* $x_n \xrightarrow{d} x$ *and* $y_n \xrightarrow{p} c$ *then*

$$\begin{aligned} x_n y_n &\xrightarrow{d} xc \\ x_n + y_n &\xrightarrow{d} x + c \\ x_n / y_n &\xrightarrow{d} x/c \quad (\textit{if } c \neq 0) \end{aligned}$$

ii. ***Continuous mapping theorem:*** *If* $x_n \xrightarrow{d} x$ *and* g *is a continuous function then* $g(x_n) \xrightarrow{d} g(x)$.

Proposition 9.13 (Implications of stochastic convergences)**.** *We have:*

$$\boxed{\xrightarrow{L^s}} \underset{1\leq r\leq s}{\Rightarrow} \boxed{\xrightarrow{L^r}} \Downarrow \boxed{\xrightarrow{a.s.}} \Rightarrow \boxed{\xrightarrow{p}} \Rightarrow \boxed{\xrightarrow{d}}.$$

Proof. (of the fact that $\left(\overset{p}{\to}\right) \Rightarrow \left(\overset{d}{\to}\right)$). Assume that $X_n \overset{p}{\to} X$. Denoting by F and F_n the c.d.f. of X and X_n, respectively:

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_n \leq x, X > x+\varepsilon) \\ &\leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon). \end{aligned}$$

Besides,

$$\begin{aligned} F(x-\varepsilon) &= \mathbb{P}(X \leq x-\varepsilon, X_n \leq x) + \mathbb{P}(X \leq x-\varepsilon, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \varepsilon), \end{aligned}$$

which implies:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x). \tag{9.3}$$

Eqs. (9.3) and (9.3) imply:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x) \leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

Taking limits as $n \to \infty$ yields

$$F(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon).$$

The result is then obtained by taking limits as $\varepsilon \to 0$ (if F is continuous at x). □

Proposition 9.14 (Convergence in distribution to a constant)**.** *If X_n converges in distribution to a constant c, then X_n converges in probability to c.*

Proof. If $\varepsilon > 0$, we have $\mathbb{P}(X_n < c-\varepsilon) \underset{n\to\infty}{\to} 0$ i.e. $\mathbb{P}(X_n \geq c-\varepsilon) \underset{n\to\infty}{\to} 1$ and $\mathbb{P}(X_n < c+\varepsilon) \underset{n\to\infty}{\to} 1$. Therefore $\mathbb{P}(c-\varepsilon \leq X_n < c+\varepsilon) \underset{n\to\infty}{\to} 1$, which gives the result. □

Example 9.1 (Convergence in probability but not L^r)**.** Let $\{x_n\}_{n\in\mathbb{N}}$ be a series of random variables defined by:

$$x_n = nu_n,$$

where u_n are independent random variables s.t. $u_n \sim \mathcal{B}(1/n)$.

We have $x_n \overset{p}{\to} 0$ but $x_n \overset{L^r}{\nrightarrow} 0$ because $\mathbb{E}(|X_n - 0|) = \mathbb{E}(X_n) = 1$.

Theorem 9.2 (Cauchy criterion (non-stochastic case))**.** *We have that $\sum_{i=0}^{T} a_i$ converges ($T \to \infty$) iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

$$\left| \sum_{i=N+1}^{M} a_i \right| < \eta.$$

Theorem 9.3 (Cauchy criterion (stochastic case))**.** *We have that $\sum_{i=0}^{T} \theta_i \varepsilon_{t-i}$ converges in mean square ($T \to \infty$) to a random variable iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

$$\mathbb{E}\left[\left(\sum_{i=N+1}^{M} \theta_i \varepsilon_{t-i} \right)^2 \right] < \eta.$$

9.3.4 Central limit theorem

Theorem 9.4 (Law of large numbers)**.** *The sample mean is a consistent estimator of the population mean.*

Proof. Let's denote by ϕ_{X_i} the characteristic function of a r.v. X_i. If the mean of X_i is μ then the Talyor expansion of the characteristic function is:

$$\phi_{X_i}(u) = \mathbb{E}(\exp(iuX)) = 1 + iu\mu + o(u).$$

The properties of the characteristic function (see Def. 9.10) imply that:

$$\phi_{\frac{1}{n}(X_1+\cdots+X_n)}(u) = \prod_{i=1}^{n}\left(1 + i\frac{u}{n}\mu + o\left(\frac{u}{n}\right)\right) \to e^{iu\mu}.$$

The facts that (a) $e^{iu\mu}$ is the characteristic function of the constant μ and (b) that a characteristic function uniquely characterises a distribution imply that the sample mean converges in distribution to the constant μ, which further implies that it converges in probability to μ. □

Theorem 9.5 (Lindberg-Levy Central limit theorem, CLT)**.** *If x_n is an i.i.d. sequence of random variables with mean μ and variance σ^2 ($\in]0, +\infty[$), then:*

$$\boxed{\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \textit{where} \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.}$$

Proof. Let us introduce the r.v. $Y_n := \sqrt{n}(\bar{X}_n - \mu)$. We have $\phi_{Y_n}(u) = \left[\mathbb{E}\left(\exp(i\frac{1}{\sqrt{n}}u(X_1 - \mu))\right)\right]^n$. We have:

$$\begin{aligned} & \left[\mathbb{E}\left(\exp\left(i\frac{1}{\sqrt{n}}u(X_1-\mu)\right)\right)\right]^n \\ = & \left[\mathbb{E}\left(1 + i\frac{1}{\sqrt{n}}u(X_1-\mu) - \frac{1}{2n}u^2(X_1-\mu)^2 + o(u^2)\right)\right]^n \\ = & \left(1 - \frac{1}{2n}u^2\sigma^2 + o(u^2)\right)^n. \end{aligned}$$

Therefore $\phi_{Y_n}(u) \underset{n\to\infty}{\to} \exp\left(-\frac{1}{2}u^2\sigma^2\right)$, which is the characteristic function of $\mathcal{N}(0, \sigma^2)$. □

9.4 Some properties of Gaussian variables

Proposition 9.15. *If* $\mathbf{A}$ *is idempotent and if* $\mathbf{x}$ *is Gaussian,* $\mathbf{Lx}$ *and* $\mathbf{x}'\mathbf{Ax}$ *are independent if* $\mathbf{LA} = \mathbf{0}$.

Proof. If $\mathbf{LA} = \mathbf{0}$, then the two Gaussian vectors $\mathbf{Lx}$ and $\mathbf{Ax}$ are independent. This implies the independence of any function of $\mathbf{Lx}$ and any function of $\mathbf{Ax}$. The results then follows from the observation that $\mathbf{x}'\mathbf{Ax} = (\mathbf{Ax})'(\mathbf{Ax})$, which is a function of $\mathbf{Ax}$. □

Proposition 9.16 (Bayesian update in a vector of Gaussian variables)**.** *If*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}\right),$$

then

$$Y_2|Y_1 \sim \mathcal{N}\left(\Omega_{21}\Omega_{11}^{-1}Y_1, \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}\right).$$
$$Y_1|Y_2 \sim \mathcal{N}\left(\Omega_{12}\Omega_{22}^{-1}Y_2, \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right).$$

Proposition 9.17 (Truncated distributions)**.** *If X is a random variable distributed according to some p.d.f. f, with c.d.f. F, with infinite support. Then the p.d.f. of $X|a \leq X < b$ is*

$$g(x) = \frac{f(x)}{F(b) - F(a)}\mathbb{I}_{\{a \leq x < b\}},$$

for any $a < b$.

In partiucular, for a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$f(X = x|a \leq X < b) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{Z}.$$

with $Z = \Phi(\beta) - \Phi(\alpha)$, where $\alpha = \dfrac{a-\mu}{\sigma}$ and $\beta = \dfrac{b-\mu}{\sigma}$.

Moreover:

$$\mathbb{E}(X|a \leq X < b) = \mu - \frac{\phi(\beta) - \phi(\alpha)}{Z}\sigma. \tag{9.4}$$

We also have:

$$\begin{aligned} &\mathbb{V}ar(X|a \leq X < b) \\ = \ &\sigma^2\left[1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{Z} - \left(\frac{\phi(\beta) - \phi(\alpha)}{Z}\right)^2\right] \end{aligned} \tag{9.5}$$

In particular, for $b \to \infty$, we get:

$$\mathbb{V}ar(X|a < X) = \sigma^2\left[1 + \alpha\lambda(-\alpha) - \lambda(-\alpha)^2\right], \tag{9.6}$$

*with $\lambda(x) = \dfrac{\phi(x)}{\Phi(x)}$ is called the **inverse Mills ratio**.*

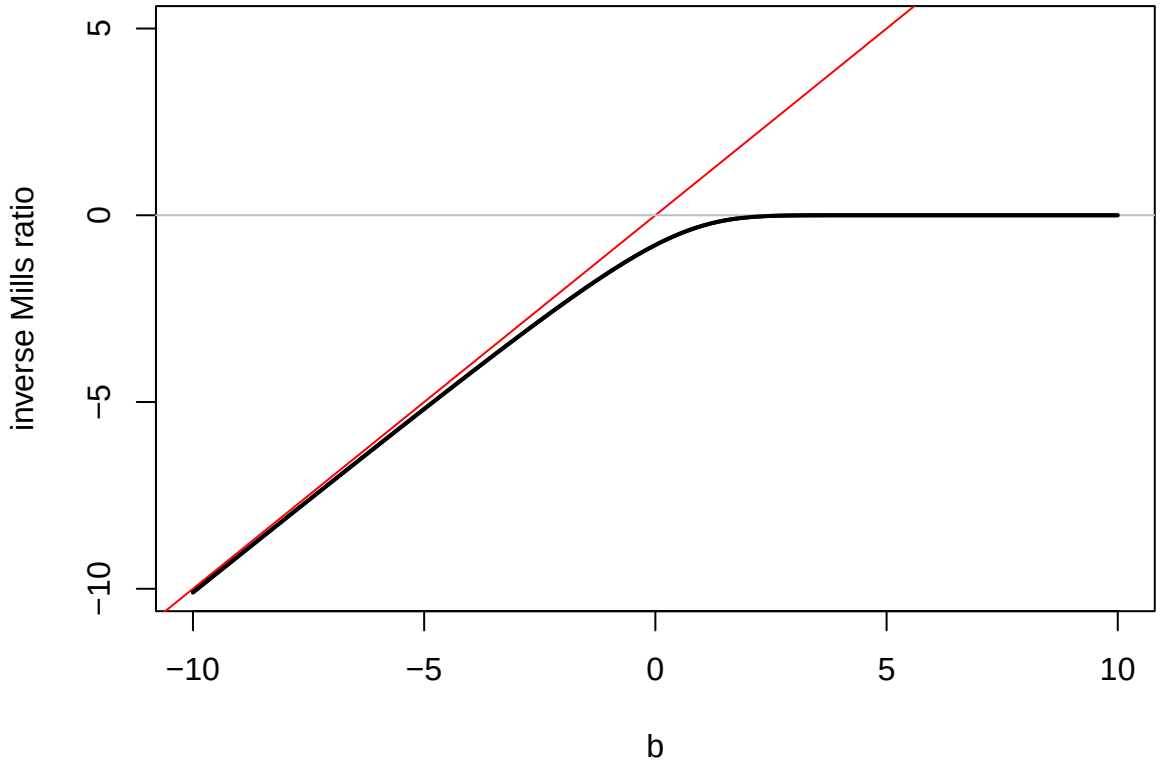


Figure 9.3: $\mathbb{E}(X|X < b)$ as a function of b when $X \sim \mathcal{N}(0,1)$ (in black).

Consider the case where $a \to -\infty$ (i.e. the conditioning set is $X < b$) and $\mu = 0$, $\sigma = 1$. Then Eq. (9.4) gives $\mathbb{E}(X|X < b) = -\lambda(b) = -\dfrac{\phi(b)}{\Phi(b)}$, where λ is the function computing the inverse Mills ratio.

Proposition 9.18 (p.d.f. of a multivariate Gaussian variable)**.** *If $Y \sim \mathcal{N}(\mu, \Omega)$ and if Y is a n-dimensional vector, then the density function of Y is:*

$$\frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left[-\frac{1}{2}(Y-\mu)' \Omega^{-1} (Y-\mu)\right].$$

9.5 Proofs

Proof of Proposition 6.4

Proof. Assumptions (i) and (ii) (in the set of Assumptions 6.1) imply that θ_{MLE} exists $(= \text{argmax}_\theta (1/n) \log \mathcal{L}(\theta; \mathbf{y}))$.

$(1/n) \log \mathcal{L}(\theta; \mathbf{y})$ can be interpreted as the sample mean of the r.v. $\log f(Y_i; \theta)$ that are i.i.d. Therefore $(1/n) \log \mathcal{L}(\theta; \mathbf{y})$ converges to $\mathbb{E}_{\theta_0}(\log f(Y; \theta))$ – which exists (Assumption iv).

Because the latter convergence is uniform (Assumption v), the solution θ_{MLE} almost surely

converges to the solution to the limit problem:

$$\text{argmax}_\theta \mathbb{E}_{\theta_0}(\log f(Y;\theta)) = \text{argmax}_\theta \int_y \log f(y;\theta) f(y;\theta_0) dy.$$

Properties of the Kullback information measure (see Prop. 9.9), together with the identifiability assumption (ii) implies that the solution to the limit problem is unique and equal to θ_0.

Consider a r.v. sequence θ that converges to θ_0. The Taylor expansion of the score in a neighborood of θ_0 yields to:

$$\frac{\partial \log \mathcal{L}(\theta;\mathbf{y})}{\partial \theta} = \frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} + \frac{\partial^2 \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta \partial \theta'}(\theta - \theta_0) + o_p(\theta - \theta_0)$$

θ_{MLE} converges to θ_0 and satisfies the likelihood equation $\frac{\partial \log \mathcal{L}(\theta;\mathbf{y})}{\partial \theta} = \mathbf{0}$. Therefore:

$$\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx -\frac{\partial^2 \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta \partial \theta'}(\theta_{MLE} - \theta_0),$$

or equivalently:

$$\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\theta_0)}{\partial \theta \partial \theta'} \right) \sqrt{n}(\theta_{MLE} - \theta_0),$$

By the law of large numbers, we have: $\left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\theta_0)}{\partial \theta \partial \theta'} \right) \to \frac{1}{n}\mathbf{I}(\theta_0) = \mathcal{I}_Y(\theta_0)$.

Besides, we have:

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} &= \sqrt{n} \left(\frac{1}{n} \sum_i \frac{\partial \log f(y_i;\theta_0)}{\partial \theta} \right) \\ &= \sqrt{n} \left(\frac{1}{n} \sum_i \left\{ \frac{\partial \log f(y_i;\theta_0)}{\partial \theta} - \mathbb{E}_{\theta_0} \frac{\partial \log f(Y_i;\theta_0)}{\partial \theta} \right\} \right) \end{aligned}$$

which converges to $\mathcal{N}(0, \mathcal{I}_Y(\theta_0))$ by the CLT.

Collecting the preceding results leads to (b). The fact that θ_{MLE} achieves the FDCR bound proves (c). □

Proof of Proposition 6.5

Proof. We have $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ (Eq. (6.10)). A Taylor expansion around θ_0 yields to:

$$\sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{9.7}$$

Under H_0, $h(\theta_0) = 0$ therefore:

$$\sqrt{n}h(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{9.8}$$

Hence

$$\sqrt{n}\left(\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial \theta}\right)^{-1/2} h(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}(0, Id).$$

Taking the quadratic form, we obtain:

$$nh(\hat{\theta}_n)'\left(\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial \theta}\right)^{-1} h(\hat{\theta}_n) \xrightarrow{d} \chi^2(r).$$

The fact that the test has asymptotic level α directly stems from what precedes. **Consistency of the test**: Consider $\theta_0 \in \Theta$. Because the MLE is consistent, $h(\hat{\theta}_n)$ converges to $h(\theta_0) \neq 0$. Eq. (9.7) is still valid. It implies that ξ_n^W converges to $+\infty$ and therefore that $\mathbb{P}_\theta(\xi_n^W \geq \chi^2_{1-\alpha}(r)) \to 1$. □

Proof of Proposition 6.6

Proof. Notations: "≈" means "equal up to a term that converges to 0 in probability". We are under H_0. $\hat{\theta}^0$ is the constrained ML estimator; $\hat{\theta}$ denotes the unconstrained one.

We combine the two Taylor expansion: $h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}(\hat{\theta}_n - \theta_0)$ and $h(\hat{\theta}_n^0) \approx \frac{\partial h(\theta_0)}{\partial \theta'}(\hat{\theta}_n^0 - \theta_0)$ and we use $h(\hat{\theta}_n^0) = 0$ (by definition) to get:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{9.9}$$

Besides, we have (using the definition of the information matrix):

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \tag{9.10}$$

and:

$$0 = \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n; \mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{9.11}$$

Taking the difference and multiplying by $\mathcal{I}(\theta_0)^{-1}$:

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \approx \mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta}\mathcal{I}(\theta_0). \tag{9.12}$$

Eqs. (9.9) and (9.12) yield to:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta}. \tag{9.13}$$

Recall that $\hat{\theta}_n^0$ is the MLE of θ_0 under the constraint $h(\theta) = 0$. The vector of Lagrange multipliers $\hat{\lambda}_n$ associated to this program satisfies:

$$\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} + \frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \hat{\lambda}_n = 0. \tag{9.14}$$

Substituting the latter equation in Eq. (9.13) gives:

$$\begin{aligned} \sqrt{n} h(\hat{\theta}_n) &\approx -\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \frac{\hat{\lambda}_n}{\sqrt{n}} \\ &\approx -\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h'(\theta_0; \mathbf{y})}{\partial \theta} \frac{\hat{\lambda}_n}{\sqrt{n}}, \end{aligned}$$

which yields:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \approx -\left(\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h'(\theta_0; \mathbf{y})}{\partial \theta} \right)^{-1} \sqrt{n} h(\hat{\theta}_n). \tag{9.15}$$

It follows, from Eq. (9.8), that:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h'(\theta_0; \mathbf{y})}{\partial \theta} \right)^{-1} \right).$$

Taking the quadratic form of the last equation gives:

$$\frac{1}{n} \hat{\lambda}_n' \frac{\partial h(\hat{\theta}_n^0)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n^0)^{-1} \frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \hat{\lambda}_n \xrightarrow{d} \chi^2(r).$$

Using Eq. (9.14), it appears that the left-hand side term of the last equation is ξ^{LM} as defined in Eq. (6.18). Consistency: see Remark 17.3 in Gouriéroux and Monfort (1995). □

Proof of Proposition 6.7

Proof. Let us first demonstrate the asymptotic equivalence of ξ^{LM} and ξ^{LR}.

The second-order Taylor expansions of $\log \mathcal{L}(\hat{\theta}_n^0, \mathbf{y})$ and $\log \mathcal{L}(\hat{\theta}_n, \mathbf{y})$ are:

$$\begin{aligned} \log \mathcal{L}(\hat{\theta}_n, \mathbf{y}) &\approx \log \mathcal{L}(\theta_0, \mathbf{y}) + \frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'} (\hat{\theta}_n - \theta_0) \\ &\quad -\frac{n}{2} (\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0) (\hat{\theta}_n - \theta_0) \\ \log \mathcal{L}(\hat{\theta}_n^0, \mathbf{y}) &\approx \log \mathcal{L}(\theta_0, \mathbf{y}) + \frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'} (\hat{\theta}_n^0 - \theta_0) \\ &\quad -\frac{n}{2} (\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0) (\hat{\theta}_n^0 - \theta_0). \end{aligned}$$

Taking the difference, we obtain:

$$\begin{aligned}\xi_n^{LR} &\approx 2\frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n - \hat{\theta}_n^0) + n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) \\ &\quad -n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0).\end{aligned}$$

Using $\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$ (Eq. (9.11)), we have:

$$\begin{aligned}\xi_n^{LR} &\approx 2n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) + n(\hat{\theta}_n^0 - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) \\ &\quad -n(\hat{\theta}_n - \theta_0)'\mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0).\end{aligned}$$

In the second of the three terms in the sum, we replace $(\hat{\theta}_n^0 - \theta_0)$ by $(\hat{\theta}_n^0 - \hat{\theta}_n + \hat{\theta}_n - \theta_0)$ and we develop the associated product. This leads to:

$$\xi_n^{LR} \approx n(\hat{\theta}_n^0 - \hat{\theta}_n)'\mathcal{I}(\theta_0)^{-1}(\hat{\theta}_n^0 - \hat{\theta}_n). \tag{9.16}$$

The difference between Eqs. (9.10) and (9.11) implies:

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0),$$

which, associated to Eq. @**??**eq:lr10), gives:

$$\xi_n^{LR} \approx \frac{1}{n}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \approx \xi_n^{LM}.$$

Hence ξ_n^{LR} has the same asymptotic distribution as ξ_n^{LM}.

Let's show that the LR test is consistent. For this, note that:

$$\begin{aligned}\frac{\log \mathcal{L}(\hat{\theta}, \mathbf{y}) - \log \mathcal{L}(\hat{\theta}^0, \mathbf{y})}{n} &= \frac{1}{n}\sum_{i=1}^{n}[\log f(y_i; \hat{\theta}_n) - \log f(y_i; \hat{\theta}_n^0)] \\ &\to \mathbb{E}_0[\log f(Y; \theta_0) - \log f(Y; \theta_\infty)],\end{aligned}$$

where θ_∞, the pseudo true value, is such that $h(\theta_\infty) \neq 0$ (by definition of H_1). From the Kullback inequality and the asymptotic identifiability of θ_0, it follows that $\mathbb{E}_0[\log f(Y; \theta_0) - \log f(Y; \theta_\infty)] > 0$. Therefore $\xi_n^{LR} \to +\infty$ under H_1.

Let us now demonstrate the equivalence of $\xi^{LM} and \xi^{W}$.

We have (using Eq. **??**eq:multiplier)):

$$\xi_n^{LM} = \frac{1}{n}\hat{\lambda}_n'\frac{\partial h(\hat{\theta}_n^0)}{\partial \theta'}\mathcal{I}(\hat{\theta}_n^0)^{-1}\frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta}\hat{\lambda}_n.$$

Since, under H_0, $\hat{\theta}_n^0 \approx \hat{\theta}_n \approx \theta_0$, Eq. (9.15) therefore implies that:

$$\xi^{LM} \approx nh(\hat{\theta}_n)' \left(\frac{\partial h(\hat{\theta}_n)}{\partial \theta'} \mathcal{J}(\hat{\theta}_n)^{-1} \frac{\partial h'(\hat{\theta}_n; \mathbf{y})}{\partial \theta} \right)^{-1} h(\hat{\theta}_n) = \xi^W,$$

which gives the result. □

Proof of Eq. (8.3)

Proof. We have:

$$\begin{aligned}
& T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] \\
= \; & T\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - \mu)\right)^2\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(y_t - \mu)^2 + 2\sum_{s<t\leq T}(y_t - \mu)(y_s - \mu)\right] \\
= \; & \gamma_0 + \frac{2}{T}\left(\sum_{t=2}^{T}\mathbb{E}\left[(y_t - \mu)(y_{t-1} - \mu)\right]\right) + \frac{2}{T}\left(\sum_{t=3}^{T}\mathbb{E}\left[(y_t - \mu)(y_{t-2} - \mu)\right]\right) + ... \\
& + \frac{2}{T}\left(\sum_{t=T-1}^{T}\mathbb{E}\left[(y_t - \mu)(y_{t-(T-2)} - \mu)\right]\right) + \frac{2}{T}\mathbb{E}\left[(y_t - \mu)(y_{t-(T-1)} - \mu)\right] \\
= \; & \gamma_0 + 2\frac{T-1}{T}\gamma_1 + \cdots + 2\frac{1}{T}\gamma_{T-1}.
\end{aligned}$$

Therefore:

$$\begin{aligned}
& T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j \\
= \; & -2\frac{1}{T}\gamma_1 - 2\frac{2}{T}\gamma_2 - \cdots - 2\frac{T-1}{T}\gamma_{T-1} - 2\gamma_T - 2\gamma_{T+1} + ...
\end{aligned}$$

And then:

$$\begin{aligned}
& \left|T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j\right| \\
\leq \; & 2\frac{1}{T}|\gamma_1| + 2\frac{2}{T}|\gamma_2| + \cdots + 2\frac{T-1}{T}|\gamma_{T-1}| + 2|\gamma_T| + 2|\gamma_{T+1}| + ...
\end{aligned}$$

For any $q \leq T$, we have:

$$\begin{aligned}
\left|T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j\right| \; \leq \; & 2\frac{1}{T}|\gamma_1| + 2\frac{2}{T}|\gamma_2| + \cdots + 2\frac{q-1}{T}|\gamma_{q-1}| + 2\frac{q}{T}|\gamma_q| + \\
& 2\frac{q+1}{T}|\gamma_{q+1}| + \cdots + 2\frac{T-1}{T}|\gamma_{T-1}| + 2|\gamma_T| + 2|\gamma_{T+1}| + ... \\
\leq \; & \frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q-1)|\gamma_{q-1}| + q|\gamma_q|\right) + \\
& 2|\gamma_{q+1}| + \cdots + 2|\gamma_{T-1}| + 2|\gamma_T| + 2|\gamma_{T+1}| + ...
\end{aligned}$$

Consider $\varepsilon > 0$. The fact that the autocovariances are absolutely summable implies that there exists q_0 such that (Cauchy criterion, Theorem 9.2):

$$2|\gamma_{q_0+1}| + 2|\gamma_{q_0+2}| + 2|\gamma_{q_0+3}| + \cdots < \varepsilon/2.$$

Then, if $T > q_0$, it comes that:

$$\left| T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty} \gamma_j \right| \leq \frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right) + \varepsilon/2.$$

If $T \geq 2\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right)/(\varepsilon/2)$ $(= f(q_0)$, say) then

$$\frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \cdots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right) \leq \varepsilon/2.$$

Then, if $T > f(q_0)$ and $T > q_0$, i.e. if $T > \max(f(q_0), q_0)$, we have:

$$\left| T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty} \gamma_j \right| \leq \varepsilon.$$

□

Proof of Proposition 8.7

Proof. We have:

$$\begin{aligned}
\mathbb{E}([y_{t+1} - y^*_{t+1}]^2) &= \mathbb{E}\left([\{y_{t+1} - \mathbb{E}(y_{t+1}|x_t)\} + \{\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}\}]^2\right) \\
&= \mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2\right) + \mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]^2\right) \\
&\quad +2\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\right). \qquad (9.17)
\end{aligned}$$

Let us focus on the last term. We have:

$$\begin{aligned}
&\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\right) \\
= \ &\mathbb{E}(\mathbb{E}([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]\underbrace{[\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]}_{\text{function of } x_t}|x_t)) \\
= \ &\mathbb{E}([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\mathbb{E}([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]|x_t)) \\
= \ &\mathbb{E}([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\underbrace{[\mathbb{E}(y_{t+1}|x_t) - \mathbb{E}(y_{t+1}|x_t)]}_{=0}) = 0.
\end{aligned}$$

Therefore, Eq. (9.17) becomes:

$$\begin{aligned}
&\mathbb{E}([y_{t+1} - y^*_{t+1}]^2) \\
= \ &\underbrace{\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2\right)}_{\geq 0 \text{ and does not depend on } y^*_{t+1}} + \underbrace{\mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]^2\right)}_{\geq 0 \text{ and depends on } y^*_{t+1}}.
\end{aligned}$$

This implies that $\mathbb{E}([y_{t+1} - y^*_{t+1}]^2)$ is always larger than $\mathbb{E}([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2)$, and is therefore minimized if the second term is equal to zero, that is if $\mathbb{E}(y_{t+1}|x_t) = y^*_{t+1}$. □

Proof of Proposition ??

Proof. Using Proposition 9.18, we obtain that, conditionally on x_1, the log-likelihood is given by

$$\begin{aligned}\log \mathcal{L}(Y_T;\theta) &= -(Tn/2)\log(2\pi) + (T/2)\log\left|\Omega^{-1}\right| \\ &\quad -\frac{1}{2}\sum_{t=1}^{T}\left[\left(y_t - \Pi' x_t\right)'\Omega^{-1}\left(y_t - \Pi' x_t\right)\right].\end{aligned}$$

Let's rewrite the last term of the log-likelihood:

$$\begin{aligned}\sum_{t=1}^{T}\left[\left(y_t - \Pi' x_t\right)'\Omega^{-1}\left(y_t - \Pi' x_t\right)\right] &= \\ \sum_{t=1}^{T}\left[\left(y_t - \hat{\Pi}' x_t + \hat{\Pi}' x_t - \Pi' x_t\right)'\Omega^{-1}\left(y_t - \hat{\Pi}' x_t + \hat{\Pi}' x_t - \Pi' x_t\right)\right] &= \\ \sum_{t=1}^{T}\left[\left(\hat{\varepsilon}_t + (\hat{\Pi} - \Pi)' x_t\right)'\Omega^{-1}\left(\hat{\varepsilon}_t + (\hat{\Pi} - \Pi)' x_t\right)\right],\end{aligned}$$

where the j^{th} element of the $(n \times 1)$ vector $\hat{\varepsilon}_t$ is the sample residual, for observation t, from an OLS regression of $y_{j,t}$ on x_t. Expanding the previous equation, we get:

$$\begin{aligned}&\sum_{t=1}^{T}\left[\left(y_t - \Pi' x_t\right)'\Omega^{-1}\left(y_t - \Pi' x_t\right)\right] = \sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t \\ &+2\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi} - \Pi)' x_t + \sum_{t=1}^{T} x_t'(\hat{\Pi} - \Pi)\Omega^{-1}(\hat{\Pi} - \Pi)' x_t.\end{aligned}$$

Let's apply the trace operator on the second term (that is a scalar):

$$\begin{aligned}\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi} - \Pi)' x_t &= Tr\left(\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi} - \Pi)' x_t\right) \\ = Tr\left(\sum_{t=1}^{T}\Omega^{-1}(\hat{\Pi} - \Pi)' x_t\hat{\varepsilon}_t'\right) &= Tr\left(\Omega^{-1}(\hat{\Pi} - \Pi)'\sum_{t=1}^{T} x_t\hat{\varepsilon}_t'\right).\end{aligned}$$

Given that, by construction (property of OLS estimates), the sample residuals are orthogonal to the explanatory variables, this term is zero. Introducing $\tilde{x}_t = (\hat{\Pi} - \Pi)' x_t$, we have

$$\sum_{t=1}^{T}\left[\left(y_t - \Pi' x_t\right)'\Omega^{-1}\left(y_t - \Pi' x_t\right)\right] = \sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t + \sum_{t=1}^{T}\tilde{x}_t'\Omega^{-1}\tilde{x}_t.$$

Since Ω is a positive definite matrix, Ω^{-1} is as well. Consequently, the smallest value that the last term can take is obtained for $\tilde{x}_t = 0$, i.e. when $\Pi = \hat{\Pi}$.

The MLE of Ω is the matrix $\hat{\Omega}$ that maximizes $\Omega \overset{\ell}{\to} L(Y_T; \hat{\Pi}, \Omega)$. We have:

$$\log \mathcal{L}(Y_T; \hat{\Pi}, \Omega) \quad = \quad -(Tn/2)\log(2\pi) + (T/2)\log\left|\Omega^{-1}\right| - \frac{1}{2}\sum_{t=1}^{T}\left[\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t\right].$$

Matrix $\hat{\Omega}$ is a symmetric positive definite. It is easily checked that the (unrestricted) matrix that maximizes the latter expression is symmetric positive definite matrix. Indeed:

$$\frac{\partial \log \mathcal{L}(Y_T; \hat{\Pi}, \Omega)}{\partial \Omega} = \frac{T}{2}\Omega' - \frac{1}{2}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t' \Rightarrow \hat{\Omega}' = \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_t',$$

which leads to the result. □

9.6 Statistical Tables

Table 9.1: Quantiles of the $\mathcal{N}(0,1)$ distribution. If a and b are respectively the row and column number; then the corresponding cell gives $\mathbb{P}(0 < X \leq a+b)$, where $X \sim \mathcal{N}(0,1)$.

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.6179	0.7257	0.8159	0.8849	0.9332	0.9641	0.9821	0.9918	0.9965
0.1	0.5040	0.6217	0.7291	0.8186	0.8869	0.9345	0.9649	0.9826	0.9920	0.9966
0.2	0.5080	0.6255	0.7324	0.8212	0.8888	0.9357	0.9656	0.9830	0.9922	0.9967
0.3	0.5120	0.6293	0.7357	0.8238	0.8907	0.9370	0.9664	0.9834	0.9925	0.9968
0.4	0.5160	0.6331	0.7389	0.8264	0.8925	0.9382	0.9671	0.9838	0.9927	0.9969
0.5	0.5199	0.6368	0.7422	0.8289	0.8944	0.9394	0.9678	0.9842	0.9929	0.9970
0.6	0.5239	0.6406	0.7454	0.8315	0.8962	0.9406	0.9686	0.9846	0.9931	0.9971
0.7	0.5279	0.6443	0.7486	0.8340	0.8980	0.9418	0.9693	0.9850	0.9932	0.9972
0.8	0.5319	0.6480	0.7517	0.8365	0.8997	0.9429	0.9699	0.9854	0.9934	0.9973
0.9	0.5359	0.6517	0.7549	0.8389	0.9015	0.9441	0.9706	0.9857	0.9936	0.9974
1	0.5398	0.6554	0.7580	0.8413	0.9032	0.9452	0.9713	0.9861	0.9938	0.9974
1.1	0.5438	0.6591	0.7611	0.8438	0.9049	0.9463	0.9719	0.9864	0.9940	0.9975
1.2	0.5478	0.6628	0.7642	0.8461	0.9066	0.9474	0.9726	0.9868	0.9941	0.9976
1.3	0.5517	0.6664	0.7673	0.8485	0.9082	0.9484	0.9732	0.9871	0.9943	0.9977
1.4	0.5557	0.6700	0.7704	0.8508	0.9099	0.9495	0.9738	0.9875	0.9945	0.9977
1.5	0.5596	0.6736	0.7734	0.8531	0.9115	0.9505	0.9744	0.9878	0.9946	0.9978
1.6	0.5636	0.6772	0.7764	0.8554	0.9131	0.9515	0.9750	0.9881	0.9948	0.9979
1.7	0.5675	0.6808	0.7794	0.8577	0.9147	0.9525	0.9756	0.9884	0.9949	0.9979
1.8	0.5714	0.6844	0.7823	0.8599	0.9162	0.9535	0.9761	0.9887	0.9951	0.9980
1.9	0.5753	0.6879	0.7852	0.8621	0.9177	0.9545	0.9767	0.9890	0.9952	0.9981
2	0.5793	0.6915	0.7881	0.8643	0.9192	0.9554	0.9772	0.9893	0.9953	0.9981
2.1	0.5832	0.6950	0.7910	0.8665	0.9207	0.9564	0.9778	0.9896	0.9955	0.9982
2.2	0.5871	0.6985	0.7939	0.8686	0.9222	0.9573	0.9783	0.9898	0.9956	0.9982
2.3	0.5910	0.7019	0.7967	0.8708	0.9236	0.9582	0.9788	0.9901	0.9957	0.9983
2.4	0.5948	0.7054	0.7995	0.8729	0.9251	0.9591	0.9793	0.9904	0.9959	0.9984
2.5	0.5987	0.7088	0.8023	0.8749	0.9265	0.9599	0.9798	0.9906	0.9960	0.9984
2.6	0.6026	0.7123	0.8051	0.8770	0.9279	0.9608	0.9803	0.9909	0.9961	0.9985
2.7	0.6064	0.7157	0.8078	0.8790	0.9292	0.9616	0.9808	0.9911	0.9962	0.9985
2.8	0.6103	0.7190	0.8106	0.8810	0.9306	0.9625	0.9812	0.9913	0.9963	0.9986
2.9	0.6141	0.7224	0.8133	0.8830	0.9319	0.9633	0.9817	0.9916	0.9964	0.9986

Table 9.2: Quantiles of the Student-t distribution. The rows correspond to different degrees of freedom (ν, say); the columns correspond to different probabilities (z, say). The cell gives q that is s.t. $\mathbb{P}(-q < X < q) = z$, with $X \sim t(\nu)$.

	0.05	0.1	0.75	0.9	0.95	0.975	0.99	0.999
1	0.079	0.158	2.414	6.314	12.706	25.452	63.657	636.619
2	0.071	0.142	1.604	2.920	4.303	6.205	9.925	31.599
3	0.068	0.137	1.423	2.353	3.182	4.177	5.841	12.924
4	0.067	0.134	1.344	2.132	2.776	3.495	4.604	8.610
5	0.066	0.132	1.301	2.015	2.571	3.163	4.032	6.869
6	0.065	0.131	1.273	1.943	2.447	2.969	3.707	5.959
7	0.065	0.130	1.254	1.895	2.365	2.841	3.499	5.408
8	0.065	0.130	1.240	1.860	2.306	2.752	3.355	5.041
9	0.064	0.129	1.230	1.833	2.262	2.685	3.250	4.781
10	0.064	0.129	1.221	1.812	2.228	2.634	3.169	4.587
20	0.063	0.127	1.185	1.725	2.086	2.423	2.845	3.850
30	0.063	0.127	1.173	1.697	2.042	2.360	2.750	3.646
40	0.063	0.126	1.167	1.684	2.021	2.329	2.704	3.551
50	0.063	0.126	1.164	1.676	2.009	2.311	2.678	3.496
60	0.063	0.126	1.162	1.671	2.000	2.299	2.660	3.460
70	0.063	0.126	1.160	1.667	1.994	2.291	2.648	3.435
80	0.063	0.126	1.159	1.664	1.990	2.284	2.639	3.416
90	0.063	0.126	1.158	1.662	1.987	2.280	2.632	3.402
100	0.063	0.126	1.157	1.660	1.984	2.276	2.626	3.390
200	0.063	0.126	1.154	1.653	1.972	2.258	2.601	3.340
500	0.063	0.126	1.152	1.648	1.965	2.248	2.586	3.310

Table 9.3: Quantiles of the χ^2 distribution. The rows correspond to different degrees of freedom; the columns correspond to different probabilities.

	0.05	0.1	0.75	0.9	0.95	0.975	0.99	0.999
1	0.004	0.016	1.323	2.706	3.841	5.024	6.635	10.828
2	0.103	0.211	2.773	4.605	5.991	7.378	9.210	13.816
3	0.352	0.584	4.108	6.251	7.815	9.348	11.345	16.266
4	0.711	1.064	5.385	7.779	9.488	11.143	13.277	18.467
5	1.145	1.610	6.626	9.236	11.070	12.833	15.086	20.515
6	1.635	2.204	7.841	10.645	12.592	14.449	16.812	22.458
7	2.167	2.833	9.037	12.017	14.067	16.013	18.475	24.322
8	2.733	3.490	10.219	13.362	15.507	17.535	20.090	26.124
9	3.325	4.168	11.389	14.684	16.919	19.023	21.666	27.877
10	3.940	4.865	12.549	15.987	18.307	20.483	23.209	29.588
20	10.851	12.443	23.828	28.412	31.410	34.170	37.566	45.315
30	18.493	20.599	34.800	40.256	43.773	46.979	50.892	59.703
40	26.509	29.051	45.616	51.805	55.758	59.342	63.691	73.402
50	34.764	37.689	56.334	63.167	67.505	71.420	76.154	86.661
60	43.188	46.459	66.981	74.397	79.082	83.298	88.379	99.607
70	51.739	55.329	77.577	85.527	90.531	95.023	100.425	112.317
80	60.391	64.278	88.130	96.578	101.879	106.629	112.329	124.839
90	69.126	73.291	98.650	107.565	113.145	118.136	124.116	137.208
100	77.929	82.358	109.141	118.498	124.342	129.561	135.807	149.449
200	168.279	174.835	213.102	226.021	233.994	241.058	249.445	267.541
500	449.147	459.926	520.950	540.930	553.127	563.852	576.493	603.446

Table 9.4: Quantiles of the $\mathcal{F}$ distribution. The columns and rows correspond to different degrees of freedom (resp. n_1 and n_2). The different panels correspond to different probabilities (α) The corresponding cell gives z that is s.t. $\mathbb{P}(X \leq z) = \alpha$, with $X \sim \mathcal{F}(n_1, n_2)$.

	1	2	3	4	5	6	7	8	9	10
alpha = 0.9										
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663
500	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644	1.612
alpha = 0.95										
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850
alpha = 0.99										
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443	2.356

Bibliography

Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10(1):465–503.

Anderson, T. (1971). *The Statistical Analysis of Time Series*. Wiley.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58(2):277–297.

Cameron, A. C. and Miller, D. L. (2014). A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50(2).

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245):32–61.

De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443–473. Twenty five years of forecasting.

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics*, 88(9):1697–1720.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Durbin, J. (1954). Errors in variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22(1/3):23–32.

Durbin, J. and Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression. I. *Biometrika*, 37(3-4):409–428.

Durbin, J. and Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, 38(1-2):159–178.

Fritsch, M., Pua, A. A. Y., and Schnurbus, J. (2019). Pdynmc - An R-package for estimating linear dynamic panel data models based on linear and nonlinear moment conditions. Passauer Diskussionspapiere, Betriebswirtschaftliche Reihe B-39-19, University of Passau, Faculty of Business and Economics.

Gouriéroux, C. and Monfort, A. (1995). *Statistics and Econometric Models*, volume 1 of *Themes in Modern Econometrics*. Cambridge University Press.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education, fifth edition.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Jordà, O., Schularick, M., and Taylor, A. M. (2017). Macrofinancial History and the New Business Cycle Facts. *NBER Macroeconomics Annual*, 31(1):213–263.

Litterman, R. and Scheinkman, J. (1991). Common Factors Affecting Bond Returns. *Journal of Fixed Income*, (1):54–61.

MacKinnon, J. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

MacKinnon, J. G., Ørregaard Nielsen, M., and Webb, M. D. (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. (1995). Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *American Economic Review*, 85(3):322–340.

Nakosteen, R. A. and Zimmer, M. (1980). Migration and income: The question of self-selection. *Southern Economic Journal*, 46(3):840–851.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415.

Stock, J. and Watson, M. W. (2003). *Introduction to Econometrics*. Prentice Hall, New York.

Stock, J. H. and Yogo, M. (2005). *Testing for Weak Instruments in Linear IV Regression*, page 80–108. Cambridge University Press.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B*, 73(3):273–282.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, 41(4):733–750.