Advanced Econometrics

Jean-Paul Renne

2022-08-19

Contents

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2022) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Below is an example borrowed from Petersen.

```
library(sandwich)
## Petersen's data
data("PetersenCL", package = "sandwich")
m <- lm(y ~ x, data = PetersenCL)
```

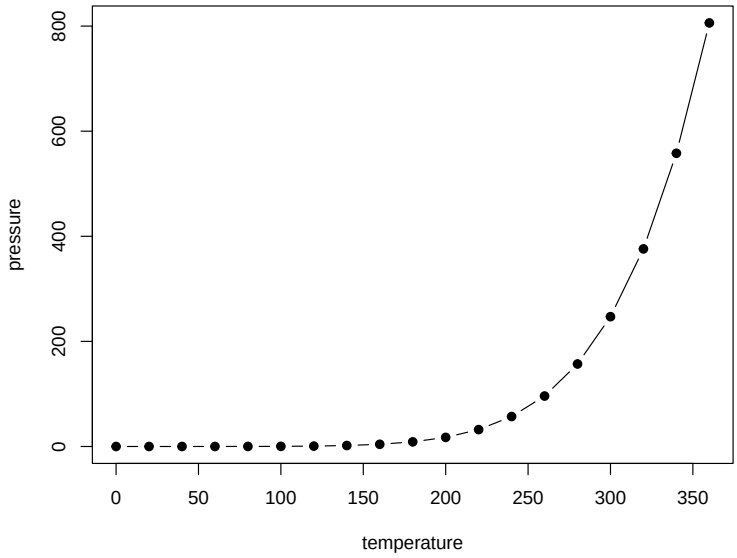


Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

```
## clustered covariances
## one-way
vcovCL(m, cluster = ~ firm)
```

```
##               (Intercept)             x
## (Intercept)  4.490702e-03 -6.473517e-05
## x           -6.473517e-05  2.559927e-03
```

```
vcovCL(m, cluster = PetersenCL$firm) ## same
```

```
##               (Intercept)             x
## (Intercept)  4.490702e-03 -6.473517e-05
## x           -6.473517e-05  2.559927e-03
```

```
## one-way with HC2
vcovCL(m, cluster = ~ firm, type = "HC2")
```

```
##               (Intercept)             x
## (Intercept)  4.494487e-03 -6.592912e-05
## x           -6.592912e-05  2.568236e-03
```

```
## two-way
vcovCL(m, cluster = ~ firm + year)
```

```
##               (Intercept)             x
## (Intercept)  4.233313e-03 -2.845344e-05
## x           -2.845344e-05  2.868462e-03
```

```
vcovCL(m, cluster = PetersenCL[, c("firm", "year")]) ## same
```

```
##               (Intercept)             x
## (Intercept)  4.233313e-03 -2.845344e-05
## x           -2.845344e-05  2.868462e-03
```

XXXX

Sargan-Hansen () test. Sargan (1958) and Hansen (1982)

Durbin-Wu-Hausman test: Durbin (1954) / Wu (1973) / Hausman (1978)

Use R! is an excellent tutorial. (notably for `plm` and the Arellano-Bond example, 140 UK firms)

Program evaluation (very good survey): Abadie and Cattaneo (2018) Mostly harmless: Angrist and Pischke (2008)

Diff-in-Diff: Card and Krueger (1994)

XXXX

Chapter 3

Linear Regressions

3.1 Specification

Definition 3.1. A linear regression is a model defined through:

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

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. If one wants to have an intercept in the specification, then set $x_{i,1} = 1$ for all i, and β_1 then corresponds to the intercept.

Hypothesis 3.1 (Full rank)**.** There is no exact linear relationship among the independent variables (the $x_{i,k}$s, for a given $i \in \{1, \dots, n\}$).

Hypothesis 3.2 (Conditional mean-zero assumption)**.**

$$\mathbb{E}(\varepsilon|\mathbf{X}) = 0. \tag{3.1}$$

Note that, in Hypothesis 3.2, ε is a n-dimensional vector (where n is the sample size), and $\mathbf{X}$ is the matrix containing all explanatory variables, of dimension $n \times K$.

Proposition 3.1. *Under Hypothesis 3.2:*

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

$\square$

Hypothesis 3.3 (Homoskedasticity)**.**

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

Panel (b) of Figure 3.1 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): $Y_i = 2 + 2X_i + \left(2\mathbb{1}_{\{X_i<0\}} + 0.2\mathbb{1}_{\{X_i\geq 0\}}\right)\varepsilon_i^*$.

```
# load data into R
data(Salaries, package = "carData")
# first six rows of the data
head(Salaries)
```

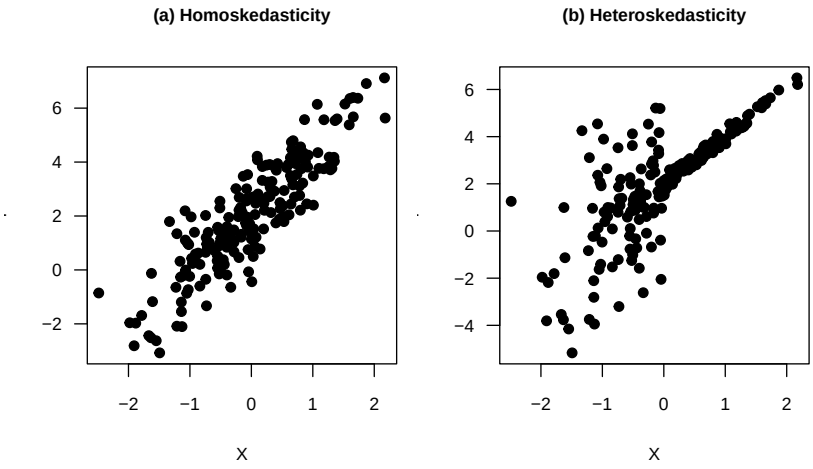


Figure 3.1: This is the caption

```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

```
# Regression:
eq <- lm(salary~.,data=Salaries)
summary(eq)
```

```
##
## Call:
## lm(formula = salary ~ ., data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65248 -13211  -1775  10384  99592
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
## rankAssocProf  12907.6     4145.3   3.114  0.00198 **
## rankProf       45066.0     4237.5  10.635  < 2e-16 ***
## disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
```

```
## yrs.since.phd    535.1      241.0   2.220  0.02698 *
## yrs.service     -489.5      211.9  -2.310  0.02143 *
## sexMale         4783.5     3858.7   1.240  0.21584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22540 on 390 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4463
## F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,1))
par(plt=c(.2,.95,.2,.95))
plot(salary/1000~yrs.since.phd,pch=19,xlab="years since PhD",ylab="Salary",data=Salari
abline(lm(salary/1000~yrs.since.phd,data=Salaries),col="red",lwd=2)
```

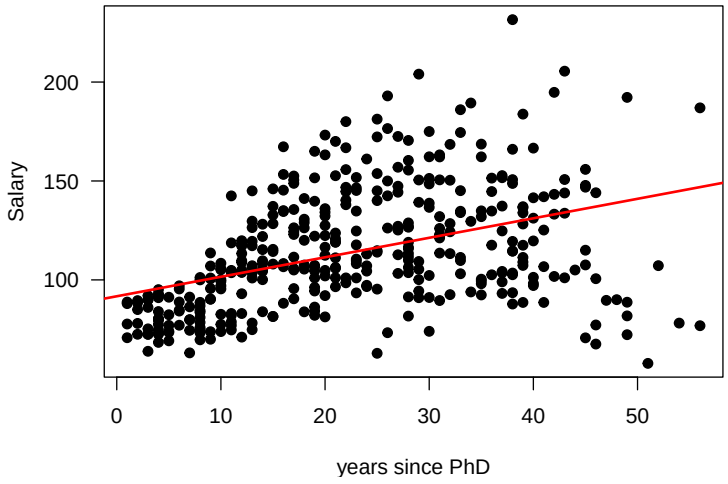


Figure 3.2: Salary versus years after PhD

Hypothesis 3.4 (Uncorrelated residuals)**.**

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

Proposition 3.2. *If 3.3 and 3.4 hold, then:*

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

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

Hypothesis 3.5 (Normal distribution)**.**

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

3.2 Least square estimation

For a given vector of coefficients $\mathbf{b} = [b_1, \dots, b_K]'$, the sum of square 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 the sum of squared residuals amounts to minimizing:

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

We have:

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

Necessary first-order condition (FOC):

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{3.2}$$

Under Assumption 3.1, $\mathbf{X}'\mathbf{X}$ is invertible. Hence:

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

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

The estimated residuals are:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} \tag{3.3}$$

where $\mathbf{M} := \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the **residual maker** matrix. Let us further define a **projection matrix** by $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. These matrices $\mathbf{M}$ and $\mathbf{P}$ are such that:

- $\mathbf{MX} = \mathbf{0}$: 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 $\mathbf{MX} = \mathbf{0}$).
- The fitted values are:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}, \tag{3.4}$$

  i.e., $\hat{\mathbf{y}}$ is the projection of the vector $\mathbf{y}$ onto the vectorial space spanned by the columns of $\mathbf{X}$.
- It can be shown that each column $\tilde{\mathbf{x}}_k$ of $\mathbf{X}$ is orthogonal to $\mathbf{e}$. $\Rightarrow$ If intercepts are included in the regression ($x_{i,1} \equiv 1$), the average of the residuals is null.

Here are some 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).

Proposition 3.3 (Properties of the OLS estimator)**.** *We have:*

i. Under Assumptions 3.1 and 3.2, the OLS estimator is linear and unbiased.

ii. Under Hypotheses 3.1 to 3.4, 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 3.1, $\mathbf{X}'\mathbf{X}$ can be inverted. We have:

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

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

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

   By Hypothesis 3.2, we have $\mathbb{E}[\,|\mathbf{X}] = 0$. Hence $\mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = 0$ 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. 3.2, if 3.3 and 3.4 hold, then we have $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \mathbb{V}ar(\varepsilon|\mathbf{X}) = \sigma^2 Id$. The result follows. □

□

3.2.1 Bivariate case

Consider a bivariate situation, where we regressy_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 $b_1 = \bar{y} - b_2\bar{w}$.

3.2.2 Gauss Markow Theorem

Theorem 3.1 (Gauss-Markov Theorem)**.** *Under Assumptions 3.1 to 3.4, 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 3.2).

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 $C\mathbf{X} = \mathbf{I}.\backslash$

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 3.3 and 3.4, see Prop. 3.2). 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 $\mathbb{V}ar(w'\mathbf{b}^*|\mathbf{X}) = w'\mathbb{V}ar(\mathbf{b}|\mathbf{X})w + \sigma^2 w'\mathbf{D}\mathbf{D}'w \geq w'\mathbb{V}ar(\mathbf{b}|\mathbf{X})w = \mathbb{V}ar(w'\mathbf{b}|\mathbf{X})$. □ □

3.2.3 Frish-Waugh

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 $\mathbf{X}$,
- $\mathbf{P}^{\mathbf{X}}$: projection matrix of any regression on $\mathbf{X}$.

Consider the case where we have two sets of explanatory variables: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$. With obvious notations: $\mathbf{b}^{\mathbf{y}/\mathbf{X}} = [\mathbf{b}_1', \mathbf{b}_2']'$.

Theorem 3.2 (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. (3.2)):

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

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:

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

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 former residuals on the latter.

```
Data <- read.csv("https://raw.githubusercontent.com/jrenne/Data4courses/master/parapluie/data4par
dummies <- as.matrix(Data[,4:14])
eq_all <- lm(parapluie~precip+dummies,data=Data)
summary(eq_all)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  -2.9534478 3.40601605 -0.8671268 0.3893855798
## precip        0.1300055 0.03594492  3.6167985 0.0006192876
## dummiesX1    -8.5990682 3.30046623 -2.6054101 0.0115966151
## dummiesX2   -13.3290386 3.30657128 -4.0310755 0.0001613897
## dummiesX3    -7.9829582 3.25949898 -2.4491366 0.0173091333
## dummiesX4    -3.3923533 3.27605582 -1.0354992 0.3046614769
## dummiesX5    -3.7038158 3.25710094 -1.1371511 0.2600724511
## dummiesX6    -3.3606412 3.27334327 -1.0266694 0.3087670632
## dummiesX7    -7.3158812 3.28491529 -2.2271141 0.0297682836
## dummiesX8    -7.7172773 3.26128164 -2.3663327 0.0212677987
## dummiesX9    -4.6491997 3.26005024 -1.4261129 0.1591057652
## dummiesX10   -5.1091987 3.25143961 -1.5713651 0.1214457733
## dummiesX11    1.9807700 3.25942144  0.6077060 0.5457145714
```

```
deseas_parapluie <- lm(parapluie~dummies,data=Data)$residuals
deseas_precip    <- lm(precip~dummies,data=Data)$residuals
eq_frac <- lm(deseas_parapluie~deseas_precip)
summary(eq_frac)$coefficients
```

```
##                    Estimate Std. Error       t value     Pr(>|t|)
## (Intercept)   -3.265931e-16 0.60898084 -5.362946e-16 1.0000000000
## deseas_precip  1.300055e-01 0.03300004  3.939557e+00 0.0001907741
```

When b_2 is scalar (and then $\mathbf{X}_2$ is of dimension $n \times 1$), Theorem 3.2 leads to:

$$b_2 = \frac{\mathbf{X}_2' M^{\mathbf{X}_1} \mathbf{y}}{\mathbf{X}_2' M^{\mathbf{X}_1} \mathbf{X}_2} \quad \text{(partial regression coefficient)}.$$

3.2.4 Goodness of fit

Define the total variation in y as the sum of squared deviations:

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

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:

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

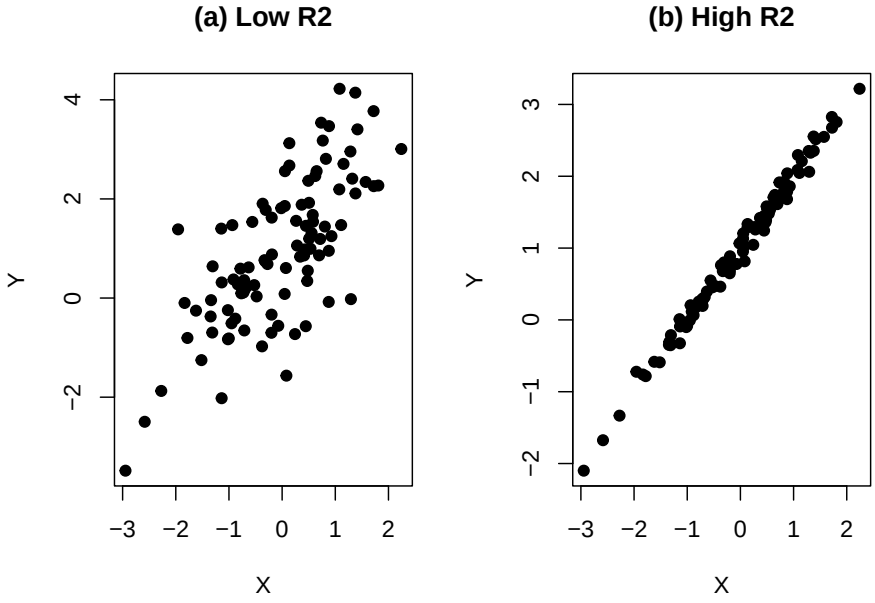
$$\boxed{\text{Coefficient of determination} = \frac{Expl.SS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}.} \quad (3.5)$$

It can be shown [Greene, 2012, Section 3.5] that:

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

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

```
par(mfrow=c(1,2))
par(plt=c(.3,.95,.2,.85))
N <- 100
eps <- rnorm(N)
X <- rnorm(N)
Y <- 1 + X + eps
plot(X,Y,pch=19,main="(a) Low R2")
Y <- 1 + X + .1*eps
plot(X,Y,pch=19,main="(b) High R2")
```



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:

$$r_{yz}^{\mathbf{X}} = \frac{\mathbf{z}^{*\prime}\mathbf{y}^*}{\sqrt{(\mathbf{z}^{*\prime}\mathbf{z}^*)(\mathbf{y}^{*\prime}\mathbf{y}^*)}}. \tag{3.6}$$

Proposition 3.4 (Change in SSR when a variable is added)**.** *We have:*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}^{*\prime}\mathbf{z}^*) \qquad (\leq \mathbf{e}'\mathbf{e}) \tag{3.7}$$

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. (3.2))

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

Hence, in particular $\mathbf{d} = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\mathbf{c}$, 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:

$$\mathbf{u}'\mathbf{u} = (\mathbf{e} - \mathbf{z}^* c)(\mathbf{e} - \mathbf{z}^* c) = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}^{*'}\mathbf{z}^*) - 2c\mathbf{z}^{*'}\mathbf{e}. \tag{3.8}$$

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 3.2), 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. (3.8) leads to the results. □ □

Proposition 3.5 (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.

Proof. Let's use the same notations as in Prop. @ref{prp:chgeR2}. Theorem 3.2 implies that $c = (\mathbf{z}^{*'}\mathbf{z}^*)^{-1}\mathbf{z}^{*'}\mathbf{y}^*$. Using this in Eq. (3.7) 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. (3.6)), 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 $\mathbf{y}'\mathbf{M}_0\mathbf{y}$. □ □

The previous theorem shows that we necessarily increase the R^2 if we add variables, **even if they are irrelevant**.

The **adjusted** R^2, denoted by $\bar{R}^2$, is a fit measure that penalizes large numbers of regressors:

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

3.2.5 Inference and Prediction

Under the normality assumption (Assumption 3.5), we know the distribution of $\mathbf{b}$ (conditional on $\mathbf{X}$). Indeed, $(\mathbf{b}|\mathbf{X}) \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is multivariate Gaussian:

$$\mathbf{b}|\mathbf{X} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \tag{3.9}$$

Problem: In practice, we do not know σ^2 (**population parameter**).

Proposition 3.6. *Under 3.1 to 3.4, an unbiased estimate of σ^2 is given by:*

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}. \tag{3.10}$$

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

Proof. $\mathbb{E}(\mathbf{e}'\mathbf{e}|\mathbf{X}) = \mathbb{E}(\varepsilon'\mathbf{M}\varepsilon|\mathbf{X}) = \mathbb{E}(\mathrm{Tr}(\varepsilon'\mathbf{M}\varepsilon)|\mathbf{X})) = \mathrm{Tr}(\mathbf{M}\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})) = \sigma^2\mathrm{Tr}(\mathbf{M})$. (Note that we have $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 Id$ by Assumptions 3.3 and 3.4, see Prop. 3.2.) Finally: $\mathrm{Tr}(\mathbf{M}) = n - \mathrm{Tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = n - \mathrm{Tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = n - \mathrm{Tr}(Id_{K\times K})$. □ □

Two results will prove important to perform hypothesis testing:

i. We know the distribution of s^2 (Prop. 3.7).
ii. s^2 and $\mathbf{b}$ are independent random variables (Prop. 3.8).

Proposition 3.7. *Under 3.1 to 3.5, we have:* $\dfrac{s^2}{\sigma^2}|\mathbf{X} \sim \chi^2(n-K)/(n-K)$.

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 3.5 – 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. Further, its trace is equal to $n-K$ (see proof of Eq. (3.10)). 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. □ □

Proposition 3.8. *Under Hypotheses 3.1 to 3.5,* $\mathbf{b}$ *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 respective independent variables, are independent. □ □

Under Hypotheses 3.1 to 3.5, let us consider b_k, the k^{th} entry of $\mathbf{b}$:

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

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

Besides, we have (Prop. 3.7):

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

As a result (using Props. 3.7 and 3.8), we have:

$$\boxed{t_k = \frac{\frac{b_k-\beta_k}{\sqrt{\sigma^2 v_k}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2(n-K)}}} = \frac{b_k-\beta_k}{\sqrt{s^2 v_k}} \sim t(n-K),} \tag{3.11}$$

where $t(n-K)$ denotes a t distribution with $n-K$ degrees of freedom.

Remark: $\frac{b_k-\beta_k}{\sqrt{\sigma^2 v_k}}|\mathbf{X} \sim \mathcal{N}(0,1)$ 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.

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. 3.6).

The previous result (Eq. (3.11)) can be extended to any linear combinations of elements of $\mathbf{b}$ (Eq. (3.11) is for its k^{th} component only).

Let us consider $\alpha'\mathbf{b}$, the OLS estimate of $\alpha'\beta$. From Eq. (3.9), 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. (3.11), one can show that Props. 3.7 and 3.8 imply that:

$$\boxed{\frac{\alpha'\mathbf{b}-\alpha'\beta}{\sqrt{s^2\alpha'(\mathbf{X}'\mathbf{X})^{-1}\alpha}} \sim t(n-K).} \tag{3.12}$$

```
par(plt=c(.2,.95,.2,.95))
xx <- seq(-3.5,3.5,by=.01)
plot(xx,dnorm(xx),xlab="X",ylab="",type="l",lwd=2)
lines(xx,dt(xx,df=3),col="red",lwd=2)
lines(xx,dt(xx,df=7),col="red",lwd=2,lty=2)
lines(xx,dt(xx,df=20),col="blue",lwd=3,lty=3)
legend("topright",
       c("N(0,1)","t(3)","t(7)","t(20)"),
       lty=c(1,1,2,3), # gives the legend appropriate symbols (lines)
       lwd=c(2,2,2,3), # line width
       col=c("black","red","red","blue"))
```

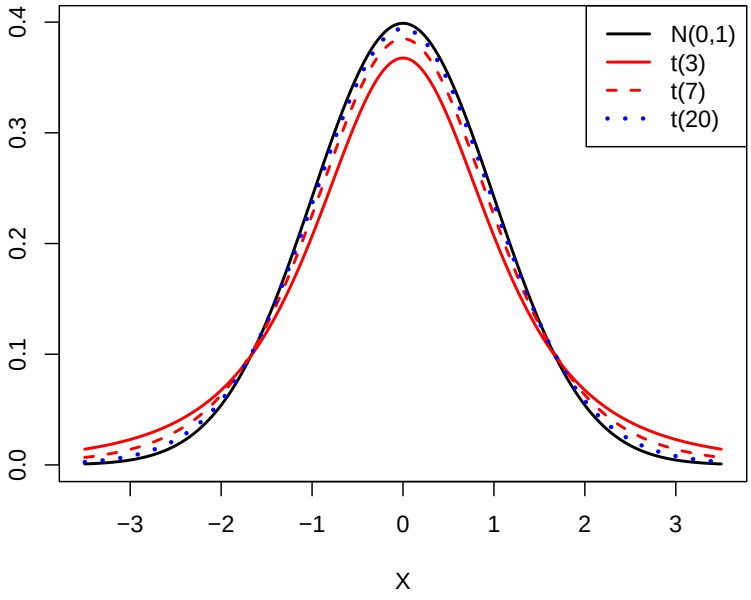


Figure 3.3: The chart shows that the higher the degree of freedom, the closer the distribution of $t(u)$ gets to the normal distribution.

3.2.6 Confidence interval of β_k

Assume we want to compute a (symmetrical) confidence interval $[I_{d,1-\alpha}, I_{u,1-\alpha}]$ that is such that $\mathbb{P}(\beta_k \in [I_{d,1-\alpha}, I_{u,1-\alpha}]) = 1-\alpha$. In particular, we want to have: $\mathbb{P}(\beta_k < I_{d,1-\alpha}) = \frac{\alpha}{2}$.

For this purpose, we make use of $t_k = \frac{b_k - \beta_k}{\sqrt{s^2 v_k}} \sim t(n-K)$ (Eq. (3.11)).

We have:

$$\mathbb{P}(\beta_k < I_{d,1-\alpha}) = \frac{\alpha}{2} \Leftrightarrow$$

$$\mathbb{P}\left(\frac{b_k - \beta_k}{\sqrt{s^2 v_k}} > \frac{b_k - I_{d,1-\alpha}}{\sqrt{s^2 v_k}}\right) = \frac{\alpha}{2} \Leftrightarrow \mathbb{P}\left(t_k > \frac{b_k - I_{d,1-\alpha}}{\sqrt{s^2 v_k}}\right) = \frac{\alpha}{2} \Leftrightarrow$$

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

where $\Phi_{t(n-K)}(\alpha)$ is the c.d.f. of the $t(n-K)$ distribution (Table @ref{tab:Studenttable}).

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

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

3.2.7 Example

The following example is based on the HRS dataset (Health and Retirement Study). We only consider only a subset of this large dataset, focusing on a few variables, and for year 2018 (wave 14). This R script builds the reduced dataset.

```
reducedHRS <- read.csv("https://raw.githubusercontent.com/jrenne/Data4courses/master/HI
eq <- lm(riearn~raedyrs+ragey_b+I(ragey_b^2)+rfemale,data=reducedHRS)
print(summary(eq))
```

```
##
## Call:
## lm(formula = riearn ~ raedyrs + ragey_b + I(ragey_b^2) + rfemale,
##     data = reducedHRS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82512 -29447  -8144  18083 394724
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.336e+05  3.245e+04  -4.116 3.92e-05 ***
## raedyrs       5.216e+03  2.384e+02  21.876  < 2e-16 ***
## ragey_b       4.758e+03  1.086e+03   4.382 1.20e-05 ***
## I(ragey_b^2) -4.441e+01  9.097e+00  -4.882 1.09e-06 ***
## rfemale      -1.499e+04  1.498e+03 -10.007  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50130 on 4540 degrees of freedom
## Multiple R-squared:  0.1173, Adjusted R-squared:  0.1165
## F-statistic: 150.8 on 4 and 4540 DF,  p-value: < 2.2e-16
```

The last two columns give the test statistic and p-values associated to the test whose null hypothesis is:

$$H_0 : \beta_k = 0.$$

The **t-statistics**, that is $b_k/\sqrt{s^2 v_k}$, is the test statistic of the test. Under H_0, the t-statistic is $t(n-K)$ (see Eq. (3.11)). Hence, the **critical region** for the test of size α is:

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

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, 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.

3.2.8 Set of linear restrictions

We consider the following model:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim i.i.d.\mathcal{N}(0, \sigma^2).$$

and we want to test for the *joint* validity of a set of restrictions involving the components of β in a linear way.

Set of linear restrictions:

$$\begin{array}{rcl} r_{1,1}\beta_1 + \cdots + r_{1,K}\beta_K & = & q_1 \\ \vdots & & \vdots \\ r_{J,1}\beta_1 + \cdots + r_{J,K}\beta_K & = & q_J, \end{array} \tag{3.13}$$

that can be written in matrix form:

$$\mathbf{R}\beta = \mathbf{q}. \tag{3.14}$$

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

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

Under Hypotheses 3.1 to 3.4, $\mathbb{V}ar(\mathbf{m}|\mathbf{X}) = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ (see Prop. 3.3).

Consider the test:

$$\boxed{H_0 : \mathbf{R}\beta - \mathbf{q} = 0 \text{ against } H_1 : \mathbf{R}\beta - \mathbf{q} \neq 0.} \tag{3.15}$$

We could perform a **Wald test**. Under 3.1 to 3.5 –we need the normality assumption– and under H_0, it can be shown that we have:

$$W = \mathbf{m}'\mathbb{V}ar(\mathbf{m}|\mathbf{X})^{-1}\mathbf{m} \sim \chi^2(J). \tag{3.16}$$

However, σ^2 is unknown. Hence we cannot compute W.

We can however approximate it be replacing σ^2 by s^2. The distribution of this new statistic is not $\chi^2(J)$ any more; it is an $\mathcal{F}$ **distribution**, and the test is called F **test**.

Proposition 3.9. *Under Hypotheses 3.1 to 3.5 and if Eq.* (3.15) *holds, we have:*

$$F = \frac{W}{J}\frac{\sigma^2}{s^2} = \frac{\mathbf{m}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{m}}{s^2 J} \sim \mathcal{F}(J, n-K), \tag{3.17}$$

where $\mathcal{F}$ is the distribution of the F-statistic.

Proof. According to Eq. (3.16), $W/J \sim \chi^2(J)/J$. 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 3.1 to 3.4, the covariance between $\mathbf{T}\varepsilon$ and $\mathbf{M}\varepsilon$ is $\sigma^2\mathbf{T}\mathbf{M} = \mathbf{0}$. Therefore, under 3.5, these variables are Gaussian variables with 0 covariance. Hence they are independent. □ □

Remark: For large $n - K$, the $\mathcal{F}_{J,n-K}$ distribution converges to $\mathcal{F}_{J,\infty} = \chi^2(J)/J$.

Proposition 3.10. *The F-statistic defined by Eq.* (3.17) *is also equal to:*

$$F = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)} = \frac{(SSR_{restr} - SSR_{unrestr})/J}{SSR_{unrestr}/(n-K)}, \tag{3.18}$$

where R_*^2 is the coef. of determination (Eq. ??eq:RR2)) 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 Prop. @ref(prp:constrained_LS), we know that $\mathbf{b}_* - \mathbf{b} = -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$. 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. 3.9 is also equal to:

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

□

The null hypothesis H_0 (Eq. (3.15)) of the F-test is rejected if F –defined by Eq. (3.17) or (3.18)– is higher than $\mathcal{F}_{1-\alpha}(J, n-K)$. (Hence, this test is a one-sided test.)

3.2.9 Common pitfalls

3.2.10 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:

$$\sum_i x_{i,1}^2 - \frac{(\sum_i x_{i,1}x_{i,2})^2}{\sum_i x_{i,2}^2} = \sum_i x_{i,1}^2(1 - correl_{1,2}^2), \tag{3.19}$$

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 $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$).

3.2.11 Omitted variables

Consider the following model, called "True model'':

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

Then, if one computes $\mathbf{b}_1$ by regressing $\mathbf{y}$ on $\mathbf{X}_1$ only, we get:

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

Hence, we obtain the omitted-variable formula:

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

(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 $\mathbf{X}_1$).

Example 3.1. Consider the "true model'':

$$wage_i = \beta_0 + \beta_1 edu_i + \beta_2 ability_i + \varepsilon_i, \quad \varepsilon_i \sim i.i.d.\,\mathcal{N}(0,\sigma_\varepsilon^2) \tag{3.20}$$

Further, we assume that the *edu* variable is correlated to the *ability*. Specifically:

$$edu_i = \alpha_0 + \alpha_1 ability_i + \eta_i, \quad \eta_i \sim i.i.d.\,\mathcal{N}(0,\sigma_\eta^2).$$

Assume we mistakingly run the regression omitting the *ability* variable:

$$wage_i = \gamma_0 + \gamma_1 edu_i + \xi_i. \tag{3.21}$$

It can be seen that $\xi_i = \varepsilon_i - (\beta_2/\alpha_1)\eta_i \sim i.i.d.\,\mathcal{N}(0,\sigma_\varepsilon^2 + (\beta_2/\alpha_1)^2\sigma_\eta^2)$ and that the population regression coefficient is $\gamma_1 = \beta_1 + \beta_2/\alpha_1 \neq \beta_1$.

Example 3.2. 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 folowing 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)
eq <- lm(testscr~str,data=CASchools)
summary(eq)$coefficients
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 698.932949  9.4674911 73.824516 6.569846e-242
## str          -2.279808  0.4798255 -4.751327  2.783308e-06
```

```
eq <- lm(testscr~str+lunch,data=CASchools)
summary(eq)$coefficients
```

```
##                 Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 702.9113020 4.70024626 149.547760  0.000000e+00
## str          -1.1172255 0.24035528  -4.648225  4.498554e-06
## lunch        -0.5997501 0.01676439 -35.775242 3.709097e-129
```

```
eq <- lm(testscr~str+lunch+english,data=CASchools)
summary(eq)$coefficients
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 700.1499572 4.68568672 149.423126 0.000000e+00
## str          -0.9983090 0.23875428  -4.181324 3.535873e-05
## lunch        -0.5473454 0.02159885 -25.341418 2.303048e-86
## english      -0.1215735 0.03231728  -3.761872 1.928369e-04
```

3.2.12 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. (3.19).

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 of type $H_0 : \beta_1 = 0$ due to the loss of power, so we might infer that they are no relevant variables when they truly are (Type-II error; False Negative).

3.3 Large Sample Properties

Even if we relax the normality assumption (Hypothesis 3.5), we can approximate the finite-sample behavior of the estimators by using *large-sample* or *asymptotic properties*.

To begin with, we proceed under Hypothesis 3.1 to 3.4. (We will see later how to deal with –partial– relaxations of Hypothesis 3.3 and 3.4.)

Under regularity assumptions, under 3.1 to 3.4, even if the residuals are not normally-distributed, the least square estimators can be *asymptotically normal* and inference can be performed as in small samples when 3.1 to 3.5 hold. This derives from Prop. 3.11 (below). The F-test (Prop. **??**prp:Ftest)) and the t-test can then be performed.

Proposition 3.11. *Under Assumptions 3.1 to 3.4, and assuming further that:*

$$Q = plim_{n\to\infty} \frac{\mathbf{X}'\mathbf{X}}{n}, \tag{3.22}$$

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

$$\sqrt{n}(\mathbf{b} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 Q^{-1}\right). \tag{3.23}$$

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 $A \neq \mathbf{0}$), $\text{plim}_{n\to\infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} = \mathbf{Q}^{-1}$. 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_is 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 the V_is 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 then leads to the results. □ □

In practice, σ^2 is estimated with $\frac{\mathbf{e}'\mathbf{e}}{n-K}$ (Eq. (3.10)) and $\mathbf{Q}^{-1}$ with $\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$.

Eqs. (3.22) and (3.23) respectively correspond to convergences in probability and in distribution.

3.4 Instrumental Variables

Here, we want to relax Hypothesis 3.2 –conditional mean zero assumption, implying in particular that $\mathbf{x}_i$ and ε_i are uncorrelated.

We consider the following model:

$$y_i = \mathbf{x_i}'\beta + \varepsilon_i, \quad \text{where } \mathbb{E}(\varepsilon_i) = 0 \text{ and } \mathbf{x_i} \not\perp \varepsilon_i. \tag{3.24}$$

Definition 3.2. 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.

Example. Let us make the assumption $\mathbf{x}_i \not\perp \varepsilon_i$ (in (3.24)) more precise:

$$\mathbb{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathbb{E}(\varepsilon_i \mathbf{x_i}) = \gamma. \tag{3.25}$$

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

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

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 funct.: $\left[\text{plim } \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1} = \text{plim } \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}$. The Slutsky Theorem 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 3.12. *If $\mathbf{z}_i$ is a L-dimensional random variable that constitutes a valid set of instruments (see Def. 3.2) 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. 3.11, 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:

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

And when $L > K$? Idea: First regress $\mathbf{X}$ on the space spanned by $\mathbf{Z}$ and then regress $\mathbf{y}$ on the fitted values $\hat{\mathbf{X}} := \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$. That is $\mathbf{b}_{iv} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$:

$$\boxed{\mathbf{b}_{iv} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}.} \tag{3.26}$$

In this case, Prop. 3.12 still holds, with $\mathbf{b}_{iv}$ given by Eq. (3.26).

$\mathbf{b}_{iv}$ is also the result of the regression of $\mathbf{y}$ on $\mathbf{X}^*$, where the columns of $\mathbf{X}^*$ are the (othogonal) projections of those of $\mathbf{X}$ on $\mathbf{Z}$, i.e. $\mathbf{X}^* = \mathbf{P}^{\mathbf{Z}}\mathbf{X}$ (using the notations introduced in Eq. (3.4)). Hence the other names of this estimator: **Two-Stage Least Squares (TSLS)**.

If the instruments do not properly satisfy Condition (a) in Def. 3.2 (i.e. if $\mathbf{x}_i$ and $\mathbf{z}_i$ are only loosely related), the instruments are said to be **weak**.

Relevant citation: Andrews et al. (2019).

This problem is for instance discussed in Stock and Yogo (2003). See also Stock and Watson pp.,489-490.

The Hausman test 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).

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 3.3. 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:

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

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. (3.25)).

Estimation of the price elasticity of cigarette demand. Instrument: real tax on cigarettes arising from the state's general sales tax. Presumption: in states with a larger general sales tax, cigarette prices are higher, 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
fm <- ivreg(log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff + I(tax/cpi),
```

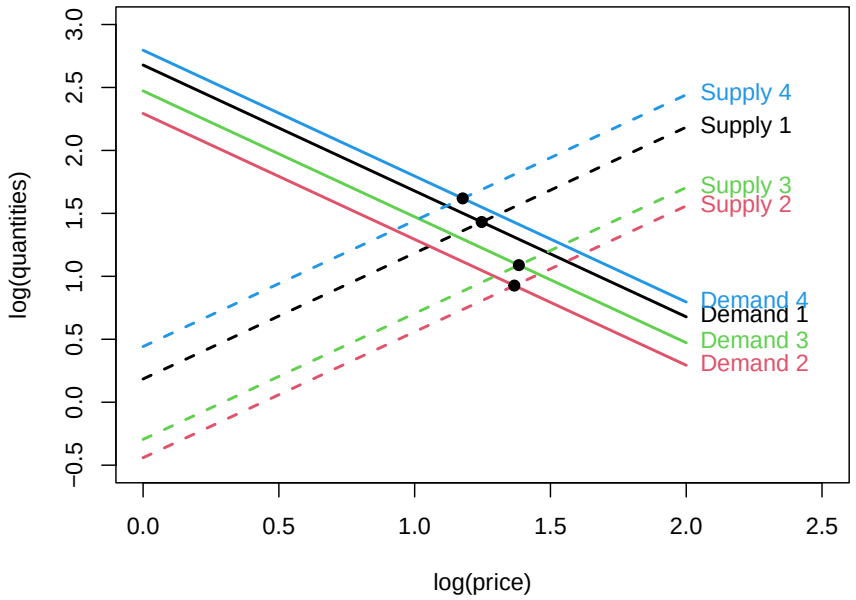


Figure 3.4: This figure illustrates the situation prevailing when estimating a price-elasticity (and the price is endogenous).

```
            data = CigarettesSW, subset = year == "1995")
eq.no.IV <- lm(log(packs) ~ log(rprice) + log(rincome),
               data = CigarettesSW, subset = year == "1995")
summary(fm, vcov = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = log(packs) ~ log(rprice) + log(rincome) | log(rincome) +
##     tdiff + I(tax/cpi), data = CigarettesSW, subset = year ==
##     "1995")
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.6006931 -0.0862222 -0.0009999  0.1164699  0.3734227
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.8950     0.9288  10.654 6.89e-14 ***
## log(rprice)   -1.2774     0.2417  -5.286 3.54e-06 ***
## log(rincome)   0.2804     0.2458   1.141     0.26
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2  44   228.738  <2e-16 ***
## Wu-Hausman         1  44     3.823  0.0569 .
## Sargan             1  NA     0.333  0.5641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1879 on 45 degrees of freedom
## Multiple R-Squared: 0.4294,  Adjusted R-squared: 0.4041
## Wald test: 17.25 on 2 and 45 DF,  p-value: 2.743e-06
```

```
summary(eq.no.IV)
```

```
##
## Call:
## lm(formula = log(packs) ~ log(rprice) + log(rincome), data = CigarettesSW,
##     subset = year == "1995")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59077 -0.07856 -0.00149  0.11860  0.35442
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.3420     1.0227  10.113 3.66e-13 ***
## log(rprice)   -1.4065     0.2514  -5.595 1.24e-06 ***
## log(rincome)   0.3439     0.2350   1.463     0.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1873 on 45 degrees of freedom
## Multiple R-squared:  0.4327, Adjusted R-squared:  0.4075
## F-statistic: 17.16 on 2 and 45 DF,  p-value: 2.884e-06
```

```
fm2 <- ivreg(log(packs) ~ log(rprice) | tdiff, data = CigarettesSW, subset = year == "
anova(fm, fm2)
```

```
## Analysis of Variance Table
##
## Model 1: log(packs) ~ log(rprice) + log(rincome) | log(rincome) + tdiff +
##     I(tax/cpi)
## Model 2: log(packs) ~ log(rprice) | tdiff
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 1.5880
## 2     46 1.6668 -1 -0.078748 1.3815  0.246
```

```
library(sem)
data("CollegeDistance", package = "AER")
simple.ed.1s<- lm(education ~ urban + gender + ethnicity + unemp + distance,
                  data = CollegeDistance)
CollegeDistance$ed.pred<- predict(simple.ed.1s)
simple.ed.2s<- lm(wage ~ urban + gender + ethnicity + unemp + ed.pred ,
                  data = CollegeDistance)

simple.comp<- encomptest(wage ~ urban + gender + ethnicity + unemp + ed.pred ,
                         wage ~ urban + gender + ethnicity + unemp + education ,
                         data = CollegeDistance)
fsttest<- encomptest(education ~ tuition + gender + ethnicity + urban ,
                     education ~ distance ,
                     data = CollegeDistance)

eqOLS <- lm(wage ~ urban + gender + ethnicity + unemp + education,
            data=CollegeDistance)

summary(eqOLS)
```

```
##
```

```
## Call:
## lm(formula = wage ~ urban + gender + ethnicity + unemp + education,
##     data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3484 -0.8408  0.1808  0.8119  3.9875
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.641490   0.157008  55.039   <2e-16 ***
## urbanyes           0.070117   0.044727   1.568   0.1170
## genderfemale      -0.085242   0.037069  -2.300   0.0215 *
## ethnicityafam     -0.556056   0.052167 -10.659   <2e-16 ***
## ethnicityhispanic -0.544007   0.048670 -11.177   <2e-16 ***
## unemp              0.133101   0.006711  19.834   <2e-16 ***
## education          0.005369   0.010362   0.518   0.6044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.268 on 4732 degrees of freedom
## Multiple R-squared:  0.1098, Adjusted R-squared:  0.1087
## F-statistic: 97.27 on 6 and 4732 DF,  p-value: < 2.2e-16
```

```
summary(simple.ed.2s)
```

```
##
## Call:
## lm(formula = wage ~ urban + gender + ethnicity + unemp + ed.pred,
##     data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1692 -0.8294  0.1502  0.8482  3.9537
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.359032   1.412087  -0.254  0.79931
## urbanyes           0.046144   0.044691   1.033  0.30188
## genderfemale      -0.070753   0.036978  -1.913  0.05576 .
## ethnicityafam     -0.227240   0.072984  -3.114  0.00186 **
## ethnicityhispanic -0.351291   0.057021  -6.161 7.84e-10 ***
## unemp              0.139163   0.006748  20.622  < 2e-16 ***
## ed.pred            0.647099   0.100592   6.433 1.38e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.263 on 4732 degrees of freedom
## Multiple R-squared:  0.1175, Adjusted R-squared:  0.1163
## F-statistic:   105 on 6 and 4732 DF,  p-value: < 2.2e-16
```

```
eqTSLS <- tsls(wage ~ urban + gender + ethnicity + unemp + education,
               ~ urban + gender + ethnicity + unemp + distance,
               data=CollegeDistance)

eqTSLS <- ivreg(wage ~ urban + gender + ethnicity + unemp + education|
                  urban + gender + ethnicity + unemp + distance,
                data=CollegeDistance)

summary(eqTSLS, vcov = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = wage ~ urban + gender + ethnicity + unemp + education |
##     urban + gender + ethnicity + unemp + distance, data = CollegeDistance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.20896 -1.14578 -0.02361  1.33303  4.77571
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.35903    1.91755  -0.187   0.8515
## urbanyes           0.04614    0.05926   0.779   0.4362
## genderfemale      -0.07075    0.04974  -1.422   0.1550
## ethnicityafam     -0.22724    0.09539  -2.382   0.0172 *
## ethnicityhispanic -0.35129    0.07577  -4.636 3.64e-06 ***
## unemp              0.13916    0.00934  14.899  < 2e-16 ***
## education          0.64710    0.13691   4.727 2.35e-06 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    1 4732     50.19 1.60e-12 ***
## Wu-Hausman          1 4731     40.30 2.38e-10 ***
## Sargan              0   NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 4732 degrees of freedom
## Multiple R-Squared: -0.6118, Adjusted R-squared: -0.6138
```

```
## Wald test: 57.08 on 6 and 4732 DF,  p-value: < 2.2e-16
```

3.5 General Regression Model

We want to relax the assumption according to which the disturbances are uncorrelated with each other (Hypothesis @ref(hyp:noncorrel_resid)) or the homoskedasticity Hypothesis 3.3.

We replace the latter two assumptions by the general formulation:

$$\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) \quad = \quad \Sigma. \tag{3.27}$$

Note that Eq. ((3.27)) is more general than Hypothesis 3.3 and @ref(hyp:noncorrel_resid) because the diagonal entries of Σ may be different (not the case under Hypothesis 3.3), and the non-diagonal entries of Σ can be $\neq 0$ (contrary to Hypothesis 3.4).

Definition 3.3. Hypothesis 3.1 and 3.2, together with Eq. (3.27), form the **General Regression Model** (GRM) framework.

Note that a regression model where Hypotheses 3.1 to 3.4 hold is a specific case of the GRM framework.

The GRM context notably allows to model **heteroskedasticity** and **autocorrelation**.

- Heteroskedasticity:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_n^2 \end{bmatrix}. \tag{3.28}$$

- Autocorrelation:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho_{2,1} & \dots & \rho_{n,1} \\ \rho_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \rho_{n,n-1} \\ \rho_{n,1} & \rho_{n,2} & \dots & 1 \end{bmatrix}. \tag{3.29}$$

Example 3.4. Autocorrelation is, in particular, a recurrent problem when time-series data are used (see Section @ref(section:TS}).

In a time-series context, subscript i refers to a date. Assume for instance that:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i \tag{3.30}$$

with

$$\varepsilon_i = \rho\varepsilon_{i-1} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2). \tag{3.31}$$

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

$$\Sigma = \frac{\sigma_v^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & ... & \rho^{n-1} \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & ... & 1 \end{bmatrix}. \tag{3.32}$$

3.5.1 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 $\mathbf{I}$. In the transformed model, OLS is BLUE (Gauss-Markow Theorem 3.1).

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

$$\boxed{\mathbf{b}_{GLS} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}}. \tag{3.33}$$

We have:

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

When Σ is unknown, the GLS estimator is 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 $\Sigma(\hat{\theta})$.

If $\hat{\theta}$ is a consistent estimator of θ, then the FGLS is asymptotically efficient (see Example **??**).

By contrast, when Σ has no obvious structure: the OLS (or IV) is the only estimator available. It remains unbiased, consistent, and asymptotically normally distributed, but not efficient. Standard inference procedures are not appropriate any longer.

Autocorrelation in the time-series context. Consider the case presented in Example 3.4. Because the OLS estimate $\mathbf{b}$ of β is consistent, the estimates e_is of the ε_is also are. Consistent estimators of ρ and σ_v are then obtained by regressing the e_is on the e_{i-1}s. Using these estimates in Eq. (3.32) provides a consistent estimate of Σ.

See Cochrane and Orcutt (2012).

Proposition 3.13. *Conditionally on* $\mathbf{X}$*, we have:*

$$\mathbb{V}ar(\mathbf{b}|\mathbf{X}) = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{3.34}$$

Under Hypothesis 3.5, since $\mathbf{b}$ *is linear in* ε*, we have:*

$$\mathbf{b}|\mathbf{X} \sim \mathcal{N}\left(\beta, (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Sigma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\right). \tag{3.35}$$

Note that the variance of the estimator is not $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ any more, so using $s^2(\mathbf{X}'\mathbf{X})^{-1}$ for inference may be misleading.

Proposition 3.14. *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. (3.34) implies that $\mathbb{V}ar(\mathbf{b}|\mathbf{X}) \to 0$. Hence $\mathbf{b}$ converges in mean square and therefore in probability. □ □

Proposition 3.15. *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:*

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

Proposition 3.16. *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. 3.13, 3.15 or 3.16.

Idea: instead of estimating Σ (dimension $n \times n$) directly, one can estimate $\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}$, of dimension $K \times K$ (or $\frac{1}{n}\mathbf{Z}'\Sigma\mathbf{Z}$ in the IV case). Indeed, this is this expression ($\mathbf{X}'\Sigma\mathbf{X}$) that eventually appears in the formulas – for instance in Eq. (3.34).

We have:

$$\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'. \tag{3.36}$$

Robust estimation of asymptotic covariance matrices look for estimates of the previous matrix. Their computation is based on the fact that if $\mathbf{b}$ is consistent, then the e_is are consistent (pointwise) estimators of the ε_is.

Example 3.5. Heteroskedasticity.

This is the case of Eq. (3.28).

We then need to estimate $\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'$. White (1980): Under general conditions:

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'\right). \tag{3.37}$$

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

$$\frac{1}{n}\mathbf{X}'\mathbf{E}^2\mathbf{X}, \tag{3.38}$$

where $\mathbf{E}$ is an $n \times n$ diagonal matrix whose diagonal elements are the estimated residuals e_i.

Illustration: Figure 3.2.

Let us illustrate the influence of heteroskedasticity using simulations.

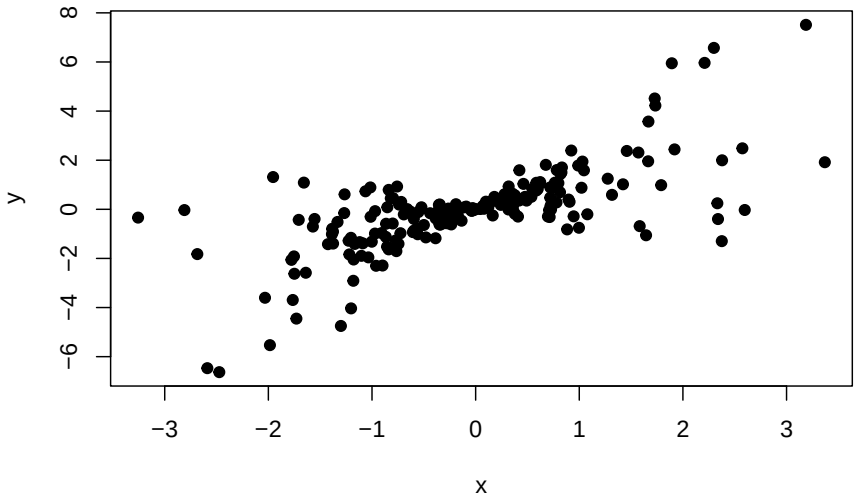
We consider the following model:

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

where the x_is are i.i.d. $t(4)$.

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

```
n <- 200
x <- rt(n,df=5)
y <- x + x*rnorm(n)
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). Using these 1000 estimates of b, we construct an approximated *(kernel-based) distribution of this OLS estimator* (in red on the figure).

For each of the 1000 OLS estimations, we employ *the standard OLS variance formula* ($s^2(\mathbf{X}'\mathbf{X})^{-1}$) to estimate the variance of b. The blue curve is a normal distribution centred on 1 and whose variance is the average of the 1000 previous variance estimates.

The variance of the simulated b is of 0.040 (that is the *true* one); the average of the estimated variances based on the standard OLS formula is of 0.005 (*bad* estimate); the average of the estimated variances based on the White robust covariance matrix is of 0.030 (better estimate).

The standard OLS formula for the variance of b overestimates the precision of this estimator.

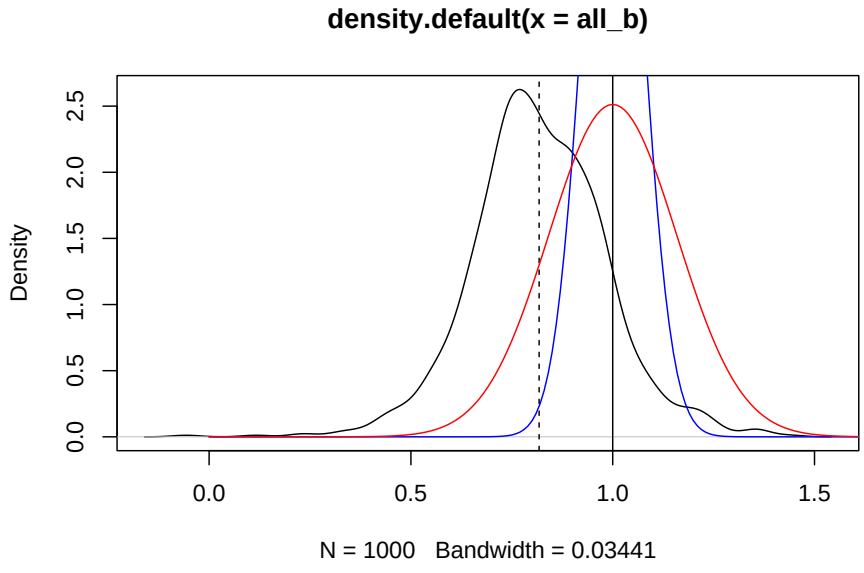
For almost 50% of the simulations, 1 is not included in the 95% confidence interval of β when the computation of the interval is based on the standard OLS formula for the variance of b.

When the White robust covariance matrix is used, 1 is not in the 95% confidence interval of β for less than 10% of the simulations.

```
n <- 200
N <- 1000
XX <- matrix(rt(n*N,df=5),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)
}
plot(density(all_b))
abline(v=mean(all_b),lty=2)
abline(v=1)
x <- seq(0,2,by=.01)
lines(x,dnorm(x,mean = 1,sd = mean(sqrt(all_V_OLS))),col="blue")
lines(x,dnorm(x,mean = 1,sd = mean(sqrt(all_V_White))),col="red")
```



3.5.2 Heteroskedasticity and Autocorrelation (HAC)

This includes the cases of Eqs. (3.28) and (3.29).

Newey and West (1987): If the correlation between terms i and j gets sufficiently small when $|i-j|$ increases:

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'\right) = \tag{3.39}$$
$$\text{plim}\left(\frac{1}{n}\sum_{t=1}^{n}e_t^2\mathbf{x}_t\mathbf{x}_t' + \frac{1}{n}\sum_{\ell=1}^{L}\sum_{t=\ell+1}^{n}w_\ell e_t e_{t-\ell}(\mathbf{x}_t\mathbf{x}_{t-\ell}' + \mathbf{x}_{t-\ell}\mathbf{x}_t')\right)$$

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

Let us illustrate the influence of autocorrelation using simulations.

We consider the following model:

$$y_i = x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, x_i^2), \tag{3.40}$$

where the x_is and the ε_is are such that:

$$x_i = 0.8x_{i-1} + u_i \quad \textit{and} \quad \varepsilon_i = 0.8\varepsilon_{i-1} + v_i, \tag{3.41}$$

where the u_is and the v_is are i.i.d. $\mathcal{N}(0,1)$.

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

We simulate 1000 samples of the same model with $n = 200$.

For each sample, we compute the OLS estimate of β (=1).

Using these 1000 estimates of b, we construct an approximated (kernel-based) distribution of this OLS estimator (in red on the figure).

For each of the 1000 OLS estimations, we employ the standard OLS variance formula ($s^2(\mathbf{X}'\mathbf{X})^{-1}$) to estimate the variance of b. The blue curve is a normal distribution centred on 1 and whose variance is the average of the 1000 previous variance estimates.

The variance of the simulated b is of 0.020 (that is the *true* one); the average of the estimated variances based on the standard OLS formula is of 0.005 (*bad* estimate); the average of the estimated variances based on the White robust covariance matrix is of 0.015 (*better* estimate).

The standard OLS formula for the variance of b overestimates the precision of this estimator.

For about 35% of the simulations, 1 is not included in the 95% confidence interval of β when the computation of the interval is based on the standard OLS formula for the variance of b.

When the Newey-West robust covariance matrix is used, 1 is not in the 95% confidence interval of β for about 13% of the simulations.

For the sake of comparison, let us consider a model with no auto-correlation ($x_i \sim i.i.d.\mathcal{N}(0, 2.8)$ and $\varepsilon_i \sim i.i.d.\mathcal{N}(0, 2.8)$).

3.5.3 How to detect autocorrelation in residuals?

Consider the usual regression (say Eq. (3.30)).

The **Durbin-Watson test** is a typical autocorrelation test. Its 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_is 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. (3.31).)

Critical values depend only on T and K: see e.g. tables CHECK.

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 $d_U(T, K)$:

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

3.6 Summary

		Under Assumptions 3.1+	**b** normal in small sample (Eq. (3.9))	**b** is BLUE (Thm 3.1)	**b** unbiased in small sample (Prop. 3.3)	**b** consistent (Prop. 3.14)*	**b** $\sim$ normal in large sample (Prop. 3.15)*
Normality of disturbances	Uncorrelated residuals Homoskedasticity Condit. mean-zero	3.2	X	X	X	X	X
		3.3	X	X			
		3.4	X	X			
		3.5	X				

*: see however Prop. 3.14 and Prop. 3.15 for additional hypotheses. Specifically $\mathbf{X}'\mathbf{X}/n$ and $\mathbf{X}'\Sigma\mathbf{X}/n$ must converge in proba. to finite positive definite matrices (Σ is defined in Eq. (3.27)).

```r
a <- 1
```

Alors, combien vaut `a`? Answer: 1.

3.7 Clusters

MacKinnon, Nielsen, and Webb (2022)

A nice reference is MacKinnon et al. (2022)

Another one is Cameron and Miller (2014)

See package fwildclusterboot for wild cluster bootstrap.

XXXXXX

Based on MacKinnon et al. (2022):

We have:

$$\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon. \tag{3.42}$$

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. (3.42) rewrites:

$$\mathbf{b} - \beta = \left(\sum_{g=1}^{G} \mathbf{X}_g'\mathbf{X}_g \right)^{-1} \mathbf{X}' \sum_{g=1}^{G} \mathbf{s}_g, \tag{3.43}$$

where $\mathbf{s}_g = \mathbf{X}_g'\varepsilon_g$ denotes the score vector (of dimension $K \times 1$) associoated 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. (3.43) is valid for any partition of $\{1, \dots, n\}$. Nevertheless, dividing the sample into **clusters** really becomes meaningful if we assume that the following hypothesis holds:

Hypothesis 3.6. We have:

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

The real assumotion here is (ii). The first one simply gives a notation for the covariance matrix of the score assiciated 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.

While the choice of clustering structure is sometimes debatable, the structure is generally assumed known in both theoretical and applied work.

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 3.6, it comes that the covariance matrix of $\mathbf{b}$ is:

$$(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^{G} \Sigma_g \right) (\mathbf{X}'\mathbf{X})^{-1} \tag{3.44}$$

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. (3.44) gives the standard expression $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.

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 Eq. (3.38)), i.e.:

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

The natural way to estimate Eq. (3.44) consists in replacing the Σ_g 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 degrees of freedom, this leads to the following estimate of the covariance matrix of $\mathbf{b}$:

$$\frac{G(n-1)}{(G-1)(n-K)} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{g=1}^{G} \widehat{\Sigma}_g \right) (\mathbf{X}'\mathbf{X})^{-1}. \tag{3.45}$$

The previous estimate is CRCV1 in MacKinnon et al. (2022).

Note that we indeed find the White estimator when $G = n$ (see Eq. (3.38)).

Remark, if only one cluster, and neglecting the degree-of-freedom correction, we would have, for $G = 1$:

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

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

3.7.1 Two-way clustering

Let's add a second dimension to the data (e.g., time). There are now two partitions of the data: one through index g, with $g \in \{1, ..., G\}$, 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).

Consider the follwing hypothesis:

Hypothesis 3.7. We have:

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

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 tow 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, the less reliable asymptotic inference based on Eq. (3.45), especially when a very few clusters are unusually large, or when the distribution of the data is heavy-tailed (has fewer moments). These issues are somehow mitigated when the clusters have an approximate factor structure.

In practice, Σ is estimated by:

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

and we then 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 packge `fwildclusterboot` allows to implement such approaches (see, e.g., this tutorial by Alexander Fischer).

3.8 Shrinkage method

Chosing the right variables is often a compicated matter, especially in the presence of many potentially relevant covariates. Keeping a large number of covariates results in large standard deviations for the estimated parameters. In order to address this issue, shrinkage methods have been designed. The objective of these methods is to help to select of 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.

James et al. (2013) (Chapter 6.2)

Example: use credit data (interest rates on XXXX)?

Chapter 4

Panel regressions

The standard panel situation is the following: we have a lot of entities ($i \in \{1, \dots, n\}$) 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 regression then reads:

$$y_{i,t} = \mathbf{x}_{i,t}' \underbrace{\beta}_{K\times 1} + \underbrace{\mathbf{z}_i'\alpha}_{\text{Individual effects}} + \varepsilon_{i,t}. \tag{4.1}$$

Our objective is to estimate the previous equation.

Figure 4.1. 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 to the same entity i.

```
T <- 2 # 2 periods
n <- 12 # 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]))
}
```

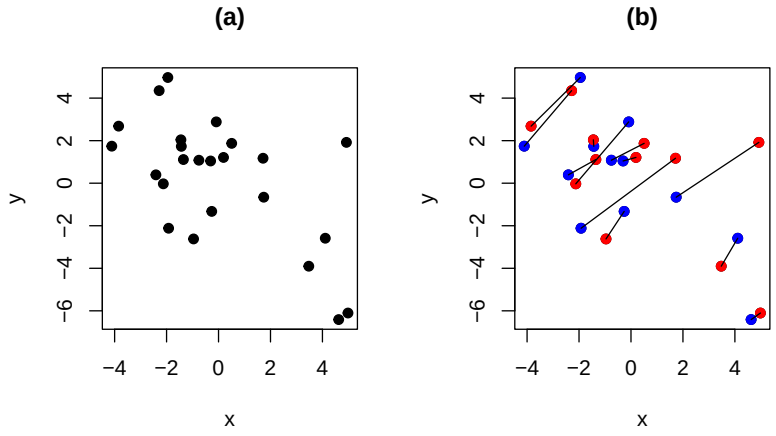


Figure 4.1: The data are the same for both panels. On Panel (b), blue dots are for $t = 1$, red dots are for $t = 2$. The lines relate the dots associated to the same entity i.

Let us know use the Cigarette Consumption Panel dataset of Stock and Watson (2007)

```
data("CigarettesSW", package = "AER")
CigarettesSW$rprice  <- with(CigarettesSW, price/cpi)
CigarettesSW$rincome <- with(CigarettesSW, income/population/cpi)
T <- length(levels(CigarettesSW$year))
n <- length(levels(CigarettesSW$state))
eq.pooled <- lm(log(packs)~log(rprice)+log(rincome),data=CigarettesSW)
print(summary(eq.pooled)$coefficients)
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  10.0671678  0.5156035 19.525020 1.158630e-34
## log(rprice)  -1.3341612  0.1353614 -9.856288 4.120472e-16
## log(rincome)  0.3181371  0.1361194  2.337191 2.157508e-02
```

```
eq.LSDV <- lm(log(packs)~log(rprice)+log(rincome)+state,
              data=CigarettesSW)
print(summary(eq.LSDV)$coefficients[1:3,])
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   9.9543751  0.2641889  37.67901 3.156258e-36
## log(rprice)  -1.2103380  0.1138384 -10.63207 5.590562e-14
## log(rincome)  0.1209004  0.1901069   0.63596 5.279541e-01
```

```
CigarettesSW$year <- as.factor(CigarettesSW$year)
eq.LSDV2 <- lm(log(packs)~log(rprice)+log(rincome)+state+year,
               data=CigarettesSW)
print(summary(eq.LSDV2)$coefficients[1:3,])
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   8.3597463  1.0485390  7.972757 3.776672e-10
## log(rprice)  -1.0559739  0.1490905 -7.082770 7.682380e-09
## log(rincome)  0.4974424  0.3042306  1.635084 1.090085e-01
```

Cigarettes data from Stock and Watson (2003). Data for U.S. states, 2 years: 1985 and 1995. Each colour corresponds to a given State.

Airlines: cost versus fuel price

Airlines data from Greene (2003). Each colour corresponds to a given airline.

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

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

4.0.1 Three standard cases

- **Pooled regression**: $\mathbf{z}_i \equiv 1$.
- **Fixed Effects**: $\mathbf{z}_i$ 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 XXX).

- **Random Effects**: $\mathbf{z}_i$ is unobserved, but uncorrelated with $\mathbf{x}_i$. The model writes:
$$y_{i,t} = \mathbf{x}_{i,t}'\beta + \alpha + \underbrace{u_i + \varepsilon_{i,t}}_{\text{compound disturbance}},$$
where $\alpha = \mathbb{E}(\mathbf{z}_i'\alpha)$ and $u_i = \mathbf{z}_i'\alpha - \mathbb{E}(\mathbf{z}_i'\alpha) \perp \mathbf{x}_i$.

Least squares are consistent (but inefficient, see GLS XXXX).

4.1 Estimation of Fixed Effects Models

Hypothesis 4.1 (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) $\leftrightarrow$ 3.1, (ii) $\leftrightarrow$ 3.2, (iii) $\leftrightarrow$ 3.3 + 3.4.

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:

$$\boxed{\mathbf{y} = [\mathbf{X} \quad \mathbf{D}] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \varepsilon,} \tag{4.2}$$

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. (4.2)) –with the dummy variables– satisfies the Gauss-Markov conditions (Theorem 3.1). Hence, in this context, the OLS estimator is the **best linear unbiased estimator**.

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:

$$\boxed{\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y}.} \tag{4.3}$$

The asymptotical distribution of $[\mathbf{b}', \mathbf{a}']'$ derives from the standard OLS context: Prop. 3.11 can be used after having replaced $\mathbf{X}$ by $\mathbf{Z} = [\mathbf{X} \quad \mathbf{D}]$.

We have:

$$\boxed{\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} \beta \\ \alpha \end{bmatrix}, \sigma^2 \frac{Q^{-1}}{nT}\right)} \tag{4.4}$$

where

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

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 is based on matrix $\mathbf{M_D} = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$, 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 Theorem 3.2, we get:

$$\boxed{\mathbf{b} = [\mathbf{X}'\mathbf{M_D}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{M_D}\mathbf{y}.} \tag{4.5}$$

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 estimates of α are given by:

$$\boxed{\mathbf{a} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathbf{b}),} \tag{4.6}$$

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. (3.2)).

XXXX Regression of the log of airline costs on log(output), log(fuel price) and capacity utilization of the fleet (Data from Greene (2003), see Fig. **??**.

XXXX Regression of the number of cigarette packs (log) on real income (log) and real price of cigarettes (log)

Extension: Fixed time and group effects

Time effects are easily introduced:

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

The LSDV (Eq. (4.2)) can be extended:

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{D} \quad \mathbf{C}] \begin{bmatrix} \beta \\ \alpha \\ \gamma \end{bmatrix} + \varepsilon, \tag{4.7}$$

with:

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

where the T-dimensional vector δ_t is

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

XXXX Geo-located data from Airbnb, Z"urich

Source: Airbnb, date 22 June 2017. Regression of price for Entire home/apt on number of bedrooms, number of people that can be accommodated.

4.2 Estimation of random effects models

Here, the individual effects are assumed not to be correlated to other variables (the $\mathbf{x}_i$s).

Random-effect models write:

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

with

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

Notation: $\eta_{i,t} = u_i + \varepsilon_{i,t}$ and $\eta_i = [\eta_{i,1}, \eta_{i,2}, ..., \eta_{i,T}]'$.

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. (3.33)):

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

(recall that 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

$$\Gamma^{-1/2} = \frac{1}{\sigma_\varepsilon}\left(\mathbf{I} - \frac{\theta}{T}\mathbf{1}\mathbf{1}'\right)$$

with

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}.$$

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?

Idea: taking deviations from group means removes heterogeneity:

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

The previous equation can be consistently estimated by OLS

(the residuals are correlated across t within an entity but the OLS remain consistent though, see Prop. @ref{prp:XXX}).

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 $\mathbf{b}$ is a consistent estimator of β; adjustment for the degrees of freedom:

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

And for 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.

4.2.1 Example: Macro-History database

Date from Jordà et al. (2017) see website.

```
JST <- read.csv("https://raw.githubusercontent.com/jrenne/Data4courses/master/JST/JSTda
JST <- subset(JST,year>1950)
N <- dim(JST)[1]
JST$hpreal <- JST$hpnom/JST$cpi # real house price index
JST$chge_ctry <- c(NaN,JST$iso[2:N]!=JST$iso[1:(N-1)])# will be use to put NA's when c
JST$dhpreal <- log(JST$hpreal/c(NaN,JST$hpreal[1:(N-1)]))
JST$dhpreal[JST$chge_ctry==1] <- NaN
JST$dhpreal[abs(JST$dhpreal)>.3] <- NaN # remove extreme price change
# JST$ltrate_1 <- c(NaN,JST$ltrate[1:(N-1)])
# JST$ltrate_1[JST$chge_ctry==1] <- NaN
# JST$stir_1 <- c(NaN,JST$stir[1:(N-1)])
# JST$stir_1[JST$chge_ctry==1] <- NaN
JST$YEAR <- as.factor(JST$year) # to have time fixed effects
eq <- lm(dhpreal ~ stir + ltrate + iso + YEAR,data=JST)
eq <- lm(dhpreal ~ I(ltrate-stir) + iso + YEAR,data=JST)
lmtest::coeftest(eq)[1:2,]
```

```
##                      Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)      -0.009956382 0.020728411 -0.4803254 0.6310957161
## I(ltrate - stir) -0.004746647 0.001254511 -3.7836614 0.0001632386
```

```
library(sandwich)
library(lmtest)
eq <- lm(dhpreal ~ I(ltrate-stir) + iso,data=JST)
vcov_cluster <- vcovCL(eq, cluster = JST[, c("iso")])
coeftest(eq, vcov = vcov_cluster)[1:2,]
```

```
##                      Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       0.028199744 0.001814513 15.541217 1.778760e-49
## I(ltrate - stir) -0.004732743 0.002349209 -2.014611 4.418367e-02
```

```
vcov_cluster <- vcovCL(eq, cluster = JST[, c("iso","YEAR")])
coeftest(eq, vcov = vcov_cluster)[1:2,]
```

```
##                      Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)       0.028199744 0.002182053 12.923490 1.020962e-35
## I(ltrate - stir) -0.004732743 0.002720606 -1.739592 8.220498e-02
```

4.2.2 Example Spatial Data

```
library(rgdal)
```

```
## Loading required package: sp
```

```
## Please note that rgdal will be retired by the end of 2023,
## plan transition to sf/stars/terra functions using GDAL and PROJ
## at your earliest convenience.
##
## rgdal: version: 1.5-30, (SVN revision 1171)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.4.2, released 2022/03/08
## Path to GDAL shared files: /Library/Frameworks/R.framework/Versions/4.2/Resources/library/rgda
## GDAL binary built with GEOS: FALSE
## Loaded PROJ runtime: Rel. 8.2.1, January 1st, 2022, [PJ_VERSION: 821]
## Path to PROJ shared files: /Library/Frameworks/R.framework/Versions/4.2/Resources/library/rgda
## PROJ CDN enabled: FALSE
## Linking to sp version:1.4-6
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading sp or rgdal.
```

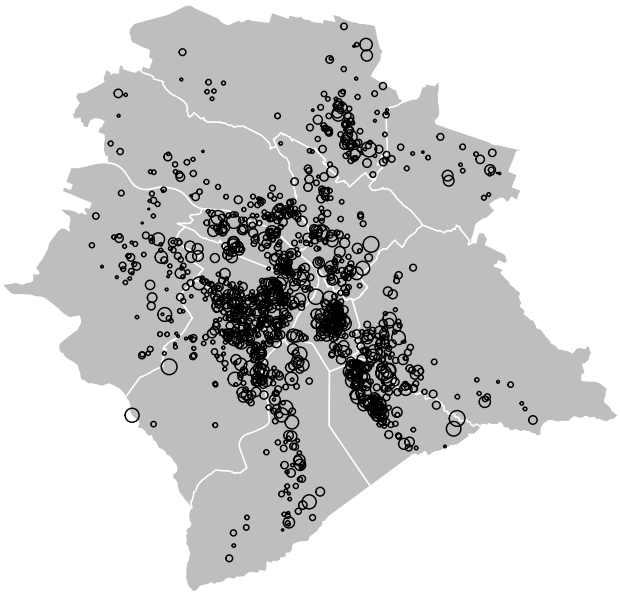
```
library(AEC)
library(sandwich)
library(lmtest)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics T
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(RColorBrewer)
par(mar=c(0,0,0,0))
plot(zurich_districts,col="gray",border="white")
for(i in 1:dim(airbnb)[1]){
  points(airbnb$longitude[i],airbnb$latitude[i],pch=1,cex=(airbnb$price[i]/200))
}
```



```
eq_noFE <- lm(price~bedrooms+accommodates,data=airbnb)
eq_FE <- lm(price~bedrooms+accommodates+neighborhood,data=airbnb)
# stargazer(
```

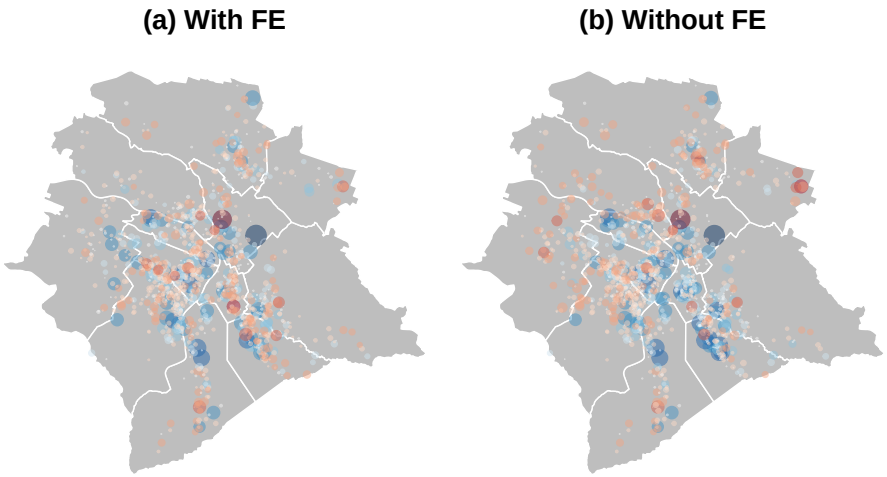
```
#   eq_FE,eq_noFE, type = "text", column.labels = c("FE", "No FE"),
#   omit = c("neighborhood"),
#   omit.labels = c("District FE")#,keep.stat = "n"
# )
# lmtest::coeftest(eq_FE)[1:2,]
# lmtest::coeftest(eq_noFE)[1:2,]

# 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 output (with and without RSE)
stargazer(eq_FE, eq_noFE, eq_FE, eq_noFE, type = "text",
          column.labels = c("FE (no HAC)", "No FE (no HAC)",
                            "FE (with HAC)", "No FE (with HAC)"),
          omit = c("neighborhood"),
          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
```

```
# Plot residuals
par(mfrow=c(1,2))
for(j in 1:2){# with/without FE
  par(plt=c(0,1,0,.8))
  if(j==1){
    residuals <- eq_FE$residuals
    titl <- "(a) With FE"
  }else{
    residuals <- eq_noFE$residuals
    titl <- "(b) Without FE"
  }
  plot(zurich_districts,col="gray",border="white",main=titl)
  # create categories based on resisuals:
  nb_categ <- 11
  categ <- round((nb_categ-1)*(residuals-min(residuals))/(max(residuals)-min(residuals)
  colour <- paste(brewer.pal(n = nb_categ, name = "RdBu"),"77",sep="")
  for(i in 1:dim(airbnb)[1]){
    points(airbnb$longitude[i],airbnb$latitude[i],pch=19,
           cex=(abs(residuals[i])/100),col=colour[categ[i]])
  }
}
```



4.3 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 3.7).

In several cases, auto-correlation in the variable of interest may stem from an auto-regressive specification of the variable of interest. Eq. (4.1) is then replaced by:

$$y_{i,t} = \rho y_{i,t-1} + \mathbf{x}_{i,t}' \underbrace{\beta}_{K\times 1} + \underbrace{\alpha_i}_{\text{Individual effects}} + \varepsilon_{i,t}. \tag{4.8}$$

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 regress $\tilde{\mathbf{y}}$ on $\widetilde{\mathbf{X}}$ (see Eq. (4.5)), 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:

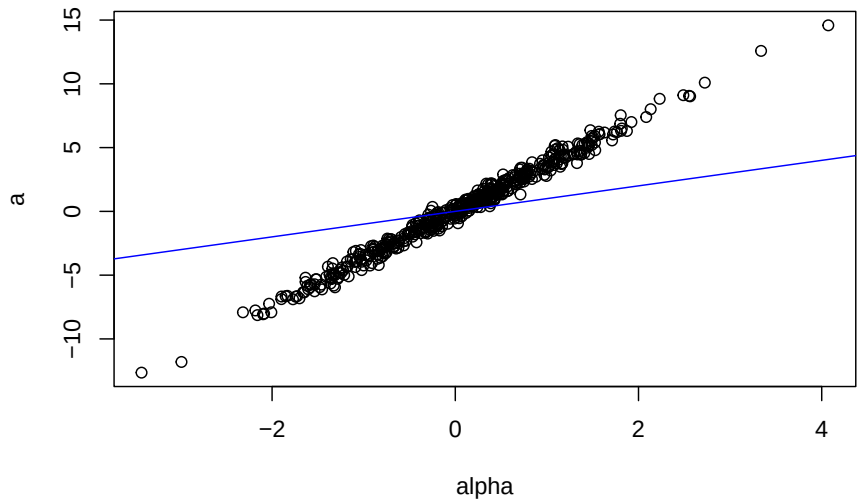
$$\tilde{\varepsilon}_{i,t} = \varepsilon_{i,t} - \frac{1}{T}\sum_{s=1}^{T} \varepsilon_{i,s}.$$

The previous equation shows that the *within-group* estimator (LSDV) introduces all realisations 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 are strictly exogenous (i.e., do not correlate to past, present, and future errors $\varepsilon_{i,t}$).[1]

```
n <- 500;T <- 10
rho <- 0.9;sigma <- .5
alpha <- rnorm(n)
y <- alpha /(1-rho) + sigma^2/(1 - rho^2) * rnorm(n)
all_y <- y
for(t in 2:T){
  eps <- rnorm(n)
  y <- rho * y + alpha + sigma * eps
```

[1] The bias may vanish for large T's.

```
  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")
```




```
#plot(all_y[,1],type="l")
```

How to address this? One can resort to instrumental-variable regressions.

Anderson and Hsiao (1982) proposed a first-differenced Two Stage Least Squares (2SLS) estimator. This estimation is based on the following transformation of the model:

$$\Delta y_{i,t} = \rho \Delta y_{i,t-1} + (\Delta \mathbf{x}_{i,t})' \beta + \Delta \varepsilon_{i,t}. \tag{4.9}$$

The OLS estimates of the parameters are biased because ε_{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 both correlated with $\Delta y_{i,t}$ and 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 assumed to be exogenous (i.e., do not contemporaneous correlate with any of the errors $\varepsilon_{i,s}$, $\forall s$), then $\mathbf{x}_{i,t}$ also is a valid instrument.

```
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_hat  <- Pz %*% Dy
Dy_1hat <- Pz %*% Dy_1
rho_TSLS <- solve(t(Dy_1hat) %*% Dy_1hat) %*% t(Dy_1hat) %*% Dy_hat
```

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]

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

with $\Delta \mathbf{v}_i = [\Delta v_{i,3}, \dots, \Delta v_{i,T}]'$.

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:

$$\text{argmin} \left(\frac{1}{n} \sum_{i=1}^{n} Z'_i \Delta \mathbf{v}_i \right)' W_n \left(\frac{1}{n} \sum_{i=1}^{n} Z'_i \Delta \mathbf{v}_i \right),$$

using the weighting matrix

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

[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 $\mathbf{z}_{i,t} = [y_{i,t}, \mathbf{x}_{i,t+2}, \mathbf{x}_{i,t+1}, \mathbf{x}'_{i,t}]'$).

where the $\widehat{\Delta \mathbf{v}_i}$'s are consistent esitmates of the $\Delta \mathbf{v}_i$ that result from a preliminary estimation. Hence, this estimator is a two-step GMM one.

If the disturbances are homoskedastic, then it can be shown that an asymtotically equilvalent (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:

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

It is straightforward to extend these GMM methods extend 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.
             use.mc.nonlin = FALSE, # use nonlinear (quadratic) moment conditions?
             include.y = TRUE, # instruments should be derived from the lags of the de
             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 fro
             fur.con.lev = FALSE, # include further control variables in equations fro
             varname.reg.fur = c("wage", "capital", "output"), # covariate(s) -in the
             lagTerms.reg.fur = c(1,2,2), # number of lags of the further controls
             include.dum = TRUE, # A logical variable indicating whether dummy variabl
             dum.diff = TRUE, # A logical variable indicating whether dummy variables
             dum.lev = FALSE, # A logical variable indicating whether dummy variables
```

```
             varname.dum = "year",
             w.mat = "iid.err", # One of the character strings c('"iid.err"', '"identity"', '"zer
             std.err = "corrected",
             estimation = "onestep", # One of the character strings c('"onestep"', '"twostep"',
             opt.meth = "none" # numerical optimization procedure. When no nonlinear moment condi
             )
summary(m1)
```

```
##
## 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
```

Chapter 5

Estimation Methods

Context and Objective:

- You observe a sample $\mathbf{y} = \{y_1, ..., y_n\}$.
- You know that these data have been generated by a model parameterized by $\theta_0 \in \mathbb{R}^K$.

5.1 Generalized Method of Moments (GMM)

5.1.1 Framework

We denote by x_t a $p \times 1$ vector of (stationary) variables observed at date t; by θ an $a \times 1$ vector of parameters, and by $h(x_t; \theta)$ 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:

$$\mathbb{E}[h(x_t; \theta_0)] = 0.$$

We denote by $\underline{x_t}$ the information contained in the current and past observations of x_t, that is: $\underline{x_t} = \{x_t, x_{t-1}, ..., x_1\}$. We denote by $g(\underline{x_T}; \theta)$ the sample average of $h(x_t; \theta)$, i.e.:

$$g(\underline{x_T}; \theta) = \frac{1}{T} \sum_{t=1}^{T} h(x_t; \theta).$$

Intuition behind GMM: Choose θ so as to make the sample moment as close as possible to 0.

Definition 5.1. A GMM estimator of θ_0 is given by:

$$\hat{\theta}_T = \text{argmin}_\theta \quad g(\underline{x_T};\theta)' \, W_T \, g(\underline{x_T};\theta),$$

where W_T is a positive definite matrix (that may depend on $\underline{x_T}$).

If $a = r$ (the dimension of θ is the same as that of $h(x_t;\theta)$, or $g(\underline{x_T};\theta)$), $\hat{\theta}_T$ is such that:

$$g(\underline{x_T};\hat{\theta}_T) = 0.$$

Under regularity and identification conditions:

$$\hat{\theta}_T \xrightarrow{p} \theta_0,$$

i.e. $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_T - \theta_0| > \varepsilon) = 0$.

Optimal weighting matrix. The GMM estimator achieving the minimum asymptotic variance is obtained when W_T is the inverse of the matrix S defined by:

$$S := \sum_{\nu=-\infty}^{\infty} \Gamma_\nu,$$

where $\Gamma_\nu := \mathbb{E}[h(x_t;\theta_0)h(x_{t-\nu};\theta_0)']$.

For $\nu \geq 0$, let us define $\hat{\Gamma}_{\nu,T}$ by:

$$\hat{\Gamma}_{\nu,T} = \frac{1}{T} \sum_{t=\nu+1}^{T} h(x_t;\hat{\theta}_T)h(x_{t-\nu};\hat{\theta}_T)',$$

S can be approximated by:

$$\begin{aligned} &\hat{\Gamma}_{0,T} \qquad \text{if the } h(x_t;\theta_0) \text{ are serially uncorrelated and} \\ &\hat{\Gamma}_{0,T} \quad + \sum_{\nu=1}^{q} [1 - \nu/(q+1)](\hat{\Gamma}_{\nu,T} + \hat{\Gamma}'_{\nu,T}) \quad \text{otherwise.} \end{aligned} \tag{5.1}$$

Asymptotic distribution of $\hat{\theta}_T$

We have:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V),$$

where $V = (DS^{-1}D')^{-1}$.

V can be approximated by $(\hat{D}_T \hat{S}_T^{-1} \hat{D}'_T)^{-1}$, where $\hat{S}_T$ is given by Eq. (5.1) and

$$\hat{D}'_T := \left. \frac{\partial g(\underline{x_T};\theta)}{\partial \theta'} \right|_{\theta=\hat{\theta}_T}.$$

5.1.2 Example: Estimation of the Stochastic Discount Factor (s.d.f.)

Under the no-arbitrage assumption, there exists a random variable $\mathcal{M}_{t,t+1}$ 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.:

$$\mathcal{M}_{t,t+1} = 1 - \mathbf{b}_M'(F_{t+1} - \mathbb{E}_t(F_{t+1})), \tag{5.2}$$

where F_t is a vector of factors. Eq. (5.2) 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:

$$\begin{aligned} & \mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]) \\ = \; & \mathbb{E}(\mathbb{E}_t\{z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]\}) \\ = \; & \mathbb{E}(z_{i,t}\underbrace{\mathbb{E}_t\{\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1\}}_{=0}) = 0. \end{aligned} \tag{5.3}$$

We have then converted a conditional moment condition into a unconditional one (which we need to implement the theory above). However, at that stage, we cannot 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. (5.3), 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.

At that stage, we count 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 Corporate Index Total Return Index
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 $g(\underline{x_T};\theta)$; 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. We consider 5 iterations (where W is updated).

```
theta <- c(rep(0,n_F)) # inital value
for(i in 1:5){# recursion on W
  res <- f_aux(theta)
  W <- solve(1/T * t(res$H) %*% res$H)
  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.

```
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,] -0.6774355  0.3182614 -2.1285502
## [2,]  4.3612746 13.1163581  0.3325065
```

The Hansen statistic can be used to test the model. If the model is correct, we have:

$$Tg(\underline{x_T};\theta)'\, S^{-1}\, g(\underline{x_T};\theta) \sim\ i.i.d.\, \chi^2(J-K),$$

where J is the number of moment contraints ($n_z \times n_r$ here) and K is the number of estimated parameters ($= n_F$ here).

```
g <- f_aux(theta)$g
Hanse_stat <- T * t(g) %*% W %*% g
pvalue <- pchisq(q = Hanse_stat,df = n_R*n_z - n_F)
```

5.2 Maximum Likelihood Estimation

Intuition behind the Maximum Likelihood Estimation: Estimator = the value of θ that is such that the probability of having observed $\mathbf{y}$ is the highest possible.

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. $y_i \sim \mathcal{E}(\lambda)$.

You have observed these arrivals for some time, thereby constituting a sample $\{y_1, \dots, y_n\}$. You want to estimate λ (i.e. in that case, the vector of parameters is simply $\theta = \lambda$).

The density of Y is $f(y;\lambda) = \frac{1}{\lambda}\exp(-y/\lambda)$. Fig. 5.1 represents that density functions for different values of λ.

Your 200 observations are reported at the bottom of Fig. 5.1 (red). You build the histogram and report it on the same chart.

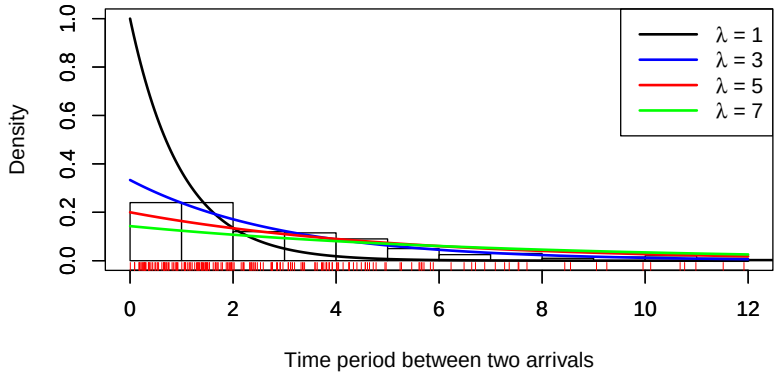


Figure 5.1: The red ticks, at the bottom, indicate observations (there are 200 of them). The historgram is based on these 200 observations

What is your estimate of λ?

Now, 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 observing, for a small ε,

- $1.1 - \varepsilon \leq Y_1 < 1.1 + \varepsilon$,
- $2.2 - \varepsilon \leq Y_2 < 2.2 + \varepsilon$,
- $0.7 - \varepsilon \leq Y_3 < 0.7 + \varepsilon$ and
- $5.0 - \varepsilon \leq Y_4 < 5.0 + \varepsilon$?

Because the y_is 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 $16\varepsilon^4$) as a function of λ.

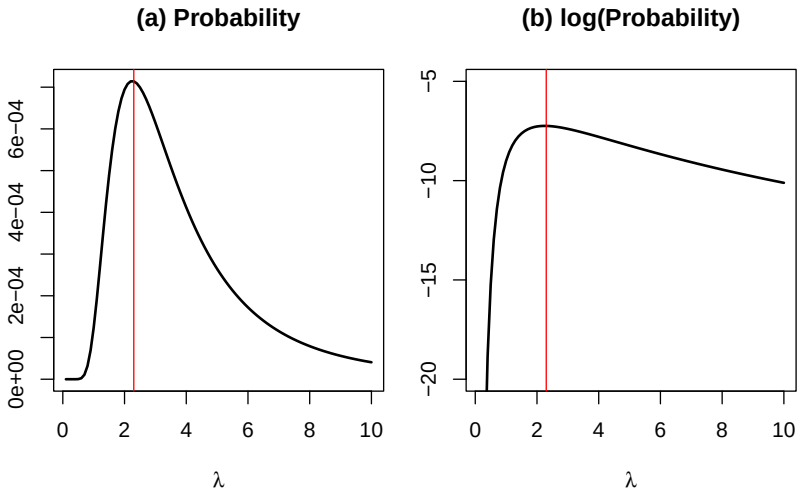


Figure 5.2: Proba. that $y_i - \varepsilon \leq Y_i < y_i + \varepsilon$, $i \in \{1, 2, 3, 4\}$. The vertical red line indicates the maximum of the function.

Back to the example with 200 observations:

5.2.1 Notations

$f(y; \theta)$ denotes the probability density function (p.d.f.) of a random variable Y which depends on a set of parameters θ.

Density of n independent and identically distributed (i.i.d.) observations of Y:

$$f(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta).$$

$\mathbf{y}$ denotes the vector of observations; $\mathbf{y} = \{y_1, \ldots, y_n\}$.

Definition 5.2 (Likelihood function)**.** $\mathcal{L} : \theta \to \mathcal{L}(\theta; \mathbf{y}) = f(y_1, \ldots, y_n; \theta)$ is the **likelihood function**.

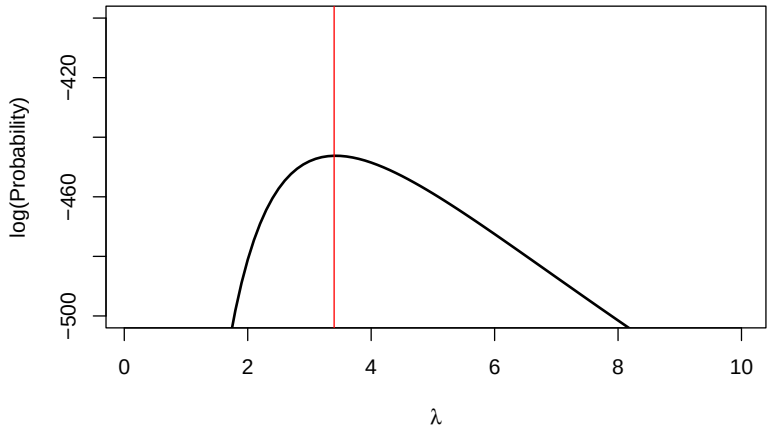


Figure 5.3: Log-likelihood function associated with the 200 i.i.d. observations. The vertical red line indicates the maximum of the function.

We often work with $\log \mathcal{L}$, the **log-likelihood function**.

Example 5.1 (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 5.3 (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 5.1), 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 5.1 (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 5.4 (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 5.2. *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 5.2. 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 5.3 (Additive property of the Info. mat.)**.** *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} Immediately obtained from the definition (see Def. @ref{def:Fisher}).
:::

Theorem 5.1 (Fréchet-Darmois-Cramér-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. 5.2 leads to the result. □

Definition 5.5. 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 5.6 (Maximum Likelihood Estimator (MLE))**.** The maximum likelihood estimator (MLE) is the vector θ that maximizes the likelihood function. Formally:

$$\theta_{MLE} = \arg\max_\theta \mathcal{L}(\theta;\mathbf{y}) = \arg\max_\theta \log \mathcal{L}(\theta;\mathbf{y}). \tag{5.4}$$

Definition 5.7 (Likelihood equation)**.** Necessary condition for maximizing the likelihood function:

$$\frac{\partial \log \mathcal{L}(\theta;\mathbf{y})}{\partial\theta} = \mathbf{0}. \tag{5.5}$$

Hypothesis 5.1 (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 $\mathbf{I}(\theta_0) = -\mathbb{E}_0\left(\frac{\partial^2 \log \mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'}\right)$ (Fisher Information matrix) exists and is nonsingular.

Proposition 5.4 (Properties of MLE)**.** *Under regularity conditions (Assumptions 5.1), the MLE is:*

a. ***Consistent****: plim* $\theta_{MLE} = \theta_0$ *(θ_0 is the true vector of parameters).*

b. ***Asymptotically normal****:* $\theta_{MLE} \sim \mathcal{N}(\theta_0, \mathbf{I}(\theta_0)^{-1})$*, where*

$$\mathbf{I}(\theta_0) = -\mathbb{E}_0\left(\frac{\partial^2 \log \mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'}\right) = n\mathcal{I}_Y(\theta_0). \quad \textit{(Fisher Info. matrix)}$$

c. ***Asymptotically efficient****:* θ_{MLE} *is asymptotically efficient and achieves the Fr'echet-Darmois-Cram'er-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 Online additional material. □

Note that (b) also writes:

$$\sqrt{n}(\theta_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y(\theta_0)^{-1}). \tag{5.6}$$

The asymptotic covariance matrix of the MLE 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:

$$\hat{\mathbf{I}}_1^{-1} = \left(-\frac{\partial^2 \log \mathcal{L}(\theta_{MLE};\mathbf{y})}{\partial\theta\partial\theta'}\right)^{-1}, \tag{5.7}$$

$$\hat{\mathbf{I}}_2^{-1} = \left(\sum_{i=1}^{n} \frac{\partial \log \mathcal{L}(\theta_{MLE};y_i)}{\partial\theta}\frac{\partial \log \mathcal{L}(\theta_{MLE};y_i)}{\partial\theta'}\right)^{-1}. \tag{5.8}$$

5.2.2 To sum up – MLE in practice

- A parametric model (depending on the vector of parameters θ whose "true" value is θ_0) is specified.
- 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. (5.4)):
$$\theta_{MLE} = \arg\max_{\theta} \log \mathcal{L}(\theta;\mathbf{y}). \tag{5.9}$$
- 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. (5.7) or Eq. (5.8). Most of the time, this computation is numerical.

5.2.3 Example: MLE estimation of a mixture of Gaussian distribution

Consider the returns of the SMI index. Let's 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 dist.)

The maximum likelihood estimate is $\theta_{MLE} = [0.30, 1.40, -1.45, 3.61, 0.87]'$.

The first two entries of the diagonal of $\hat{\mathbf{I}}_1^{-1}$ are 0.00528 and 0.00526. They are the estimates of $\mathbb{V}ar(\mu_{1,MLE})$ and of $\mathbb{V}ar(\sigma_{1,MLE})$, respectively.

95% confidence intervals for μ_1 and σ_1 are, respectively:

$$0.30 \pm 1.96\underbrace{\sqrt{0.00528}}_{=0.0726} \quad \text{and} \quad 1.40 \pm 1.96\underbrace{\sqrt{0.00526}}_{=0.0725}.$$

```
smi <- read.csv("https://raw.githubusercontent.com/jrenne/Data4courses/master/SMI/SMI.c
                dec = ".",header = TRUE, na.strings = "null")
smi$Date <- as.Date(smi$Date,"%m/%d/%y")
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))
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")
```

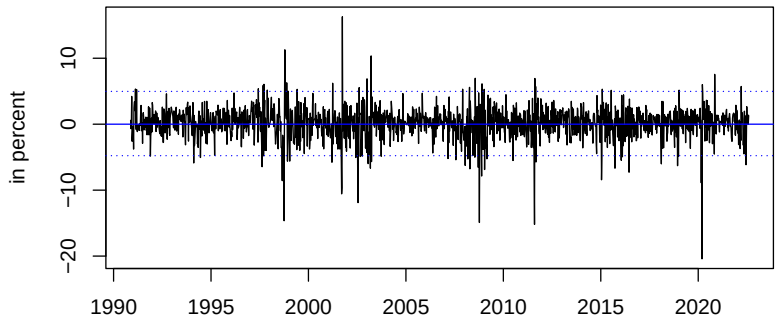


Figure 5.4: Time series of SMI weekly returns (source: Yahoo Finance).

```
f <- function(theta,y){ # Likelihood function
  mu.1 <- theta[1]; mu.2 <- theta[2]
  sigma.1 <- theta[3]; sigma.2 <- theta[4]
  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
```

Now, let us compute estimates of the covariance matrix of the MLE:

```
# Hessian approach:
J <- res.optim$hessian
I.1 <- solve(J)
# Outer-product of gradient approach:
log.f.0 <- log(f(theta,smi$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)
}
V <- t(d.log.f) %*% d.log.f
I.2 <- solve(t(d.log.f) %*% d.log.f)
# Misspecification-robust approach (sandwich formula):
I.3 <- solve(J) %*% V %*% solve(J)
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]).

```
x <- seq(-5,5,by=.01)
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, parametri
       lty=c(3,1,2), # gives the legend appropriate symbols (lines)
       lwd=c(2), # line width
       col=c("black","black","red"), # gives the legend lines the correct color and width
       pt.bg=c(1),
       pt.cex = c(1),
       bg="white",
       seg.len = 4
)
```

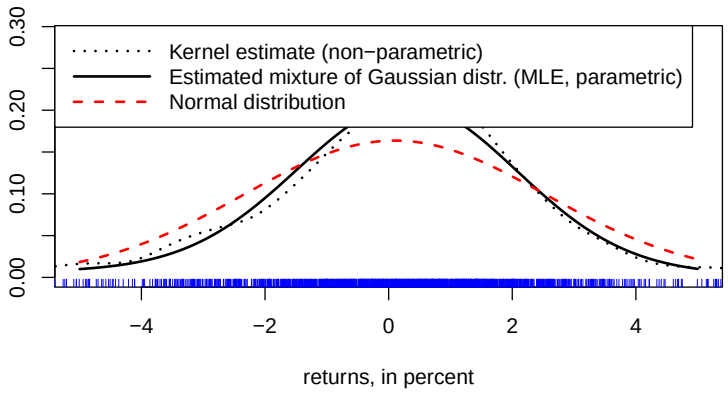


Figure 5.5: Comparison of different estimates of the distribution of returns.

Chapter 6

Microeconometrics

Chapter 7

Time Series

blabla

Chapter 8

Appendix

8.1 Statitical Tables

Table 8.1: 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 8.2: 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 8.3: 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 8.4: 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. 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).

Durbin, J. (1954). Errors in variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22(1/3):23–32.

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.

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.

MacKinnon, J. G., Ørregaard Nielsen, M., and Webb, M. D. (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415.

Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica*, 41(4):733–750.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2022). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.27.