Advanced Econometrics

Jean-Paul Renne

2022-09-09

Contents

Chapter 1

Before starting

This course covers various econometric topics, including linear regression models, discrete-choice models, and time series analysis. It provides examples or simulations based on R codes.

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

```
library(devtools)
install_github("jrenne/AEC")
library(AEC)
```

Useful (R) links:

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

Chapter 2

Linear Regressions

Definition 2.1. A linear regression is a model defined through:

$$y_i = \beta' \mathbf{x}_i + \varepsilon_i, \tag{2.1}$$

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

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

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

To have an intercept in the specification (2.1), one has to set $x_{i,1} = 1$ for all i; β_1 then corresponds to the intercept.

2.1 Hypotheses

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

Hypothesis 2.1 (Full rank)**.** There is no exact linear relationship among the independent variables (the $x_{i,k}$s, for a given $i \in \{1, \dots, n\}$).

Intuitively, when Hypothesis 2.1 is satisfied, then the estimation of the model parameters is unfeasible since, for any value of β, some changes in the explanatory variables will be exactly compensated by other changes in another set of explanatory variables, preventing the identification of these effects.

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

Hypothesis 2.2 (Conditional mean-zero assumption)**.**

$$\mathbb{E}(\varepsilon|\mathbf{X}) = 0. \tag{2.2}$$

The following proposition states some implications of Hypothesis 2.2:

Proposition 2.1. *Under Hypothesis 2.2:*

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

□

Let us now present two hypotheses (**??** and **??**) concerning the stochastic properties of the errors ε_i:

Hypothesis 2.3 (Homoskedasticity)**.**

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

The following lines of code generate a figure comparing two situations: Panel (a) of Figure 2.1 corresponds to a situation of homoskedasticity, and Panel (b) corresponds to a situation of heteroskedasticity. Let us be more specific. In the two plots, we have $X_i \sim \mathcal{N}(0,1)$ and $\varepsilon_i^* \sim \mathcal{N}(0,1)$. In Panel (a) (homoskedasticity): $Y_i = 2 + 2X_i + \varepsilon_i^*$. In Panel (b) (heteroskedasticity): $Y_i = 2 + 2X_i + \left(2\mathbb{1}_{\{X_i<0\}} + 0.2\mathbb{1}_{\{X_i\ge 0\}}\right)\varepsilon_i^*$.

Figure 2.2 shows a situation of heteroskedasticity, based on data taken from the Swiss Household Panel. The sample is restricted to persons younger than 35 year in 2019, and that have completed at least 19 years of study. The figure shows that the dispersion of yearly income increases with age.

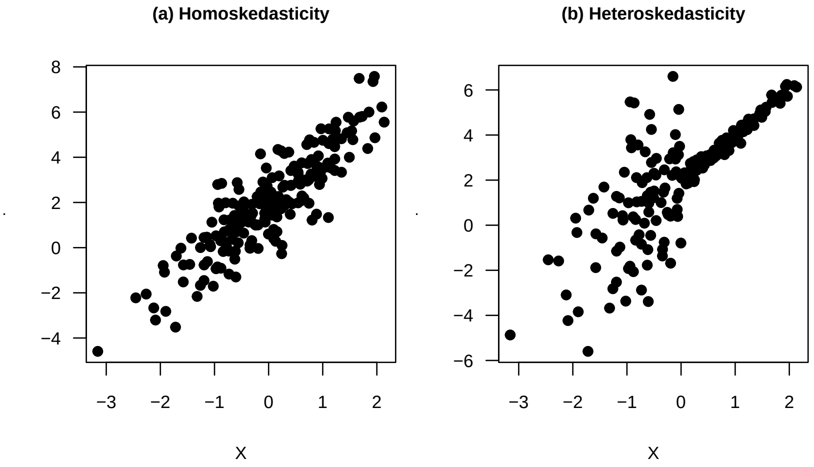


Figure 2.1: Homoskedasticity vs heteroskedasticity.

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

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

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

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

Hypothesis 2.4 (Uncorrelated errors)**.**

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

We will often work with covariance matrices. Proposition 2.2 give the specific form of the conditional covariance of the errors when both Hpoytheses 2.3 and 2.4 are satisfied.

Proposition 2.2. *If Hpoytheses 2.3 and 2.4 hold, then:*

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

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

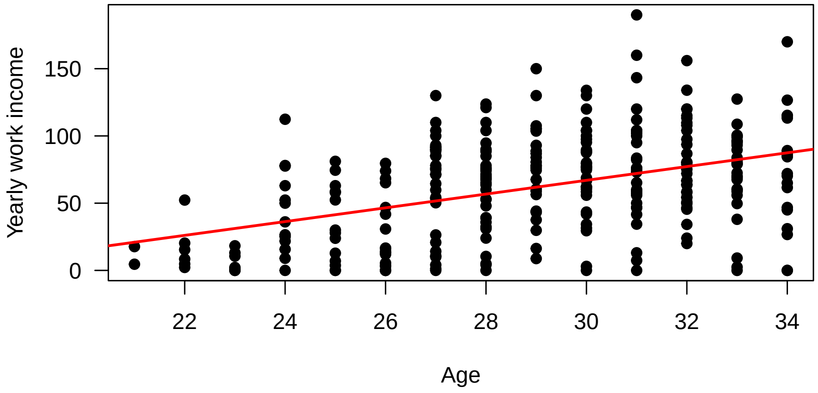


Figure 2.2: Income versus age. Data are from the Swiss Household Panel. The sample is restricted to persons that have completed at least 19 years of study. The figure shows that the dispersion of yearly income increases with age.

We will sometimes assume that errors are Gaussian — or normal. We will then work under Hypothesis 2.5:

Hypothesis 2.5 (Normal distribution)**.**

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

2.2 Least square estimation

2.2.1 Derivation of the OLS formula

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

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

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

Minimizing this sum amounts to minimizing:

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

Since:

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

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

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}. \tag{2.3}$$

Under Assumption 2.1, $\mathbf{X}'\mathbf{X}$ is invertible. Hence:

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

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

We have:

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

The estimated residuals are:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y} \tag{2.4}$$

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

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

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}, \tag{2.5}$$

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

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

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

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

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

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

Example 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:

$$
\begin{aligned}
\mathbf{X}'\mathbf{X} &= \left[\begin{array}{cc} n & \sum_i w_i \\ \sum_i w_i & \sum_i w_i^2 \end{array}\right], \\
(\mathbf{X}'\mathbf{X})^{-1} &= \frac{1}{n\sum_i w_i^2 - (\sum_i w_i)^2} \left[\begin{array}{cc} \sum_i w_i^2 & -\sum_i w_i \\ -\sum_i w_i & n \end{array}\right], \\
(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \frac{1}{n\sum_i w_i^2 - (\sum_i w_i)^2} \left[\begin{array}{c} \sum_i w_i^2 \sum_i y_i - \sum_i w_i \sum_i w_i y_i \\ -\sum_i w_i \sum_i y_i + n \sum_i w_i y_i \end{array}\right] \\
&= \frac{1}{\frac{1}{n}\sum_i (w_i - \bar{w})^2} \left[\begin{array}{c} \frac{\bar{y}}{n} \sum_i w_i^2 - \frac{\bar{w}}{n} \sum_i w_i y_i \\ \frac{1}{n} \sum_i (w_i - \bar{w})(y_i - \bar{y}) \end{array}\right].
\end{aligned}
$$

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 $b_1 = \bar{y} - b_2\bar{w}$.

2.2.2 Properties of the OLS estimate (small sample)

Proposition 2.3 states first properties of the OLS estimator:

Proposition 2.3 (Properties of the OLS estimator)**.** *We have:*

i. *Under Assumptions 2.1 and 2.2, the OLS estimator is linear and unbiased.*

ii. *Under Hypotheses 2.1 to 2.4, the conditional covariance matrix of* **b** *is:* $\mathbb{V}ar(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Proof. Under Hypothesis 2.1, $\mathbf{X}'\mathbf{X}$ 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 2.2, 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. 2.2, if 2.3 and 2.4 hold, then we have $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \mathbb{V}ar(\varepsilon|\mathbf{X}) = \sigma^2 Id$.

□

Together, Hypotheses 2.1 to 2.4 form the Gauss-Markov set of assumptions. Under these assumptions, the OLS estimator feature the lowest possible variance:

Theorem 2.1 (Gauss-Markov Theorem)**.** *Under Assumptions 2.1 to 2.4, 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 2.2).

Therefore $\mathbf{b}^*$ is unbiased if $\mathbb{E}(C\mathbf{X})\beta = \beta$. This has to be the case for any β, which implies that we must have $C\mathbf{X} = \mathbf{I}$.

Let us compute $\mathbb{V}ar(\mathbf{b}^*|\mathbf{X})$. For this, we introduce $D = C - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which is such that $D\mathbf{y} = \mathbf{b}^* - \mathbf{b}$. The fact that $C\mathbf{X} = \mathbf{I}$ implies that $D\mathbf{X} = \mathbf{0}$.

We have $\mathbb{V}ar(\mathbf{b}^*|\mathbf{X}) = \mathbb{V}ar(C\mathbf{y}|\mathbf{X}) = \mathbb{V}ar(C\varepsilon|\mathbf{X}) = \sigma^2 CC'$ (by Assumptions 2.3 and 2.4, see Prop. 2.2). 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 $\mathbb{V}ar(w'\mathbf{b}^*|\mathbf{X}) = w'\mathbb{V}ar(\mathbf{b}|\mathbf{X})w + \sigma^2 w'\mathbf{D}\mathbf{D}'w \geq w'\mathbb{V}ar(\mathbf{b}|\mathbf{X})w = \mathbb{V}ar(w'\mathbf{b}|\mathbf{X})$. □

The Frish-Waugh theorem (Theorem 2.2) reveals the relationship between the OLS estimator and the notion of partial correlation coefficient. Consider the linear least square regression of $\mathbf{y}$ on $\mathbf{X}$. We introduce the notations:

- $\mathbf{b}^{\mathbf{y}/\mathbf{X}}$: OLS estimates of β,
- $\mathbf{M}^{\mathbf{X}}$: residual-maker matrix of any regression on $\mathbf{X}$,
- $\mathbf{P}^{\mathbf{X}}$: projection matrix of any regression on $\mathbf{X}$.

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

Theorem 2.2 (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. (2.3)):

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

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

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

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 residuals associated with the former regression on the ones associated withe the latter regressions.

This is illustrated by the following code, where we run different regressions involving the number of Google searches for "parapluie" (*umbrella* in French) In the broad specification, we regress it on French precipitations and month dummies. Next, we deseasonalize both the dependent variable and the precipitations by regressing them on the month dummies. As stated by Theorem 2.2, regressing deseasonalized Google searches on deseasonalized precipitations give the same coefficient as in the baseline regression.

```r
library(AEC)
dummies <- as.matrix(parapluie[,4:14])
eq_all <- lm(parapluie~precip+dummies,data=parapluie)
deseas_parapluie <- lm(parapluie~dummies,data=parapluie)$residuals
deseas_precip    <- lm(precip~dummies,data=parapluie)$residuals
eq_frac <- lm(deseas_parapluie~deseas_precip)
rbind(eq_all$coefficients,
      c(eq_frac$coefficients,rep(NaN,11)))
```

```
##        (Intercept)    precip  dummies1  dummies2  dummies3  dummies4  dummies5
## [1,] -2.953448e+00 0.1300055 -8.599068 -13.32904 -7.982958 -3.392353 -3.703816
## [2,] -1.038345e-15 0.1300055       NaN       NaN       NaN       NaN       NaN
##       dummies6  dummies7  dummies8 dummies9 dummies10 dummies11
## [1,] -3.360641 -7.315881 -7.717277  -4.6492 -5.109199   1.98077
## [2,]       NaN       NaN       NaN      NaN       NaN       NaN
```

When b_2 is scalar (and then $\mathbf{X}_2$ is of dimension $n \times 1$), Theorem 2.2 gives the expression of the **partial regression coefficient** b_2:

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

2.2.3 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:

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

We can now define the coefficient of determination:

$$\boxed{\text{Coefficient of determination} = \frac{Expl.SS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{M}^0\mathbf{y}}.} \qquad (2.6)$$

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

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

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

The hgher the R^2, the higher the goodness of fit of a model. One however has to be cautious with R^2. Indeed, it is easy to increase it: it suffices to add explanatory variables. As stated by Proposition 2.5, adding an explanatory variable (even if it does not truly relate to the dependent variable) results in an increase in the R^2. In the limit, taking any set of n non-linearly-dependent explanatory variables (i.e., variables satisfying Hypothesis 2.1) results in a R^2 equal to one.

Proposition 2.4 (Change in SSR when a variable is added)**.** *We have:*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}^{*\prime}\mathbf{z}^*) \qquad (\leq \mathbf{e}'\mathbf{e}) \qquad (2.7)$$

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. (2.3))

$$\left[\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{z} \\ \mathbf{z}'\mathbf{X} & \mathbf{z}'\mathbf{z} \end{array}\right]\left[\begin{array}{c} \mathbf{d} \\ \mathbf{c} \end{array}\right] = \left[\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{z}'\mathbf{y} \end{array}\right].$$

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:

$$\mathbf{u}'\mathbf{u} = (\mathbf{e} - \mathbf{z}^*c)(\mathbf{e} - \mathbf{z}^*c) = \mathbf{e}'\mathbf{e} + c^2(\mathbf{z}^{*\prime}\mathbf{z}^*) - 2c\mathbf{z}^{*\prime}\mathbf{e}. \qquad (2.8)$$

Now $\mathbf{z}^{*\prime}\mathbf{e} = \mathbf{z}^{*\prime}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{z}^{*\prime}\mathbf{y}$ because $\mathbf{z}^*$ are the residuals in an OLS regression on $\mathbf{X}$. Since $c = (\mathbf{z}^{*\prime}\mathbf{z}^*)^{-1}\mathbf{z}^{*\prime}\mathbf{y}^*$ (by an application of Theorem 2.2), we have $(\mathbf{z}^{*\prime}\mathbf{z}^*)c = \mathbf{z}^{*\prime}\mathbf{y}^*$ and, therefore, $\mathbf{z}^{*\prime}\mathbf{e} = (\mathbf{z}^{*\prime}\mathbf{z}^*)c$. Inserting this in Eq. (2.8) leads to the results. □

Proposition 2.5 (Change in the coefficient of determination when a variable is added)**.** *Denoting by R^2_W the coefficient of determination in the regression of* $\mathbf{y}$ *on some variable* $\mathbf{W}$*, we have:*

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

where $r^{\mathbf{X}}_{yz}$ is the coefficient of partial correlation (see Definition 3.1).

Proof. Let's use the same notations as in Prop. @ref{prp:chgeR2}. Theorem 2.2 implies that $c = (\mathbf{z}^{*\prime}\mathbf{z}^*)^{-1}\mathbf{z}^{*\prime}\mathbf{y}^*$. Using this in Eq. (2.7) gives $\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - (\mathbf{z}^{*\prime}\mathbf{y}^*)^2/(\mathbf{z}^{*\prime}\mathbf{z}^*)$. Using the definition of the partial correlation (Eq. (3.1)), we get $\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e}\left(1 - (r^{\mathbf{X}}_{yz})^2\right)$. The results is obtained by dividing both sides of the previous equation by $\mathbf{y}'\mathbf{M}_0\mathbf{y}$. □

Figure 2.3, below, illustrates the fact that one can obtain an R^2 of one by regressing a sample of length n on any set of n linearly-independent variables.

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

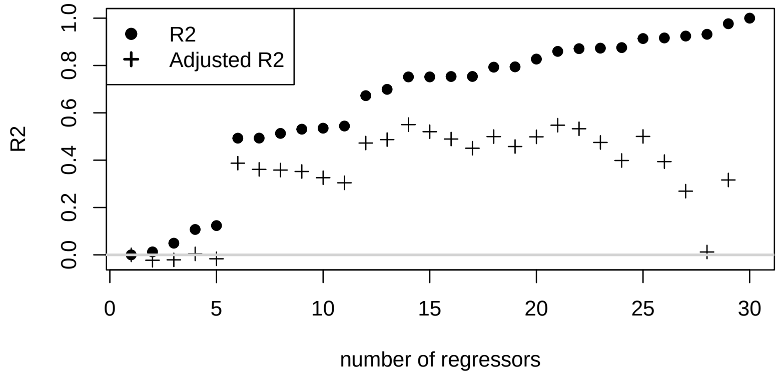


Figure 2.3: This figure illustrates the monotonous increase in the R^2 as a function of the number of explanatory variables. In the true model, there is no explanatory variables, i.e., $y_i = \varepsilon_i$. We then take (independent) regressors and regress y on the latter, progresively increasing the set of regressors.

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

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

2.2.4 Inference and confidence intervals (in small sample)

Under the normality assumption (Assumption 2.5), 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:

$$\mathbf{b}|\mathbf{X} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \tag{2.9}$$

Eq. (2.9) can be used to conduct inference and tests. However, in practice, we do not know σ^2 (which is a population parameter). The following proposition gives an unbiased estimate of σ^2.

Proposition 2.6. *Under 2.1 to 2.4, an unbiased estimate of σ^2 is given by:*

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}. \tag{2.10}$$

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

Proof. $\mathbb{E}(\mathbf{e}'\mathbf{e}|\mathbf{X}) = \mathbb{E}(\varepsilon'\mathbf{M}\varepsilon|\mathbf{X}) = \mathbb{E}(\text{Tr}(\varepsilon'\mathbf{M}\varepsilon)|\mathbf{X})) = \text{Tr}(\mathbf{M}\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})) = \sigma^2\text{Tr}(\mathbf{M})$. (Note that we have $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2 Id$ by Assumptions 2.3 and 2.4, see Prop. 2.2.) Finally:

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

which leads to the result. □

Two results will prove important to produce inference:

i. We know the distribution of s^2 (Prop. 2.7).
ii. s^2 and $\mathbf{b}$ are independent random variables (Prop. 2.8).

Proposition 2.7. *Under 2.1 to 2.5, 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 2.5— and uncorrelated). The variance of each of these i.i.d. Gaussian variable is σ^2. Because $\mathbf{M}$ is an idempotent symmetric matrix, its eigenvalues are either 0 or 1, and its rank equals its trace (see Propositions 3.3 and 3.4). Further, its trace is equal to $n-K$ (see proof of Eq. (2.10)). 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 2.8. *Under Hypotheses 2.1 to 2.5,* $\mathbf{b}$ *and* s^2 *are independent.*

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

Consistently with Eq. (2.9), under Hypotheses 2.1 to 2.5, the k^{th} entry of $\mathbf{b}$ satisfies:

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

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

Moreover, we have (Prop. 2.7):

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

As a result (using Propositions 2.7 and 2.8), we have:

$$\boxed{t_k = \frac{\frac{b_k-\beta_k}{\sqrt{\sigma^2 v_k}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2(n-K)}}} = \frac{b_k-\beta_k}{\sqrt{s^2 v_k}} \sim t(n-K),} \tag{2.11}$$

where $t(n-K)$ denotes a t distribution with $n-K$ degrees of freedom.[1]

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. 2.6).

The previous result (Eq. (2.11)) can be extended to any linear combinations of elements of $\mathbf{b}$ (Eq. (2.11) is for its k^{th} component only).

Let us consider $\alpha'\mathbf{b}$, the OLS estimate of $\alpha'\beta$. From Eq. (2.9), 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. (2.11), one can show that Props. 2.7 and 2.8 also imply that:

$$\boxed{\frac{\alpha'\mathbf{b}-\alpha'\beta}{\sqrt{s^2\alpha'(\mathbf{X}'\mathbf{X})^{-1}\alpha}} \sim t(n-K).} \tag{2.12}$$

What precedes is widely exploited for statistical inference in the context of linear regressions. Indeed, Eq. (2.11) gives a sense of the distances between b_k and β_k that can be deemed as "likely". For instance, it implies that, if $\sqrt{v_k s^2}$ is equal to 1 (say), then the probability to obtain b_k smaller than β_k− 4.587 ×$\sqrt{v_k s^2}$ or larger than β_k+ 4.587 ×$\sqrt{v_k s^2}$ is equal to 0.1% when $n-K=10$.

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

$$H_0 : \beta_k = 0.$$

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

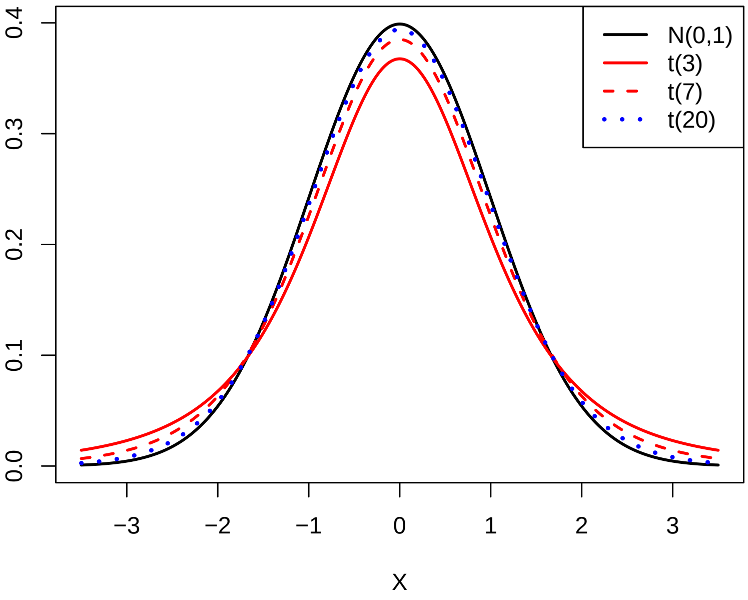


Figure 2.4: The higher the degree of freedom, the closer the distribution of $t(\nu)$ gets to the normal distribution. (Convergence in distribution.)

Under the null hypothesis, the test statistic follows a Student-t distribution with $n-K$ degrees of freedom. The **t-statistic** is therefore of particular importance, and, as a result, it is routinely reported in regression outputs. This is illustrated below, in a regression that aims at determining covariates of households' income. This example makes use of data from the Swiss Household Panel (SHP). `edyear19` is the number of years of education and `age19` is the age of the respondent, as of 2019.

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

The last two columns give the **t-statistic** and p-values associated with the t-test, whose **critical region** for the test of size α is:

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

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

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

For this purpose, we make use of Eq. (2.11), i.e., $t_k=\frac{b_k-\beta_k}{\sqrt{s^2v_k}}\sim t(n-K)$. We have:

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

where $\Phi_{t(n-K)}(\alpha)$ is the c.d.f. of the $t(n-K)$ distribution (Table 3.2).

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

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

Using the results of the previous regression, we compute lower and upper bounds of 95/% confidence intervals for the estimated parameters as follows:

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

```
##                 lower.b      upper.b
## (Intercept) -74.8848532 -69.06276141
## edyear19      4.7334839   4.95504836
## age19         3.1272532   3.34998983
## I(age19^2)   -0.0300164  -0.02788323
## female      -32.5523381 -31.06546311
```

2.2.5 Testing a set of linear restrictions

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

$$\begin{array}{rcl} r_{1,1}\beta_1 + \dots + r_{1,K}\beta_K &=& q_1 \\ \vdots && \vdots \\ r_{J,1}\beta_1 + \dots + r_{J,K}\beta_K &=& q_J. \end{array} \tag{2.13}$$

In matrix form, we get:

$$\mathbf{R}\beta = \mathbf{q}. \tag{2.14}$$

Define the **Discrepancy vector** $\mathbf{m} = \mathbf{R}\mathbf{b} - \mathbf{q}$. Under the null hypothesis:

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

With these notations, the assumption to test is:

$$\boxed{H_0 : \mathbf{R}\beta - \mathbf{q} = 0 \text{ against } H_1 : \mathbf{R}\beta - \mathbf{q} \neq 0.} \tag{2.15}$$

Under Hypotheses 2.1 to 2.4, we have $\mathbb{V}ar(\mathbf{m}|\mathbf{X}) = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ (see Prop. 2.3). If we add the normality assumption (Hypothesis 2.5), we have:

$$W = \mathbf{m}'\mathbb{V}ar(\mathbf{m}|\mathbf{X})^{-1}\mathbf{m} \sim \chi^2(J). \tag{2.16}$$

If σ^2 was known, we could then conduct a **Wald test**. But this is not the case in practice and we cannot compute W. We can, however, approximate it be replacing σ^2 by s^2. The distribution of this new statistic is not $\chi^2(J)$ any more; it is an $\mathcal{F}$ **distribution** (whose quantiles are shown in Table 3.4), and the test is called F **test**.

Proposition 2.9. *Under Hypotheses 2.1 to 2.5 and if Eq.* (2.15) *holds, we have:*

$$F = \frac{W}{J}\frac{\sigma^2}{s^2} = \frac{\mathbf{m}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{m}}{s^2 J} \sim \mathcal{F}(J, n-K), \qquad (2.17)$$

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

Proof. According to Eq. (2.16), $W/J \sim \chi^2(J)/J$. 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 2.1 to 2.4, the covariance between $\mathbf{T}\varepsilon$ and $\mathbf{M}\varepsilon$ is $\sigma^2\mathbf{T}\mathbf{M} = \mathbf{0}$. Therefore, under 2.5, these variables are Gaussian variables with 0 covariance. Hence they are independent. □

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

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

Proposition 2.10. *The F-statistic defined by Eq.* (2.17) *is also equal to:*

$$F = \frac{(R^2 - R_*^2)/J}{(1-R^2)/(n-K)} = \frac{(SSR_{restr} - SSR_{unrestr})/J}{SSR_{unrestr}/(n-K)}, \qquad (2.18)$$

where R_*^2 is the coef. of determination (Eq. (2.6)*) of the "restricted regression"* (SSR: sum of squared residuals.)

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

By Proposition 3.5 (in Appendix **??**), we have: $\mathbf{b}_* {-} \mathbf{b} = -(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\}^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})$. Therefore:

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

This implies that the F statistic defined in Prop. 2.9 is also equal to:

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

which leads to the result. □

The null hypothesis H_0 (Eq. (2.15)) of the F-test is rejected if F –defined by Eq. (2.17) or (2.18)– is higher than $\mathcal{F}_{1-\alpha}(J, n-K)$. (Hence, this test is a one-sided test.)

2.2.6 Large Sample Properties

Even if we relax the normality assumption (Hypothesis 2.5), we can approximate the finite-sample behavior of the estimators by using *large-sample* or *asymptotic properties.*

To begin with, we proceed under Hypothesis 2.1 to 2.4. (We will see, later on, how to deal with –partial– relaxations of Hypothesis 2.3 and 2.4.)

Under regularity assumptions, and under Hypotheses 2.1 to 2.4, 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 Hypotheses 2.1 to 2.5 hold. This derives from Prop. 2.11 (below). The F-test (Prop. 2.10) and the t-test (Eq. (2.11)) can then be performed.

Proposition 2.11. *Under Assumptions 2.1 to 2.4, and assuming further that:*

$$Q = plim_{n\rightarrow\infty}\frac{\mathbf{X}'\mathbf{X}}{n}, \tag{2.19}$$

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

$$\sqrt{n}(\mathbf{b}-\beta)\overset{d}{\rightarrow}\mathcal{N}\left(0,\sigma^2Q^{-1}\right). \tag{2.20}$$

Proof. Since $\mathbf{b}=\beta+\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\varepsilon}{n}\right)$, we have: $\sqrt{n}(\mathbf{b}-\beta)=\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon$. Since $f:A\rightarrow A^{-1}$ is a continuous function (for $A\neq\mathbf{0}$), $\text{plim}_{n\rightarrow\infty}\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}=\mathbf{Q}^{-1}$ (see Prop. 3.7). 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 $\sigma^2\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')=\sigma^2Q$. Applying the multivariate central limit theorem on vectors V_i gives $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\varepsilon_i\right)=\left(\frac{1}{\sqrt{n}}\right)\mathbf{X}'\varepsilon\overset{d}{\rightarrow}\mathcal{N}(0,\sigma^2Q)$. An application of Slutsky's theorem (Prop. 3.7) then leads to the results. □

In practice, σ^2 is estimated with $\frac{\mathbf{e}'\mathbf{e}}{n-K}$ (Eq. (2.10)) and $\mathbf{Q}^{-1}$ with $\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$. That is, the covariance matrix of the estimator is approximated by:

$$\widehat{\mathbb{V}ar}(\mathbf{b})=s^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{2.21}$$

Eqs. (2.19) and (2.20) respectively correspond to convergences in probability and in distribution (see Definitions 3.9 and 3.12, respectively).

2.3 Common pitfalls in linear regressions

2.3.1 Multicollinearity

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

$$\mathbf{X}'\mathbf{X} = \left[\begin{array}{cc} \sum_i x_{i,1}^2 & \sum_i x_{i,1}x_{i,2} \\ \sum_i x_{i,1}x_{i,2} & \sum_i x_{i,2}^2 \end{array}\right],$$

therefore:

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum_i x_{i,1}^2 \sum_i x_{i,2}^2 - (\sum_i x_{i,1}x_{i,2})^2} \left[\begin{array}{cc} \sum_i x_{i,2}^2 & -\sum_i x_{i,1}x_{i,2} \\ -\sum_i x_{i,1}x_{i,2} & \sum_i x_{i,1}^2 \end{array}\right].$$

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

$$\sum_i x_{i,1}^2 - \frac{(\sum_i x_{i,1}x_{i,2})^2}{\sum_i x_{i,2}^2} = \sum_i x_{i,1}^2(1 - correl_{1,2}^2), \tag{2.22}$$

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

Hence, the closer to one $correl_{1,2}$, the higher the variance of b_1 (recall that the variance of b_1 is the upper-left component of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$).

2.3.2 Omitted variables

Consider the following model (the “True model”):

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

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

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

This results in the omitted-variable formula:

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

(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 $\mathbf{X}_1$). Unless the variables included in $\mathbf{X}_1$ are orthogonal to those in $\mathbf{X}_2$, we obtain a bias.

Example 2.2. 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 following regressions show that the effect is lower when controls are added.

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

```
##
## ====================================================================================
##                                                 Dependent variable:
##                      ----------------------------------------------------------------
##                                                      testscr
##                               (1)                     (2)                      (3)
## ------------------------------------------------------------------------------------
## str                        -2.280***               -1.117***                -0.998**
##                             (0.480)                 (0.240)                  (0.239)
## lunch                                              -0.600***                -0.547**
##                                                     (0.017)                  (0.022)
## english                                                                     -0.122**
##                                                                              (0.032)
## Constant                  698.933***               702.911***               700.150*
##                             (9.467)                 (4.700)                  (4.686)
## ------------------------------------------------------------------------------------
## Observations                  420                     420                      420
## R2                           0.051                   0.767                    0.775
## Adjusted R2                  0.049                   0.766                    0.773
## Residual Std. Error    18.581 (df = 418)        9.222 (df = 417)         9.080 (df =
## F Statistic         22.575*** (df = 1; 418) 685.756*** (df = 2; 417) 476.306*** (df
## ====================================================================================
## Note:                                                             *p<0.1; **p<0.05;
```

2.3.3 Irrelevant variable

Consider the *True model*:

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

while the *Estimated model* is:

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

The estimates are unbiased. However, adding irrelevant explanatory variables increases the variance of the estimate of β_1 (compared to the case where one uses the correct explanatory variables). This is the case unless the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ is null, see Eq. (2.22).

In other words, the estimator is *inefficient*, i.e., there exists an alternative consistent estimator whose variance is lower. The inefficiency problem can have serious consequences when testing hypotheses 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).

2.4 Instrumental Variables

Nice of interpretation of tests

The conditional mean zero assumption (Hypothesis 2.2), according to which $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$ —which implies in particular that $\mathbf{x}_i$ and ε_i are uncorrelated— is sometimes not consistent with the considered economic framework. When it is the case, the parameters of interest may still be estimated consistently by resorting to instrumental variable techniques.

Consider the following model:

$$y_i = \mathbf{x_i}'\beta + \varepsilon_i, \quad \text{where } \mathbb{E}(\varepsilon_i) = 0 \text{ and } \mathbf{x_i} \not\perp \varepsilon_i. \tag{2.23}$$

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

$$\mathbb{E}(\varepsilon_i) = 0 \quad \text{and} \quad \mathbb{E}(\varepsilon_i \mathbf{x_i}) = \gamma, \tag{2.24}$$

in which case we have $\mathbf{x}_i \not\perp \varepsilon_i$ (consistently with Eq. (2.23)).

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

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

Let us now introduce the notion of instruments.

Definition 2.2 (Instrumental variables)**.** The L-dimensional random variable $\mathbf{z}_i$ is a valid set of instruments if:

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

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, $\frac{\mathbf{Z}'\varepsilon}{n} \overset{p}{\to} \mathbb{E}(\mathbf{z}_i\varepsilon_i) = 0$.

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 (Prop. 3.7) 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 2.12 (Asymptotic distribution of the IV estimator)**.** *If $\mathbf{z}_i$ is a L-dimensional random variable that constitutes a valid set of instruments (see Def. 2.2) and if $L = K$, then the asymptotic distribution of $\mathbf{b}_{iv}$ is:*

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

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. 2.11, the starting point being that $\mathbf{b}_{iv} = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\varepsilon$. □

When $L = K$, we have:

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

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

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

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

1. Regress $\mathbf{X}$ on the space spanned by $\mathbf{Z}$ and

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

The results in:

$$\boxed{\mathbf{b}_{iv} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}.} \tag{2.25}$$

In this case, Prop. 2.12 still holds, with $\mathbf{b}_{iv}$ given by Eq. (2.25).

$\mathbf{b}_{iv}$ 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. (2.5)). Hence the other names of this estimator: **Two-Stage Least Squares (TSLS)**.

If the instruments do not properly satisfy Condition (a) in Def. 2.2 (i.e. if $\mathbf{x}_i$ and $\mathbf{z}_i$ are only loosely related), the instruments are said to be **weak** (see, e.g., Stock and Yogo (2005), available here or Andrews et al. (2019)). A simple standard way to test for weak instruments consist in looking at the F-statistic associated with the first stage of the estimation. The easier it is to reject the null hypothesis (large test statistic), the less weak the instruments.

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

The test statistic is:

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

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

Example 2.3 (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:

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

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. (2.24)).

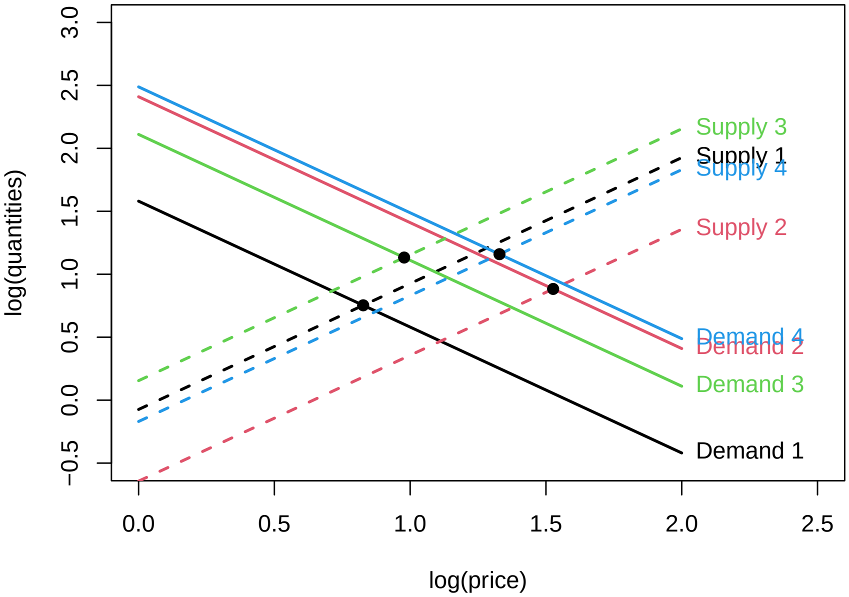


Figure 2.5: This figure illustrates the situation prevailing when estimating a price-elasticity (and the price is endogenous).

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

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

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

```
##
## ==========================================================================
##                                      Dependent variable:
##                     ------------------------------------------------------
##                                           log(packs)
##                              OLS                    instrumental
##                                                       variable
##                              (1)                 (2)             (3)
## --------------------------------------------------------------------------
## log(rprice)               -1.407***           -1.277***       -1.084***
##                            (0.251)             (0.263)         (0.317)
## log(rincome)                0.344               0.280
##                            (0.235)             (0.239)
## Constant                  10.342***           9.895***        9.720***
##                            (1.023)             (1.059)         (1.514)
## --------------------------------------------------------------------------
## Observations                 48                  48              48
## R2                          0.433               0.429           0.401
## Adjusted R2                 0.408               0.404           0.388
## Residual Std. Error    0.187 (df = 45)     0.188 (df = 45) 0.190 (df = 46)
## F Statistic         17.165*** (df = 2; 45)
## ==========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

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

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

Example 2.4 (Education and wage)**.** In this example, we make use of another dataset proposed by Stock and Watson (2003), namely the `CollegeDistance` dataset.[2] the objective is to estimate the effect of education on wages. Education

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

choice is suspected to be an endogenous variable, which calls for an IV strategy. The instrumental variable is the distance to college.

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

```
##
## ===========================================================================
##                                             Dependent variable:
##                                 -------------------------------------------
##                                 education               wage
##                                    OLS       OLS     instrumental    OLS
##                                                        variable
##                                    (1)       (2)         (3)         (4)
## ---------------------------------------------------------------------------
## urbanyes                         -0.092     0.046       0.046       0.070
##                                  (0.065)   (0.045)     (0.060)     (0.045)
## genderfemale                     -0.025    -0.071*      -0.071    -0.085**
##                                  (0.052)   (0.037)     (0.050)     (0.037)
## ethnicityafam                   -0.524*** -0.227***    -0.227**   -0.556***
##                                  (0.072)   (0.073)     (0.099)     (0.052)
## ethnicityhispanic               -0.275*** -0.351***   -0.351***   -0.544***
##                                  (0.068)   (0.057)     (0.077)     (0.049)
## unemp                             0.010    0.139***    0.139***   0.133***
##                                  (0.010)   (0.007)     (0.009)     (0.007)
## distance                        -0.087***
##                                  (0.012)
## ed.pred                                    0.647***
##                                            (0.101)
## education                                              0.647***     0.005
##                                                        (0.136)     (0.010)
## Constant                        14.061***   -0.359      -0.359    8.641***
##                                  (0.083)   (1.412)     (1.908)     (0.157)
## ---------------------------------------------------------------------------
## Observations                       4,739     4,739      4,739      4,739
## R2                                 0.023     0.117      -0.612     0.110
## Adjusted R2                        0.022     0.116      -0.614     0.109
## Residual Std. Error (df = 4732)    1.770     1.263      1.706      1.268
## F Statistic (df = 6; 4732)       18.552*** 104.971***             97.274***
## ==========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

2.5 General Regression Model (GRM) and robust covariance matrices

The statistical inference presented above relies on strong assumptions regarding the stochastic properties of the errors. Namely, they are assumed to be mutually uncorrelated (Hypothesis @ref(hyp:noncorrel_resid)) and homoskedastic (Hypothesis 2.3.

The objective of this section is to present approaches aimed at adjusting the estimate of the covariance matrix of the OLS estimator ($(\mathbf{X}'\mathbf{X})^{-1}s^2$, see Eq. (2.21)), when the previous hypotheses do not hold.

2.5.1 Presentation of the General Regression Model (GRM)

It will prove useful to introduce the following notation:

$$\mathbb{V}ar(\varepsilon|\mathbf{X}) = \mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) \quad = \quad \Sigma. \tag{2.26}$$

Note that Eq. ((2.26)) is more general than Hypothesis 2.3 and @ref(hyp:noncorrel_resid) because the diagonal entries of Σ may be different (as opposed to under Hypothesis 2.3), and the non-diagonal entries of Σ can be non-null (as opposed to under Hypothesis 2.4).

Definition 2.3 (General Regression Model (GRM))**.** Hypothesis 2.1 and 2.2, together with Eq. (2.26), form the General Regression Model (GRM) framework.

Naturally, a regression model where Hypotheses 2.1 to 2.4 hold is a specific case of the GRM framework.

The GRM context notably encompasses situations of heteroskedasticity and autocorrelation:

- Heteroskedasticity:

$$\Sigma = \left[\begin{array}{cccc} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_n^2 \end{array}\right]. \tag{2.27}$$

- Autocorrelation:

$$\Sigma = \sigma^2 \left[\begin{array}{cccc} 1 & \rho_{2,1} & \dots & \rho_{n,1} \\ \rho_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \rho_{n,n-1} \\ \rho_{n,1} & \rho_{n,2} & \dots & 1 \end{array}\right]. \tag{2.28}$$

Example 2.5 (Auto-regressive processes)**.** Autocorrelation is, in particular, a recurrent problem when time-series data are used (see Section **??**).

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

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i \tag{2.29}$$

with

$$\varepsilon_i = \rho\varepsilon_{i-1} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma_v^2). \tag{2.30}$$

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

$$\Sigma = \frac{\sigma_v^2}{1-\rho^2} \left[\begin{array}{cccc} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{array}\right]. \tag{2.31}$$

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

2.5.2 Generalized Least Squares

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

We have $\Sigma = (\mathbf{PP}')^{-1}$ 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 2.1).

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

$$\boxed{\mathbf{b}_{GLS} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}}. \tag{2.32}$$

We have:

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

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

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

Example 2.6 (GLS in the auto-correlation case)**.** Consider the case presented in Example 2.5. 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. (2.31) provides a consistent estimate of Σ. Applying these steps recursively gives an efficient estimator of β (Cochrane and Orcutt (1949)).

2.5.3 Asymptotic properties of the OLS estimator in the GRM framework

In the GRM framework, we have:

$$\mathbb{V}ar(\mathbf{b}|\mathbf{X}) = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{2.33}$$

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

Proposition 2.13 (Consistency of the OLS estimator in the GRM framework)**.** *If plim $(\mathbf{X}'\mathbf{X}/n)$ and plim $(\mathbf{X}'\Sigma\mathbf{X}/n)$ are finite positive definite matrices, then plim $(\mathbf{b}) = \beta$.*

Proof. We have $\mathbb{V}ar(\mathbf{b}) = \mathbb{E}[\mathbb{V}ar(\mathbf{b}|\mathbf{X})] + \mathbb{V}ar[\mathbb{E}(\mathbf{b}|\mathbf{X})]$. Since $\mathbb{E}(\mathbf{b}|\mathbf{X}) = \beta$, $\mathbb{V}ar[\mathbb{E}(\mathbf{b}|\mathbf{X})] = 0$. Eq. (2.33) implies that $\mathbb{V}ar(\mathbf{b}|\mathbf{X}) \to 0$. Hence $\mathbf{b}$ converges in mean square, and therefore in probability (see Prop. **??**). □

Prop. 2.14 gives the asymptotic distribution of the OLS estimator in the GRM framework.

Proposition 2.14 (Asymptotic distribution of the OLS estimator in the GRM framework)**.** *If $Q_{xx} = plim\ (\mathbf{X}'\mathbf{X}/n)$ and $Q_{x\Sigma x} = plim\ (\mathbf{X}'\Sigma\mathbf{X}/n)$ are finite positive definite matrices, then:*

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

The IV estimator also features a normal asymptotic distribution:

Proposition 2.15 (Asymptotic distribution of the OLS estimator in the GRM framework)**.** *If regressors and IV variables are "well-behaved", then:*

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

where

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

with

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

For practical purposes, one needs to have estimates of Σ in Props. 2.14 or 2.15. The complication comes from the fact that Σ is of dimension $n \times n$, and its estimation —based on a sample of length n— is therefore infeasible in the general case. Notwithstanding, looking at Eq. (2.33), it appears that one can focus on the estimation of $Q_{x\Sigma x} = \text{plim}\ (\mathbf{X}'\Sigma\mathbf{X}/n)$ (or $\text{plim}\ \left(\frac{1}{n}\mathbf{Z}'\Sigma\mathbf{Z}\right)$ in the IV case). This matrix being of dimension $K \times K$, its estimation is easier.

We have:

$$\frac{1}{n}\mathbf{X}'\Sigma\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'. \tag{2.34}$$

The so-called **robust covariance matrices** are estimates of the previous matrix. Their computation is based on the fact that if $\mathbf{b}$ is consistent, then the e_i's are consistent (pointwise) estimators of the ε_i's. Let us present such robust covariance matrices in two basic situations: heteroskedasticity (Example 2.7) and auto-correlation of the residuals (Example **??**)

Example 2.7 (Heteroskedasticity)**.** This is the case of Eq. (2.27).

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:

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{x}_i\mathbf{x}_i'\right). \tag{2.35}$$

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

$$\frac{1}{n}\mathbf{X}'\mathbf{E}^2\mathbf{X}, \tag{2.36}$$

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

Illustration: Figure **??**.

Let us illustrate the influence of heteroskedasticity using simulations.

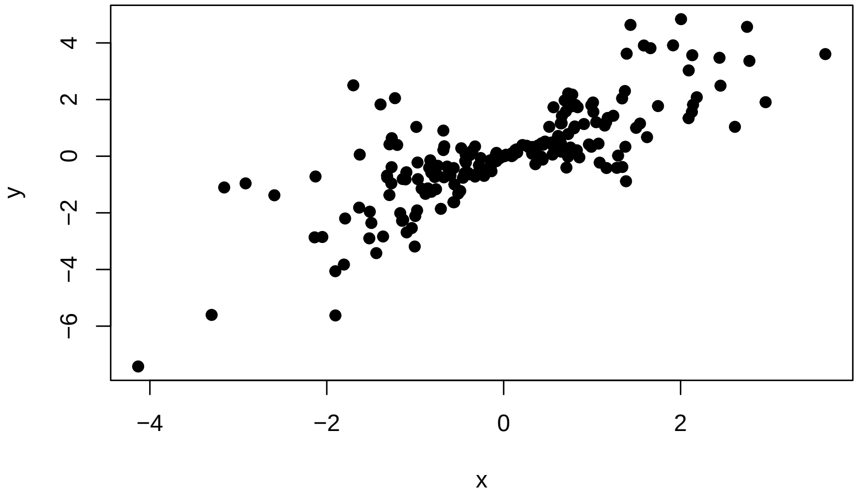
We consider the following model:

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

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).

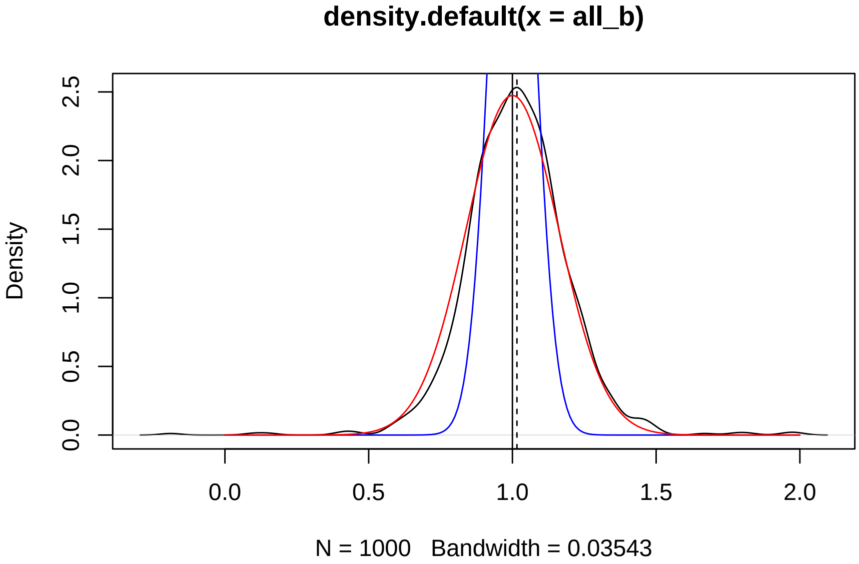
The standard OLS formula for the variance of b overestimates the precision of this estimator.

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.

```r
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")
```



Example 2.8 (Heteroskedasticity and Autocorrelation (HAC))**.** This includes the cases of Eqs. (2.27) and (2.28).

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

$$\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{i,j}\mathbf{x}_i\mathbf{x}_j'\right) = \tag{2.37}$$

$$\text{plim}\left(\frac{1}{n}\sum_{t=1}^{n}e_t^2\mathbf{x}_t\mathbf{x}_t' + \frac{1}{n}\sum_{\ell=1}^{L}\sum_{t=\ell+1}^{n}w_\ell e_t e_{t-\ell}(\mathbf{x}_t\mathbf{x}_{t-\ell}' + \mathbf{x}_{t-\ell}\mathbf{x}_t')\right)$$

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

Let us illustrate the influence of autocorrelation using simulations.

We consider the following model:

$$y_i = x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, x_i^2), \tag{2.38}$$

where the x_is and the ε_is are such that:

$$x_i = 0.8x_{i-1} + u_i \quad and \quad \varepsilon_i = 0.8\varepsilon_{i-1} + v_i, \tag{2.39}$$

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 $\varepsilon_i \sim i.i.d.\mathcal{N}(0, 2.8)$).

2.5.4 How to detect autocorrelation in residuals?

Consider the usual regression (say Eq. (2.29)).

The **Durbin-Watson test** is a typical autocorrelation test. Its test statistic is:

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

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. (2.30).)

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 $d_U(T, K)$:

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

2.6 Summary

	Under Assumptions 2.1+	**b** normal in small sample (Eq. (2.9))	**b** is BLUE (Thm 2.1)	**b** unbiased in small sample (Prop. 2.3)	**b** consistent (Prop. 2.13)*	**b** $\sim$ normal in large sample (Prop. 2.14)*
Condit. mean-zero	2.2	X	X	X	X	X
Homoskedasticity	2.3	X	X			
Uncorrelated residuals	2.4	X	X			
Normality of disturbances	2.5	X				

*: see however Prop. 2.13 and Prop. 2.14 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. (2.26)).

2.7 Clusters

```
library(AEC)
library(sandwich)
shp$income <- shp$i19ptotn/1000
shp$female <- 1*(shp$sex19==2)
eq <- lm(income ~ edyear19 + age19 + I(age19^2) + female,data=shp)
#eq <- lm(income ~ edyear19 + age19 + I(age19^2) + female + I(female*ownkid19*(age19<4
#lmtest::coeftest(eq,vcov. = sandwich)
#lmtest::coeftest(eq,vcov. = vcovHC)
#X <- cbind(1,shp$edyear19,shp$age19,shp$age19^2,shp$female)
#solve(t(X) %*% X) %*% t(X) %*% diag(eq$residuals^2) %*% X %*% solve(t(X) %*% X)
#vcovHC(eq,type="HC0")
#sandwich(eq)
#vcovHC(eq,type="HC1")
```

XXXX HC0, HC1… Davidson MacKinnon 2004 Section 5.5 XXXX

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:

$$\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon. \tag{2.40}$$

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 $\mathbf{X}$:

$$\mathbf{X} = \left[\begin{array}{c}\mathbf{X}_1\\ \mathbf{X}_1\\ \vdots\\ \mathbf{X}_G\end{array}\right].$$

With these notations, Eq. (2.40) rewrites:

$$\mathbf{b} - \beta = \left(\sum_{g=1}^{G} \mathbf{X}_g'\mathbf{X}_g\right)^{-1} \mathbf{X}' \sum_{g=1}^{G} \mathbf{s}_g, \tag{2.41}$$

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. (2.41) 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 2.6. 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 2.6, it comes that the covariance matrix of $\mathbf{b}$ is:

$$(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\Sigma_g\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{2.42}$$

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

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. (2.36)), i.e.:

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

The natural way to estimate Eq. (2.42) 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 $\mathbf{b}$:

$$\frac{G(n-1)}{(G-1)(n-K)}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\widehat{\Sigma}_g\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{2.43}$$

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

Note that we indeed find the White estimator when $G = n$ (see Eq. (2.36)).

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

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

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

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

Consider the follwing hypothesis:

Hypothesis 2.7. We have:

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

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:

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

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. (2.43), 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).

2.8 Shrinkage methods

Chosing the right variables is often complicated, especially in the presence of many potentially relevant covariates. Keeping a large number of covariates results in large standard deviations for the estimated parameters (see Section 2.3.3). 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** approach.[3]

In both cases (ridge and lasso), the OLS minimization problem (see Section 2.2), that is:

$$\mathbf{b} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta) \tag{2.44}$$

is replaced by the following:

$$\mathbf{b}_\lambda = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta) + \lambda f(\beta), \tag{2.45}$$

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

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

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

$$\begin{aligned}
f(\beta) &= \sum_{j=1}^{K} \beta_j^2 \quad \text{in the ridge case (ℓ_2 norm),}\\
f(\beta) &= \sum_{j=1}^{K} |\beta_j| \quad \text{in the lasso case (ℓ_1 norm).}
\end{aligned}$$

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

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

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

Parameter λ is a , that has to be determined separately from the minimization problem of Eq. (2.45). One can combine standard criteria (e.g., BIC or Akaike) with lasso regressions to help determine λ.

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

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

```
library(AEC)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-4
```

```
credit$owner <- 1*(credit$home_ownership=="OWN")
credit$renter <- 1*(credit$home_ownership=="MORTGAGE")
credit$verification_status <- 1*(credit$verification_status=="Not Verified")
```

```
credit$emp_length_10 <- 1*(credit$emp_length_10)
credit$log_annual_inc <- log(credit$annual_inc)
credit$log_funded_amnt <- log(credit$funded_amnt)
credit$annual_inc2 <- (credit$annual_inc)^2
credit$funded_amnt2 <- (credit$funded_amnt)^2
x <- subset(credit,select = c(delinq_2yrs, annual_inc, annual_inc2, log_annual_inc, dti, installm
```

Let us standardize the data:

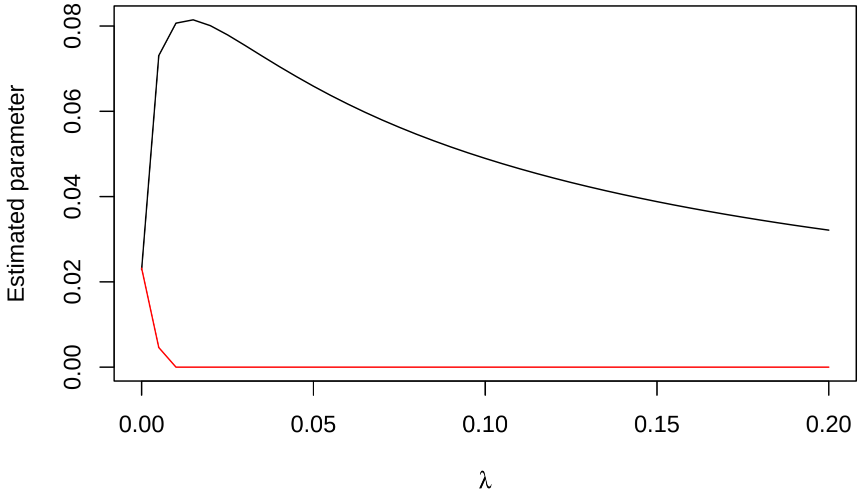
```
y <- credit$int_rate/sd(credit$int_rate,na.rm = TRUE)
stdv.x <- apply(x,2,function(a){sd(a,na.rm = TRUE)})
x <- x/t(matrix(stdv.x,dim(x)[2],dim(x)[1]))
```

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

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

The following figure shows how estimated parameters depend on λ:

```
variab <- 3
plot(result.ridge$lambda,coef(result.ridge)[variab,],type="l",
     ylim=c(min(coef(result.ridge)[variab,],coef(result.lasso)[variab,]),
            max(coef(result.ridge)[variab,],coef(result.lasso)[variab,])),
     xlab=expression(lambda),ylab="Estimated parameter")
lines(result.lasso$lambda,coef(result.lasso)[variab,],col="red")
```



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

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

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

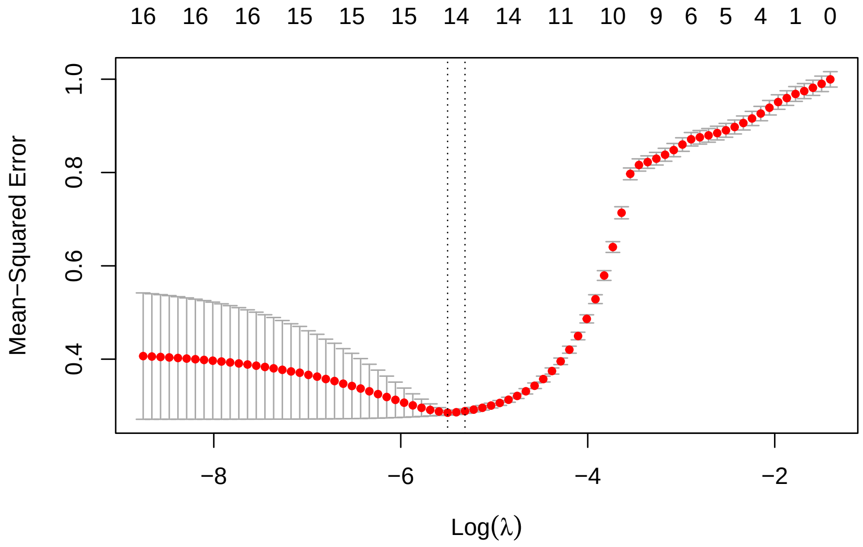
```
cbind(coef(result.lasso)[,i],coef(result.lasso)[,j])
```

```
##                           [,1]          [,2]
## (Intercept)         3.24385583  6.4716266208
## delinq_2yrs         0.06308870  0.0689352682
## annual_inc          0.00000000  0.0045956535
## annual_inc2         0.00000000  0.0000000000
## log_annual_inc      0.00000000 -0.0361238200
## dti                 0.00000000  0.0224224582
## installment         0.14767959  8.2287289816
## verification_status 0.00000000 -0.0009750047
## funded_amnt         0.00000000 -7.3091694210
## funded_amnt2        0.00000000 -0.4711846250
## log_funded_amnt     0.00000000 -0.2460932367
## pub_rec             0.03390816  0.0599725219
## emp_length_10         0.00000000 -0.0192494122
## owner                 0.00000000 -0.0244459908
## renter               -0.03882640 -0.0624308746
## pub_rec_bankruptcies  0.00000000  0.0000000000
## revol_util            0.00000000  0.0000000000
## revol_bal             0.00000000  0.0024022685
```

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

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

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



```
cvglmnet$lambda.min # value of lambda.min, that is the value of lambda that gives minimum mean cr
```

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

```
cvglmnet$lambda.1se # largest value of lambda that is such that the sample cost is wit
```

```
## [1] 0.004927671
```

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

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)           6.583140504
## delinq_2yrs           0.066258597
## annual_inc            0.006054923
## annual_inc2           .
## log_annual_inc       -0.037951509
## dti                   0.020652724
## installment           8.518960621
## verification_status  -0.002894435
## funded_amnt          -7.560852965
## funded_amnt2         -0.504192252
## log_funded_amnt      -0.253756726
## pub_rec               0.058043332
## emp_length_10        -0.018946660
## owner                -0.024841917
## renter               -0.060398310
## pub_rec_bankruptcies  .
## revol_util            .
## revol_bal             0.003093752
```

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

```
##       lambda.min
## 21529   3.807030
## 21547   3.416528
## 21579   4.026621
## 21583   3.255096
## 21608   3.340412
```

Chapter 3

Appendix

3.1 Statistical Tables

3.2 Statistics: definitions and results

Definition 3.1 (Partial correlation)**.** The **partial correlation** between y and z, controlling for some variables $\mathbf{X}$ is the sample correlation between y^* and z^*, where the latter two variables are the residuals in regressions of y on $\mathbf{X}$ and of z on $\mathbf{X}$, respectively.

This correlation is denoted by $r_{yz}^{\mathbf{X}}$. By definition, we have:

$$r_{yz}^{\mathbf{X}} = \frac{\mathbf{z}^{*\prime}\mathbf{y}^*}{\sqrt{(\mathbf{z}^{*\prime}\mathbf{z}^*)(\mathbf{y}^{*\prime}\mathbf{y}^*)}}. \tag{3.1}$$

Definition 3.2 (Skewness and kurtosis)**.** Let Y be a random variable whose fourth moment exists. The expectation of Y is denoted by μ.

- The skewness of Y is given by:

$$\frac{\mathbb{E}[(Y-\mu)^3]}{\{\mathbb{E}[(Y-\mu)^2]\}^{3/2}}.$$

- The kurtosis of Y is given by:

$$\frac{\mathbb{E}[(Y-\mu)^4]}{\{\mathbb{E}[(Y-\mu)^2]\}^2}.$$

Table 3.1: 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 3.2: 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 3.3: 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 3.4: 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

Definition 3.3 (Eigenvalues)**.** The eigenvalues of of a matrix M are the numbers λ for which:

$$|M - \lambda I| = 0,$$

where $| \bullet |$ is the determinant operator.

Proposition 3.1 (Properties of the determinant)**.** *We have:*

- $|MN| = |M| \times |N|$.
- $|M^{-1}| = |M|^{-1}$.
- *If M admits the diagonal representation $M = TDT^{-1}$, where D is a diagonal matrix whose diagonal entries are $\{\lambda_i\}_{i=1,\dots,n}$, then:*

$$|M - \lambda I| = \prod_{i=1}^{n} (\lambda_i - \lambda).$$

Definition 3.4 (Moore-Penrose inverse)**.** If $M \in \mathbb{R}^{m \times n}$, then its Moore-Penrose pseudo inverse (exists and) is the unique matrix $M^* \in \mathbb{R}^{n \times m}$ that satisfies:

i. $MM^*M = M$
ii. $M^*MM^* = M^*$
iii. $(MM^*)' = MM^*$ .iv $(M^*M)' = M^*M$.

Proposition 3.2 (Properties of the Moore-Penrose inverse)**.** • *If M is invertible then $M^* = M^{-1}$.*
- *The pseudo-inverse of a zero matrix is its transpose. **
- **

  The pseudo-inverse of the pseudo-inverse is the original matrix.

Definition 3.5 (F distribution)**.** Consider $n = n_1 + n_2$ i.i.d. $\mathcal{N}(0,1)$ r.v. X_i. If the r.v. F is defined by:

$$F = \frac{\sum_{i=1}^{n_1} X_i^2}{\sum_{j=n_1+1}^{n_1+n_2} X_j^2} \frac{n_2}{n_1}$$

then $F \sim \mathcal{F}(n_1, n_2)$. (See Table 3.4 for quantiles.)

Definition 3.6 (Student-t distribution)**.** Z follows a Student-t (or t) distribution with ν degrees of freedom (d.f.) if:

$$Z = X_0 \bigg/ \sqrt{\frac{\sum_{i=1}^{\nu} X_i^2}{\nu}}, \quad X_i \sim i.i.d.\mathcal{N}(0,1).$$

We have $\mathbb{E}(Z) = 0$, and $\mathbb{V}ar(Z) = \frac{\nu}{\nu-2}$ if $\nu > 2$. (See Table 3.2 for quantiles.)

Definition 3.7 (Chi-square distribution)**.** Z follows a χ^2 distribution with ν d.f. if $Z = \sum_{i=1}^{\nu} X_i^2$ where $X_i \sim i.i.d.\mathcal{N}(0,1)$. We have $\mathbb{E}(Z) = \nu$. (See Table 3.3 for quantiles.)

Definition 3.8 (Idempotent matrix)**.** Matrix M is idempotent if $M^2 = M$.

If M is a symmetric idempotent matrix, then $M'M = M$.

Proposition 3.3 (Roots of an idempotent matrix)**.** *The eigenvalues of an idempotent matrix are either 1 or 0.*

Proof. If λ is an eigenvalue of an idempotent matrix M then $\exists x \neq 0$ s.t. $Mx = \lambda x$. Hence $M^2 x = \lambda M x \Rightarrow (1-\lambda)Mx = 0$. Either all element of Mx are zero, in which case $\lambda = 0$ or at least one element of Mx is nonzero, in which case $\lambda = 1$. □

Proposition 3.4 (Idempotent matrix and chi-square distribution)**.** *The rank of a symmetric idempotent matrix is equal to its trace.*

Proof. The result follows from Prop. 3.3, combined with the fact that the rank of a symmetric matrix is equal to the number of its nonzero eigenvalues. □

Proposition 3.5 (Constrained least squares)**.** *The solution of the following optimisation problem:*

$$\min_{\beta} \quad ||\mathbf{y} - \mathbf{X}\beta||^2$$
$$\textit{subject to } \mathbf{R}\beta = \mathbf{q}$$

is given by:

$$\boxed{\beta^r = \beta_0 - (\mathbf{X'X})^{-1}\mathbf{R'}\{\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R'}\}^{-1}(\mathbf{R}\beta_0 - \mathbf{q}),}$$

where $\beta_0 = (\mathbf{X'X})^{-1}\mathbf{X'y}$.

Proof. See for instance Jackman, 2007. □

Proposition 3.6 (Chebychev's inequality)**.** *If* $\mathbb{E}(|X|^r)$ *is finite for some* $r > 0$ *then:*

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[|X-c|^r]}{\varepsilon^r}.$$

In particular, for $r = 2$*:*

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X - c| > \varepsilon) \leq \frac{\mathbb{E}[(X-c)^2]}{\varepsilon^2}.$$

Proof. Remark that $\varepsilon^r \mathbb{I}_{\{|X| \geq \varepsilon\}} \leq |X|^r$ and take the expectation of both sides. □

Definition 3.9 (Convergence in probability)**.** The random variable sequence x_n converges in probability to a constant c if $\forall \varepsilon$, $\lim_{n\to\infty} \mathbb{P}(|x_n - c| > \varepsilon) = 0$.

It is denoted as: $\text{plim } x_n = c$.

Definition 3.10 (Convergence in the Lr norm)**.** x_n converges in the r-th mean (or in the L^r-norm) towards x, if $\mathbb{E}(|x_n|^r)$ and $\mathbb{E}(|x|^r)$ exist and if

$$\lim_{n\to\infty} \mathbb{E}(|x_n - x|^r) = 0.$$

It is denoted as: $x_n \overset{L^r}{\to} c$.

For $r = 2$, this convergence is called **mean square convergence**.

Definition 3.11 (Almost sure convergence)**.** The random variable sequence x_n converges almost surely to c if $\mathbb{P}(\lim_{n\to\infty} x_n = c) = 1$.

It is denoted as: $x_n \overset{a.s.}{\to} c$.

Definition 3.12 (Convergence in distribution)**.** x_n is said to converge in distribution (or in law) to x if

$$\lim_{n\to\infty} F_{x_n}(s) = F_x(s)$$

for all s at which F_X –the cumulative distribution of X– is continuous.

It is denoted as: $x_n \overset{d}{\to} x$.

Proposition 3.7 (Rules for limiting distributions (Slutsky))**.** *We have:*

i. ***Slutsky's theorem:*** *If* $x_n \overset{d}{\to} x$ *and* $y_n \overset{p}{\to} c$ *then*

$$\begin{aligned} x_n y_n &\overset{d}{\to} xc \\ x_n + y_n &\overset{d}{\to} x + c \\ x_n/y_n &\overset{d}{\to} x/c \quad (\text{if } c \neq 0) \end{aligned}$$

ii. ***Continuous mapping theorem:*** *If* $x_n \overset{d}{\to} x$ *and* g *is a continuous function then* $g(x_n) \overset{d}{\to} g(x)$.

Proposition 3.8 (Implications of stochastic convergences)**.** *We have:*

$$\begin{tikzcd}
\boxed{\overset{L^s}{\to}} \arrow[r, Rightarrow, "1\leq r\leq s"'] & \boxed{\overset{L^r}{\to}} \arrow[d, Rightarrow] & \\
\boxed{\overset{a.s.}{\to}} \arrow[r, Rightarrow] & \boxed{\overset{p}{\to}} \arrow[r, Rightarrow] & \boxed{\overset{d}{\to}}.
\end{tikzcd}$$

Proof. (of the fact that $\left(\overset{p}{\to}\right) \Rightarrow \left(\overset{d}{\to}\right)$). Assume that $X_n \overset{p}{\to} X$. Denoting by F and F_n the c.d.f. of X and X_n, respectively:

$$F_n(x) = \mathbb{P}(X_n \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_n \leq x, X > x+\varepsilon) \leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon). \tag{3.2}$$

Besides,

$$F(x-\varepsilon) = \mathbb{P}(X \leq x-\varepsilon, X_n \leq x)+\mathbb{P}(X \leq x-\varepsilon, X_n > x) \leq F_n(x)+\mathbb{P}(|X_n-X| > \varepsilon),$$

which implies:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x). \tag{3.3}$$

Eqs. (3.2) and (3.3) imply:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x) \leq F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

Taking limits as $n \to \infty$ yields

$$F(x-\varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon).$$

The result is then obtained by taking limits as $\varepsilon \to 0$ (if F is continuous at x). □

Proposition 3.9 (Convergence in distribution to a constant). *If X_n converges in distribution to a constant c, then X_n converges in probability to c.*

Proof. If $\varepsilon > 0$, we have $\mathbb{P}(X_n < c - \varepsilon) \underset{n\to\infty}{\rightarrow} 0$ i.e. $\mathbb{P}(X_n \geq c - \varepsilon) \underset{n\to\infty}{\rightarrow} 1$ and $\mathbb{P}(X_n < c + \varepsilon) \underset{n\to\infty}{\rightarrow} 1$. Therefore $\mathbb{P}(c - \varepsilon \leq X_n < c + \varepsilon) \underset{n\to\infty}{\rightarrow} 1$, which gives the result. □

Example of ***plim*** **but not L^r convergence**: Let $\{x_n\}_{n\in\mathbb{N}}$ be a series of random variables defined by:

$$x_n = nu_n,$$

where u_n are independent random variables s.t. $u_n \sim \mathcal{B}(1/n)$.

We have $x_n \overset{p}{\rightarrow} 0$ but $x_n \overset{L^r}{\nrightarrow} 0$ because $\mathbb{E}(|X_n - 0|) = \mathbb{E}(X_n) = 1$.

Theorem 3.1 (Cauchy criterion (non-stochastic case)). *We have that $\sum_{i=0}^{T} a_i$ converges ($T \to \infty$) iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

$$\left|\sum_{i=N+1}^{M} a_i\right| < \eta.$$

Theorem 3.2 (Cauchy criterion (stochastic case)). *We have that $\sum_{i=0}^{T} \theta_i \varepsilon_{t-i}$ converges in mean square ($T \to \infty$) to a random variable iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

$$\mathbb{E}\left[\left(\sum_{i=N+1}^{M} \theta_i \varepsilon_{t-i}\right)^2\right] < \eta.$$

Definition 3.13 (Characteristic function)**.** For any real-valued random variable X, the characteristic function is defined by:

$$\phi_X : u \to \mathbb{E}[\exp(iuX)].$$

Theorem 3.3 (Law of large numbers)**.** *The sample mean is a consistent estimator of the population mean.*

Proof. Let's denote by ϕ_{X_i} the characteristic function of a r.v. X_i. If the mean of X_i is μ then the Talyor expansion of the characteristic function is:

$$\phi_{X_i}(u) = \mathbb{E}(\exp(iuX)) = 1 + iu\mu + o(u).$$

The properties of the characteristic function (see Def. 3.13) imply that:

$$\phi_{\frac{1}{n}(X_1+\dots+X_n)}(u) = \prod_{i=1}^{n} \left(1 + i\frac{u}{n}\mu + o\left(\frac{u}{n}\right)\right) \to e^{iu\mu}.$$

The facts that (a) $e^{iu\mu}$ is the characteristic function of the constant μ and (b) that a characteristic function uniquely characterises a distribution imply that the sample mean converges in distribution to the constant μ, which further implies that it converges in probability to μ. □

Theorem 3.4 (Lindberg-Levy Central limit theorem, CLT)**.** *If x_n is an i.i.d. sequence of random variables with mean μ and variance σ^2 ($\in]0, +\infty[$), then:*

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

Proof. Let us introduce the r.v. $Y_n := \sqrt{n}(\bar{X}_n - \mu)$. We have $\phi_{Y_n}(u) = \left[\mathbb{E}\left(\exp(i\frac{1}{\sqrt{n}}u(X_1 - \mu))\right)\right]^n$. We have:

$$\begin{aligned}\left[\mathbb{E}\left(\exp\left(i\frac{1}{\sqrt{n}}u(X_1-\mu)\right)\right)\right]^n &= \left[\mathbb{E}\left(1 + i\frac{1}{\sqrt{n}}u(X_1-\mu) - \frac{1}{2n}u^2(X_1-\mu)^2 + o(u^2)\right)\right]^n \\ &= \left(1 - \frac{1}{2n}u^2\sigma^2 + o(u^2)\right)^n.\end{aligned}$$

Therefore $\phi_{Y_n}(u) \underset{n\to\infty}{\to} \exp\left(-\frac{1}{2}u^2\sigma^2\right)$, which is the characteristic function of $\mathcal{N}(0, \sigma^2)$. □

Proposition 3.10 (Inverse of a partitioned matrix)**.** *We have:*

$$\left[\begin{array}{cc}\mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{array}\right]^{-1} =$$
$$\left[\begin{array}{cc}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\end{array}\right].$$

Proposition 3.11. *If* $\mathbf{A}$ *is idempotent and if* $\mathbf{x}$ *is Gaussian,* $\mathbf{Lx}$ *and* $\mathbf{x}'\mathbf{Ax}$ *are independent if* $\mathbf{LA} = \mathbf{0}$.

Proof. If $\mathbf{LA} = \mathbf{0}$, then the two Gaussian vectors $\mathbf{Lx}$ and $\mathbf{Ax}$ are independent. This implies the independence of any function of $\mathbf{Lx}$ and any function of $\mathbf{Ax}$. The results then follows from the observation that $\mathbf{x}'\mathbf{Ax} = (\mathbf{Ax})'(\mathbf{Ax})$, which is a function of $\mathbf{Ax}$. □

Proposition 3.12 (Inner product of a multivariate Gaussian variable)**.** *Let X be a n-dimensional multivariate Gaussian variable: $X \sim \mathcal{N}(0, \Sigma)$. We have:*

$$X'\Sigma^{-1}X \sim \chi^2(n).$$

Proof. Because Σ is a symmetrical definite positive matrix, it admits the spectral decomposition PDP' where P is an orthogonal matrix (i.e. $PP' = Id$) and D is a diagonal matrix with non-negative entries. Denoting by $\sqrt{D^{-1}}$ the diagonal matrix whose diagonal entries are the inverse of those of D, it is easily checked that the covariance matrix of $Y := \sqrt{D^{-1}}P'X$ is Id. Therefore Y is a vector of uncorrelated Gaussian variables. The properties of Gaussian variables imply that the components of Y are then also independent. Hence $Y'Y = \sum_i Y_i^2 \sim \chi^2(n)$.

It remains to note that $Y'Y = X'PD^{-1}P'X = X'\mathbb{V}ar(X)^{-1}X$ to conclude. □

Theorem 3.5 (Cauchy-Schwarz inequality)**.** *We have:*

$$|\mathbb{C}ov(X,Y)| \leq \sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}$$

and, if $X \neq=$ and $Y \neq 0$, the equality holds iff X and Y are the same up to an affine transformation.

Proof. If $\mathbb{V}ar(X) = 0$, this is trivial. If this is not the case, then let's define Z as $Z = Y - \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$. It is easily seen that $\mathbb{C}ov(X,Z) = 0$. Then, the variance of $Y = Z + \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$ is equal to the sum of the variance of Z and of the variance of $\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$, that is:

$$\mathbb{V}ar(Y) = \mathbb{V}ar(Z) + \left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}\right)^2 \mathbb{V}ar(X) \geq \left(\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}\right)^2 \mathbb{V}ar(X).$$

The equality holds iff $\mathbb{V}ar(Z) = 0$, i.e. iff $Y = \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X + cst$. □

Definition 3.14 (Matrix derivatives)**.** Consider a fonction $f : \mathbb{R}^K \rightarrow \mathbb{R}$. Its first-order derivative is:

$$\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = \left[\begin{array}{c} \frac{\partial f}{\partial b_1}(\mathbf{b}) \\ \vdots \\ \frac{\partial f}{\partial b_K}(\mathbf{b}) \end{array}\right].$$

We use the notation:

$$\frac{\partial f}{\partial \mathbf{b}'}(\mathbf{b}) = \left(\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b})\right)'.$$

Proposition 3.13. *We have:*

- *If* $f(\mathbf{b}) = A'\mathbf{b}$ *where* A *is a* $K \times 1$ *vector then* $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = A$.
- *If* $f(\mathbf{b}) = \mathbf{b}'A\mathbf{b}$ *where* A *is a* $K \times K$ *matrix, then* $\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = 2A\mathbf{b}$.

Definition 3.15 (Asymptotic level)**.** An asymptotic test with critical region Ω_n has an asymptotic level equal to α if:

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

where S_n is the test statistic and Θ is such that the null hypothesis H_0 is equivalent to $\theta \in \Theta$.

Definition 3.16 (Asymptotically consistent test)**.** An asymptotic test with critical region Ω_n is consistent if:

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

where S_n is the test statistic and Θ^c is such that the null hypothesis H_0 is equivalent to $\theta \notin \Theta^c$.

Definition 3.17 (Kullback discrepancy)**.** Given two p.d.f. f and f^*, the Kullback discrepancy is defined by:

$$I(f, f^*) = \mathbb{E}^*\left(\log \frac{f^*(Y)}{f(Y)}\right) = \int \log \frac{f^*(y)}{f(y)} f^*(y)dy.$$

Proposition 3.14 (Properties of the Kullback discrepancy)**.** *We have:*

i. $I(f, f^*) \geq 0$

ii. $I(f, f^*) = 0$ *iff* $f \equiv f^*$.

Proof. $x \to -\log(x)$ is a convex function. Therefore $\mathbb{E}^*(-\log f(Y)/f^*(Y)) \geq -\log \mathbb{E}^*(f(Y)/f^*(Y)) = 0$ (proves (i)). Since $x \to -\log(x)$ is strictly convex, equality in (i) holds if and only if $f(Y)/f^*(Y)$ is constant (proves (ii)). □

Proposition 3.15 (Square and absolute summability)**.** *We have:*

$$\underbrace{\sum_{i=0}^{\infty}|\theta_i| < +\infty}_{\textit{Absolute summability}} \quad \Rightarrow \quad \underbrace{\sum_{i=0}^{\infty}\theta_i^2 < +\infty}_{\textit{Square summability}} \quad .$$

Proof. See Appendix 3.A in Hamilton. Idea: Absolute summability implies that there exist N such that, for $j > N$, $|\theta_j| < 1$ (deduced from Cauchy criterion, Theorem 3.1 and therefore $\theta_j^2 < |\theta_j|$. □

3.3 Some properties of Gaussian variables

Proposition 3.16 (Bayesian update in a vector of Gaussian variables)**.** *If*

$$\left[\begin{array}{c} Y_1 \\ Y_2 \end{array}\right] \sim \mathcal{N}\left(0, \left[\begin{array}{cc} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{array}\right]\right),$$

then

$$\begin{aligned} Y_2|Y_1 &\sim \mathcal{N}\left(\Omega_{21}\Omega_{11}^{-1}Y_1, \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}\right). \\ Y_1|Y_2 &\sim \mathcal{N}\left(\Omega_{12}\Omega_{22}^{-1}Y_2, \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right). \end{aligned}$$

Proposition 3.17 (Truncated distributions)**.** *If X is a random variable distributed according to some p.d.f. f, with c.d.f. F, with infinite support. Then the p.d.f. of $X|a \le X < b$ is*

$$g(x) = \frac{f(x)}{F(b) - F(a)}\mathbb{I}_{\{a \le x < b\}},$$

for any $a < b$.

In partiucular, for a Gaussian variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$f(X = x|a \le X < b) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{Z}.$$

with $Z = \Phi(\beta) - \Phi(\alpha)$, where $\alpha = \dfrac{a-\mu}{\sigma}$ and $\beta = \dfrac{b-\mu}{\sigma}$.

Moreover:

$$\mathbb{E}(X|a \leq X < b) \quad = \quad \mu - \frac{\phi(\beta) - \phi(\alpha)}{Z}\sigma. \tag{3.4}$$

We also have:

$$\begin{aligned} & \mathbb{V}ar(X|a \leq X < b) \\ = \quad & \sigma^2 \left[1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{Z} - \left(\frac{\phi(\beta) - \phi(\alpha)}{Z}\right)^2\right] \end{aligned} \tag{3.5}$$

In particular, for $b \rightarrow \infty$, we get:

$$\mathbb{V}ar(X|a < X) = \sigma^2 \left[1 + \alpha\lambda(-\alpha) - \lambda(-\alpha)^2\right], \tag{3.6}$$

with $\lambda(x) = \dfrac{\phi(x)}{\Phi(x)}$ is called the ***inverse Mills ratio****.*

Consider the case where $a \rightarrow -\infty$ (i.e. the conditioning set is $X < b$) and $\mu = 0$, $\sigma = 1$. Then Eq. (3.4) gives $\mathbb{E}(X|X < b) = -\lambda(b) = -\dfrac{\phi(b)}{\Phi(b)}$, where λ is the function computing the inverse Mills ratio.

Proposition 3.18 (p.d.f. of a multivariate Gaussian variable)**.** *If $Y \sim \mathcal{N}(\mu, \Omega)$ and if Y is a n-dimensional vector, then the density function of Y is:*

$$\frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \exp\left[-\frac{1}{2}\left(Y - \mu\right)' \Omega^{-1}\left(Y - \mu\right)\right].$$

3.4 Proofs

3.4.1 Proof of Proposition ??

Proof. Assumptions (i) and (ii) (in the set of Assumptions **??**) imply that θ_{MLE} exists ($= \text{argmax}_\theta (1/n) \log \mathcal{L}(\theta; \mathbf{y})$).

$(1/n) \log \mathcal{L}(\theta; \mathbf{y})$ can be interpreted as the sample mean of the r.v. $\log f(Y_i; \theta)$ that are i.i.d. Therefore $(1/n) \log \mathcal{L}(\theta; \mathbf{y})$ converges to $\mathbb{E}_{\theta_0}(\log f(Y; \theta))$ – which exists (Assumption iv).

Because the latter convergence is uniform (Assumption v), the solution θ_{MLE} almost surely converges to the solution to the limit problem:

$$\text{argmax}_\theta \mathbb{E}_{\theta_0}(\log f(Y; \theta)) = \text{argmax}_\theta \int_y \log f(y; \theta) f(y; \theta_0) dy.$$

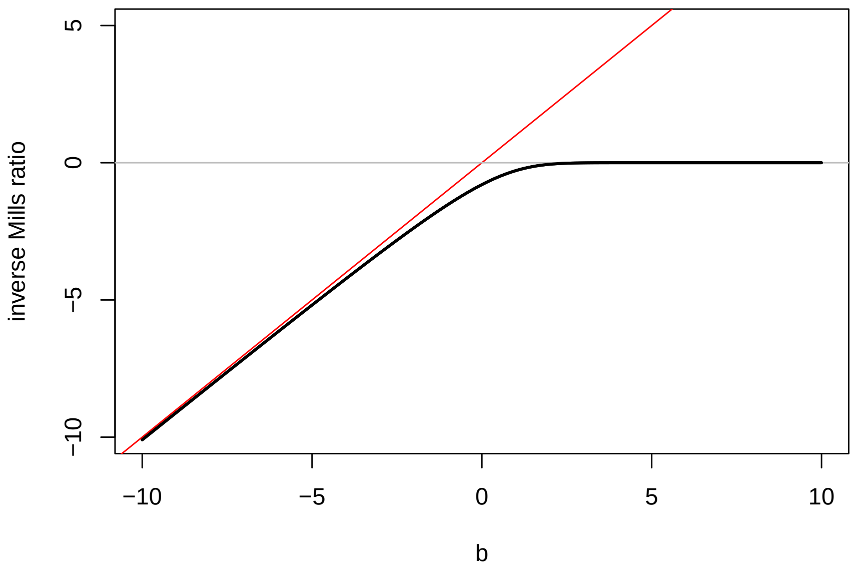


Figure 3.1: $\mathbb{E}(X|X < b)$ as a function of b when $X \sim \mathcal{N}(0, 1)$ (in black).

Properties of the Kullback information measure (see Prop. 3.14), together with the identifiability assumption (ii) implies that the solution to the limit problem is unique and equal to θ_0.

Consider a r.v. sequence θ that converges to θ_0. The Taylor expansion of the score in a neighborood of θ_0 yields to:

$$\frac{\partial \log \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} = \frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} + \frac{\partial^2 \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta \partial \theta'}(\theta - \theta_0) + o_p(\theta - \theta_0)$$

θ_{MLE} converges to θ_0 and satisfies the likelihood equation $\frac{\partial \log \mathcal{L}(\theta; \mathbf{y})}{\partial \theta} = \mathbf{0}$. Therefore:

$$\frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} \approx -\frac{\partial^2 \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta \partial \theta'}(\theta_{MLE} - \theta_0),$$

or equivalently:

$$\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} \approx \left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i; \theta_0)}{\partial \theta \partial \theta'} \right) \sqrt{n}(\theta_{MLE} - \theta_0),$$

By the law of large numbers, we have: $\left(-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i; \theta_0)}{\partial \theta \partial \theta'} \right) \to \frac{1}{n} \mathbf{I}(\theta_0) = \mathcal{I}_Y(\theta_0)$.

Besides, we have:

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial\theta} &= \sqrt{n}\left(\frac{1}{n}\sum_i \frac{\partial \log f(y_i;\theta_0)}{\partial\theta}\right)\\ &= \sqrt{n}\left(\frac{1}{n}\sum_i \left\{\frac{\partial \log f(y_i;\theta_0)}{\partial\theta} - \mathbb{E}_{\theta_0}\frac{\partial \log f(Y_i;\theta_0)}{\partial\theta}\right\}\right)\end{aligned}$$

which converges to $\mathcal{N}(0,\mathcal{I}_Y(\theta_0))$ by the CLT.

Collecting the preceding results leads to (b). The fact that θ_{MLE} achieves the FDCR bound proves (c). □

3.4.2 Proof of Proposition ??

Proof. We have $\sqrt{n}(\hat\theta_n - \theta_0) \overset{d}{\rightarrow} \mathcal{N}(0,\mathcal{I}(\theta_0)^{-1})$ (Eq. **??**eq:normMLE)). A Taylor expansion around θ_0 yields to:

$$\sqrt{n}(h(\hat\theta_n) - h(\theta_0)) \overset{d}{\rightarrow} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial\theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial\theta}\right). \tag{3.7}$$

Under H_0, $h(\theta_0)=0$ therefore:

$$\sqrt{n}h(\hat\theta_n) \overset{d}{\rightarrow} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial\theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial\theta}\right). \tag{3.8}$$

Hence

$$\sqrt{n}\left(\frac{\partial h(\theta_0)}{\partial\theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial\theta}\right)^{-1/2} h(\hat\theta_n) \overset{d}{\rightarrow} \mathcal{N}\left(0, Id\right).$$

Taking the quadratic form, we obtain:

$$nh(\hat\theta_n)'\left(\frac{\partial h(\theta_0)}{\partial\theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h(\theta_0)'}{\partial\theta}\right)^{-1} h(\hat\theta_n) \overset{d}{\rightarrow} \chi^2(r).$$

The fact that the test has asymptotic level α directly stems from what precedes. **Consistency of the test**: Consider $\theta_0 \in \Theta$. Because the MLE is consistent, $h(\hat\theta_n)$ converges to $h(\theta_0) \neq 0$. Eq. (3.7) is still valid. It implies that ξ^W_n converges to $+\infty$ and therefore that $\mathbb{P}_\theta(\xi^W_n \geq \chi^2_{1-\alpha}(r)) \rightarrow 1$. □

3.4.3 Proof of Proposition ??

Proof. Notations: "≈" means "equal up to a term that converges to 0 in probability". We are under H_0. $\hat\theta^0$ is the constrained ML estimator; $\hat\theta$ denotes the unconstrained one.

We combine the two Taylor expansion: $h(\hat{\theta}_n) \approx \dfrac{\partial h(\theta_0)}{\partial \theta'}(\hat{\theta}_n - \theta_0)$ and $h(\hat{\theta}^0_n) \approx \dfrac{\partial h(\theta_0)}{\partial \theta'}(\hat{\theta}^0_n - \theta_0)$ and we use $h(\hat{\theta}^0_n) = 0$ (by definition) to get:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}\sqrt{n}(\hat{\theta}_n - \hat{\theta}^0_n). \tag{3.9}$$

Besides, we have (using the definition of the information matrix):

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}^0_n - \theta_0) \tag{3.10}$$

and:

$$0 = \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n;\mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{3.11}$$

Taking the difference and multiplying by $\mathcal{I}(\theta_0)^{-1}$:

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}^0_n) \approx \mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial \theta}\mathcal{I}(\theta_0). \tag{3.12}$$

Eqs. (3.9) and (3.12) yield to:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial \theta}. \tag{3.13}$$

Recall that $\hat{\theta}^0_n$ is the MLE of θ_0 under the constraint $h(\theta) = 0$. The vector of Lagrange multipliers $\hat{\lambda}_n$ associated to this program satisfies:

$$\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial \theta} + \frac{\partial h'(\hat{\theta}^0_n;\mathbf{y})}{\partial \theta}\hat{\lambda}_n = 0. \tag{3.14}$$

Substituting the latter equation in Eq. (3.13) gives:

$$\sqrt{n}h(\hat{\theta}_n) \approx -\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\hat{\theta}^0_n;\mathbf{y})}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}} \approx -\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\theta_0;\mathbf{y})}{\partial \theta}\frac{\hat{\lambda}_n}{\sqrt{n}}$$

which yields:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \approx -\left(\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\theta_0;\mathbf{y})}{\partial \theta}\right)^{-1}\sqrt{n}h(\hat{\theta}_n). \tag{3.15}$$

It follows, from Eq. (3.8), that:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \overset{d}{\rightarrow} \mathcal{N}\left(0, \left(\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\theta_0;\mathbf{y})}{\partial \theta}\right)^{-1}\right).$$

Taking the quadratic form of the last equation gives:

$$\frac{1}{n}\hat{\lambda}_n' \frac{\partial h(\hat{\theta}_n^0)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n^0)^{-1} \frac{\partial h'(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta} \hat{\lambda}_n \xrightarrow{d} \chi^2(r).$$

Using Eq. (3.14), it appears that the left-hand side term of the last equation is ξ^{LM} as defined in Eq. (**??**). Consistency: see Remark 17.3 in Gouriéroux and Monfort (1995). □

3.4.4 Proof of Proposition ??

Proof. We have (using Eq. **??**eq:multiplier)):

$$\xi_n^{LM} = \frac{1}{n}\hat{\lambda}_n' \frac{\partial h(\hat{\theta}_n^0)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n^0)^{-1} \frac{\partial h'(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta} \hat{\lambda}_n.$$

Since, under H_0, $\hat{\theta}_n^0 \approx \hat{\theta}_n \approx \theta_0$, Eq. (3.15) therefore implies that:

$$\xi^{LM} \approx n h(\hat{\theta}_n)' \left(\frac{\partial h(\hat{\theta}_n)}{\partial \theta'} \mathcal{I}(\hat{\theta}_n)^{-1} \frac{\partial h'(\hat{\theta}_n;\mathbf{y})}{\partial \theta} \right)^{-1} h(\hat{\theta}_n) = \xi^W,$$

which gives the result. □

3.4.5 Proof of Proposition ??

Proof. The second-order taylor expansions of $\log \mathcal{L}(\hat{\theta}_n^0, \mathbf{y})$ and $\log \mathcal{L}(\hat{\theta}_n, \mathbf{y})$ are:

$$\begin{aligned}
\log \mathcal{L}(\hat{\theta}_n, \mathbf{y}) &\approx \log \mathcal{L}(\theta_0, \mathbf{y}) + \frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n - \theta_0) - \frac{n}{2}(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0) \\
\log \mathcal{L}(\hat{\theta}_n^0, \mathbf{y}) &\approx \log \mathcal{L}(\theta_0, \mathbf{y}) + \frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n^0 - \theta_0) - \frac{n}{2}(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0).
\end{aligned}$$

Taking the difference, we obtain:

$$\xi_n^{LR} \approx 2\frac{\partial \log \mathcal{L}(\theta_0, \mathbf{y})}{\partial \theta'}(\hat{\theta}_n - \hat{\theta}_n^0) + n(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) - n(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0).$$

Using $\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$ (Eq. (3.11)), we have:

$$\xi_n^{LR} \approx 2n(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \hat{\theta}_n^0) + n(\hat{\theta}_n^0 - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n^0 - \theta_0) - n(\hat{\theta}_n - \theta_0)' \mathcal{I}(\theta_0)(\hat{\theta}_n - \theta_0).$$

In the second of the three terms in the sum, we replace $(\hat{\theta}_n^0 - \theta_0)$ by $(\hat{\theta}_n^0 - \hat{\theta}_n + \hat{\theta}_n - \theta_0)$ and we develop the associated product. This leads to:

$$\xi_n^{LR} \approx n(\hat{\theta}_n^0 - \hat{\theta}_n)' \mathcal{I}(\theta_0)^{-1}(\hat{\theta}_n^0 - \hat{\theta}_n). \tag{3.16}$$

The difference between Eqs. (3.10) and (3.11) implies:

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial\theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \hat{\theta}^0_n),$$

which, associated to Eq. @**??**eq:lr10), gives:

$$\xi_n^{LR} \approx \frac{1}{n}\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial\theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial \log \mathcal{L}(\hat{\theta}^0_n;\mathbf{y})}{\partial\theta} \approx \xi_n^{LM}.$$

Hence ξ_n^{LR} has the same asymptotic distribution as ξ_n^{LM}.

Let's show that the test is consistent. For this, note that:

$$\frac{\log \mathcal{L}(\hat{\theta},\mathbf{y}) - \log \mathcal{L}(\hat{\theta}^0,\mathbf{y})}{n} = \frac{1}{n}\sum_{i=1}^{n}[\log f(y_i;\hat{\theta}_n) - \log f(y_i;\hat{\theta}^0_n)] \rightarrow \mathbb{E}_0[\log f(Y;\theta_0) - \log f(Y;\theta_\infty)],$$

where θ_∞, the pseudo true value, is such that $h(\theta_\infty) \neq 0$ (by definition of H_1). From the Kullback inequality and the asymptotic identifiability of θ_0, it follows that $\mathbb{E}_0[\log f(Y;\theta_0) - \log f(Y;\theta_\infty)] > 0$. Therefore $\xi_n^{LR} \rightarrow +\infty$ under H_1. □

3.4.6 Proof of Eq. (??)

Proof. We have:

$$\begin{aligned}
& T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] \\
= \; & T\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - \mu)\right)^2\right] = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(y_t-\mu)^2 + 2\sum_{s<t\le T}(y_t-\mu)(y_s-\mu)\right] \\
= \; & \gamma_0 + \frac{2}{T}\left(\sum_{t=2}^{T}\mathbb{E}\left[(y_t-\mu)(y_{t-1}-\mu)\right]\right) + \frac{2}{T}\left(\sum_{t=3}^{T}\mathbb{E}\left[(y_t-\mu)(y_{t-2}-\mu)\right]\right) + \dots \\
& + \frac{2}{T}\left(\sum_{t=T-1}^{T}\mathbb{E}\left[(y_t-\mu)(y_{t-(T-2)}-\mu)\right]\right) + \frac{2}{T}\mathbb{E}\left[(y_t-\mu)(y_{t-(T-1)}-\mu)\right] \\
= \; & \gamma_0 + 2\frac{T-1}{T}\gamma_1 + \dots + 2\frac{1}{T}\gamma_{T-1}.
\end{aligned}$$

Therefore:

$$T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j \quad = \quad -2\frac{1}{T}\gamma_1 - 2\frac{2}{T}\gamma_2 - \dots - 2\frac{T-1}{T}\gamma_{T-1} - 2\gamma_T - 2\gamma_{T+1} + \dots$$

And then:

$$\left|T\mathbb{E}\left[(\bar{y}_T - \mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j\right| \le 2\frac{1}{T}|\gamma_1| + 2\frac{2}{T}|\gamma_2| + \dots + 2\frac{T-1}{T}|\gamma_{T-1}| + 2|\gamma_T| + 2|\gamma_{T+1}| + \dots$$

For any $q \leq T$, we have:

$$\begin{aligned}
\left| T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j \right| &\leq 2\frac{1}{T}|\gamma_1| + 2\frac{2}{T}|\gamma_2| + \dots + 2\frac{q-1}{T}|\gamma_{q-1}| + 2\frac{q}{T}|\gamma_q| + \\
&\quad 2\frac{q+1}{T}|\gamma_{q+1}| + \dots + 2\frac{T-1}{T}|\gamma_{T-1}| + 2|\gamma_T| + 2|\gamma_{T+1}| + \dots \\
&\leq \frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \dots + (q-1)|\gamma_{q-1}| + q|\gamma_q|\right) + \\
&\quad 2|\gamma_{q+1}| + \dots + 2|\gamma_{T-1}| + 2|\gamma_T| + 2|\gamma_{T+1}| + \dots
\end{aligned}$$

Consider $\varepsilon > 0$. The fact that the autocovariances are absolutely summable implies that there exists q_0 such that (Cauchy criterion, Theorem 3.1):

$$2|\gamma_{q_0+1}| + 2|\gamma_{q_0+2}| + 2|\gamma_{q_0+3}| + \dots < \varepsilon/2.$$

Then, if $T > q_0$, it comes that:

$$\left| T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j \right| \leq \frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \dots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right) + \varepsilon/2.$$

If $T \geq 2\left(|\gamma_1| + 2|\gamma_2| + \dots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right)/(\varepsilon/2)$ ($= f(q_0)$, say) then

$$\frac{2}{T}\left(|\gamma_1| + 2|\gamma_2| + \dots + (q_0-1)|\gamma_{q_0-1}| + q_0|\gamma_{q_0}|\right) \leq \varepsilon/2.$$

Then, if $T > f(q_0)$ and $T > q_0$, i.e. if $T > \max(f(q_0), q_0)$, we have:

$$\left| T\mathbb{E}\left[(\bar{y}_T-\mu)^2\right] - \sum_{j=-\infty}^{+\infty}\gamma_j \right| \leq \varepsilon.$$

□

3.4.7 Proof of Proposition ??

Proof. We have:

$$\begin{aligned}
\mathbb{E}([y_{t+1}-y^*_{t+1}]^2) &= \mathbb{E}\left([\{y_{t+1}-\mathbb{E}(y_{t+1}|x_t)\} + \{\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}\}]^2\right) \\
&= \mathbb{E}\left([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)]^2\right) + \mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]^2\right) \\
&\quad +2\mathbb{E}\left([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]\right). \qquad (3.17)
\end{aligned}$$

Let us focus on the last term. We have:

$$\begin{aligned}
&\mathbb{E}\left([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]\right) \\
=\ &\mathbb{E}(\mathbb{E}([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)]\underbrace{[\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]}_{\text{function of } x_t}|x_t)) \\
=\ &\mathbb{E}([\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]\mathbb{E}([y_{t+1}-\mathbb{E}(y_{t+1}|x_t)]|x_t)) \\
=\ &\mathbb{E}([\mathbb{E}(y_{t+1}|x_t)-y^*_{t+1}]\underbrace{[\mathbb{E}(y_{t+1}|x_t)-\mathbb{E}(y_{t+1}|x_t)]}_{=0}) = 0.
\end{aligned}$$

Therefore, Eq. (3.17) becomes:

$$\begin{aligned} & \mathbb{E}([y_{t+1} - y^*_{t+1}]^2) \\ = \quad & \underbrace{\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2\right)}_{\geq 0 \text{ and does not depend on } y^*_{t+1}} \quad + \underbrace{\mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]^2\right)}_{\geq 0 \text{ and depends on } y^*_{t+1}} . \end{aligned}$$

This implies that $\mathbb{E}([y_{t+1} - y^*_{t+1}]^2)$ is always larger than $\mathbb{E}([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2)$, and is therefore minimized if the second term is equal to zero, that is if $\mathbb{E}(y_{t+1}|x_t) = y^*_{t+1}$. □

3.4.8 Proof of Proposition ??

Proof. Using Proposition **??** (in Appendix **??**), we obtain that, conditionally on x_1, the log-likelihood is given by

$$\begin{aligned} \log \mathcal{L}(Y_T;\theta) &= -(Tn/2)\log(2\pi) + (T/2)\log\left|\Omega^{-1}\right| \\ & -\frac{1}{2}\sum_{t=1}^{T}\left[(y_t - \Pi' x_t)'\,\Omega^{-1}\,(y_t - \Pi' x_t)\right]. \end{aligned}$$

Let's rewrite the last term of the log-likelihood:

$$\begin{aligned} \sum_{t=1}^{T}\left[(y_t - \Pi' x_t)'\,\Omega^{-1}\,(y_t - \Pi' x_t)\right] &= \\ \sum_{t=1}^{T}\left[\left(y_t - \hat{\Pi}' x_t + \hat{\Pi}' x_t - \Pi' x_t\right)'\,\Omega^{-1}\left(y_t - \hat{\Pi}' x_t + \hat{\Pi}' x_t - \Pi' x_t\right)\right] &= \\ \sum_{t=1}^{T}\left[\left(\hat{\varepsilon}_t + (\hat{\Pi} - \Pi)' x_t\right)'\,\Omega^{-1}\left(\hat{\varepsilon}_t + (\hat{\Pi} - \Pi)' x_t\right)\right], \end{aligned}$$

where the j^{th} element of the $(n \times 1)$ vector $\hat{\varepsilon}_t$ is the sample residual, for observation t, from an OLS regression of $y_{j,t}$ on x_t. Expanding the previous equation, we get:

$$\begin{aligned} \sum_{t=1}^{T}\left[(y_t - \Pi' x_t)'\,\Omega^{-1}\,(y_t - \Pi' x_t)\right] &= \sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}\hat{\varepsilon}_t \\ +2\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi} - \Pi)' x_t &+ \sum_{t=1}^{T} x_t'(\hat{\Pi} - \Pi)\Omega^{-1}(\hat{\Pi} - \Pi)' x_t. \end{aligned}$$

Let's apply the trace operator on the second term (that is a scalar):

$$\begin{aligned} \sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi} - \Pi)' x_t &= Tr\left(\sum_{t=1}^{T}\hat{\varepsilon}_t'\Omega^{-1}(\hat{\Pi} - \Pi)' x_t\right) \\ = Tr\left(\sum_{t=1}^{T}\Omega^{-1}(\hat{\Pi} - \Pi)' x_t\hat{\varepsilon}_t'\right) &= Tr\left(\Omega^{-1}(\hat{\Pi} - \Pi)'\sum_{t=1}^{T} x_t\hat{\varepsilon}_t'\right). \end{aligned}$$

Given that, by construction (property of OLS estimates), the sample residuals are orthogonal to the explanatory variables, this term is zero. Introducing $\tilde{x}_t = (\hat{\Pi} - \Pi)' x_t$, we have

$$\sum_{t=1}^{T} \left[\left(y_t - \Pi' x_t \right)' \Omega^{-1} \left(y_t - \Pi' x_t \right) \right] = \sum_{t=1}^{T} \hat{\varepsilon}_t' \Omega^{-1} \hat{\varepsilon}_t + \sum_{t=1}^{T} \tilde{x}_t' \Omega^{-1} \tilde{x}_t.$$

Since Ω is a positive definite matrix, Ω^{-1} is as well. Consequently, the smallest value that the last term can take is obtained for $\tilde{x}_t = 0$, i.e. when $\Pi = \hat{\Pi}$.

The MLE of Ω is the matrix $\hat{\Omega}$ that maximizes $\Omega \overset{\ell}{\rightarrow} L(Y_T; \hat{\Pi}, \Omega)$. We have:

$$\log \mathcal{L}(Y_T; \hat{\Pi}, \Omega) \quad = \quad -(Tn/2) \log(2\pi) + (T/2) \log \left| \Omega^{-1} \right| - \frac{1}{2} \sum_{t=1}^{T} \left[\hat{\varepsilon}_t' \Omega^{-1} \hat{\varepsilon}_t \right].$$

Matrix $\hat{\Omega}$ is a symmetric positive definite. It is easily checked that the (unrestricted) matrix that maximizes the latter expression is symmetric positive definite matrix. Indeed:

$$\frac{\partial \log \mathcal{L}(Y_T; \hat{\Pi}, \Omega)}{\partial \Omega} = \frac{T}{2} \Omega' - \frac{1}{2} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t' \Rightarrow \hat{\Omega}' = \frac{1}{T} \sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t',$$

which leads to the result. □

3.4.9 Proof of Proposition ??

Proof. Let us drop the i subscript. Rearranging Eq. (**??**), we have:

$$\sqrt{T}(\mathbf{b} - \beta) = (X'X/T)^{-1} \sqrt{T}(X'\varepsilon/T).$$

Let us consider the autocovariances of $\mathbf{v}_t = x_t \varepsilon_t$, denoted by γ_j^v. Using the fact that x_t is a linear combination of past ε_ts and that ε_t is a white noise, we get that $\mathbb{E}(\varepsilon_t x_t) = 0$. Therefore

$$\gamma_j^v = \mathbb{E}(\varepsilon_t \varepsilon_{t-j} x_t x_{t-j}').$$

If $j > 0$, we have $\mathbb{E}(\varepsilon_t \varepsilon_{t-j} x_t x_{t-j}') = \mathbb{E}(\mathbb{E}[\varepsilon_t \varepsilon_{t-j} x_t x_{t-j}' | \varepsilon_{t-j}, x_t, x_{t-j}]) = \mathbb{E}(\varepsilon_{t-j} x_t x_{t-j}' \mathbb{E}[\varepsilon_t | \varepsilon_{t-j}, x_t, x_{t-j}]) = 0$. Note that we have $\mathbb{E}[\varepsilon_t | \varepsilon_{t-j}, x_t, x_{t-j}] = 0$ because $\{\varepsilon_t\}$ is an i.i.d. white noise sequence. If $j = 0$, we have:

$$\gamma_0^v = \mathbb{E}(\varepsilon_t^2 x_t x_t') = \mathbb{E}(\varepsilon_t^2) \mathbb{E}(x_t x_t') = \sigma^2 \mathbf{Q}.$$

The convergence in distribution of $\sqrt{T}(X'\varepsilon/T) = \sqrt{T} \frac{1}{T} \sum_{t=1}^{T} v_t$ results from the Central Limit Theorem for covariance-stationary processes, using the γ_j^v computed above. □

3.5 Additional codes

3.5.1 Simulating GEV distributions

The following lines of code have been used to generate Figure **??**.

```
n.sim <- 4000
par(mfrow=c(1,3),
    plt=c(.2,.95,.2,.85))
all.rhos <- c(.3,.6,.95)
for(j in 1:length(all.rhos)){
  theta <- 1/all.rhos[j]
  v1 <- runif(n.sim)
  v2 <- runif(n.sim)
  w <- rep(.000001,n.sim)
  # solve for f(w) = w*(1 - log(w)/theta) - v2 = 0
  for(i in 1:20){
    f.i <- w * (1 - log(w)/theta) - v2
    f.prime <- 1 - log(w)/theta - 1/theta
    w <- w - f.i/f.prime
  }
  u1 <- exp(v1^(1/theta) * log(w))
  u2 <- exp((1-v1)^(1/theta) * log(w))

  # Get eps1 and eps2 using the inverse of the Gumbel distribution's cdf:
  eps1 <- -log(-log(u1))
  eps2 <- -log(-log(u2))
  cbind(cor(eps1,eps2),1-all.rhos[j]^2)
  plot(eps1,eps2,pch=19,col="#FF000044",
       main=paste("rho = ",toString(all.rhos[j]),sep=""),
       xlab=expression(epsilon[1]),
       ylab=expression(epsilon[2]),
       cex.lab=2,cex.main=1.5)
}
```

3.5.2 Computing the covariance matrix of IRF using the delta method

```
irf.function <- function(THETA){
  c <- THETA[1]
  phi <- THETA[2:(p+1)]
  if(q>0){
    theta <- c(1,THETA[(1+p+1):(1+p+q)])
  }else{
    theta <- 1
  }
  sigma <- THETA[1+p+q+1]
  r <- dim(Matrix.of.Exog)[2] - 1
  beta <- THETA[(1+p+q+1+1):(1+p+q+1+(r+1))]

  irf <- sim.arma(0,phi,beta,sigma=sd(Ramey$ED3_TC,na.rm=TRUE),T=60,y.0=rep(0,length(x$
                  X=NaN,beta=NaN)
  return(irf)
}

IRF.0 <- 100*irf.function(x$THETA)
eps <- .00000001
d.IRF <- NULL
for(i in 1:length(x$THETA)){
  THETA.i <- x$THETA
  THETA.i[i] <- THETA.i[i] + eps
  IRF.i <- 100*irf.function(THETA.i)
  d.IRF <- cbind(d.IRF,
                 (IRF.i - IRF.0)/eps
                 )
}
mat.var.cov.IRF <- d.IRF %*% x$I %*% t(d.IRF)
```

Bibliography

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.

Cameron, A. C. and Miller, D. L. (2014). A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50(2).

Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245):32–61.

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

Gouriéroux, C. and Monfort, A. (1995). *Statistics and Econometric Models*, volume 1 of *Themes in Modern Econometrics*. Cambridge University Press.

Greene, W. H. (2003). *Econometric Analysis*. Pearson Education, fifth edition.

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.

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

Stock, J. and Watson, M. W. (2003). *Introduction to Econometrics*. Prentice Hall, New York.

Stock, J. H. and Yogo, M. (2005). *Testing for Weak Instruments in Linear IV Regression*, page 80–108. Cambridge University Press.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B*, 73(3):273–282.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

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