Micro-Econometrics

Jean-Paul Renne

2023-01-28

Contents

Micro-Econometrics

In microeconometric models, the variables of interest often feature restricted distributions—for instance with discontinuous support—, which necessitates specific models. Typical examples are discrete-choice models (binary, multinomial, ordered outcomes), sample selection models (censored or truncated outcomes), and count-data models (integer outcomes). The course describes the estimation and interpretation of these models. It also shows how the discrete-choice models can emerge from (structural) random-utility frameworks.

The R codes use various packages that can be obtained from CRAN. This `AEC` package is available on GitHub. To install it, one need to employ the `devtools` library:

```
install.packages("devtools") # in case this library has not been loaded yet
library(devtools)
install_github("jrenne/AEC")
library(AEC)
```

Useful (R) links:

- Download R:
  - R software: https://cran.r-project.org (the basic R software)
  - RStudio: https://www.rstudio.com (a convenient R editor)
- Tutorials:
  - Rstudio: https://dss.princeton.edu/training/RStudio101.pdf (by Oscar Torres-Reyna)

- R: https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf (by Emmanuel Paradis)
- My own tutorial: https://jrenne.shinyapps.io/Rtuto_publiShiny/

Chapter 1

Panel regressions

1.1 Specification and notations

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

The linear panel regression model is:

$$y_{i,t} = \mathbf{x}'_{i,t} \underbrace{\beta}_{K\times 1} + \underbrace{\mathbf{z}'_i \alpha}_{\text{Individual effects}} + \varepsilon_{i,t}. \tag{1.1}$$

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

Figure 1.1 illustrates a panel-data situation. The model is $y_i = \alpha_i + \beta x_{i,t} + \varepsilon_{i,t}$, $t \in \{1, 2\}$. On Panel (b), blue dots are for $t = 1$, red dots are for $t = 2$. The lines relate the dots associated with the same entity i. What is remarkable in the simulated model is that, while the unconditional correlation between y and x is negative, the conditional correlation (conditional on α_i) is positive. Indeed, the sign of this conditional correlation is the sign of β, which is positive in th simulated example ($\beta = 5$). In other words, if one did not know the panel nature of the data, that would be tempting to say that $\beta < 0$, but this is not the case, due to **fixed effects** (the α_i's) that are negatively correlated to the x_i's.

```
T <- 2; n <- 12 # 2 periods and 12 entities
alpha <- 5*rnorm(n) # draw fixed effects
x.1 <- rnorm(n) - .5*alpha # note: x_i's correlate to alpha_i's
x.2 <- rnorm(n) - .5*alpha
beta <- 5; sigma <- .3
y.1 <- alpha + x.1 + sigma*rnorm(n);y.2 <- alpha + x.2 + sigma*rnorm(n)
x <- c(x.1,x.2) # pooled x
y <- c(y.1,y.2) # pooled y
par(mfrow=c(1,2))
plot(x,y,col="black",pch=19,xlab="x",ylab="y",main="(a)")
plot(x,y,col="black",pch=19,xlab="x",ylab="y",main="(b)")
points(x.1,y.1,col="blue",pch=19);points(x.2,y.2,col="red",pch=19)
for(i in 1:n){lines(c(x.1[i],x.2[i]),c(y.1[i],y.2[i]))}
```

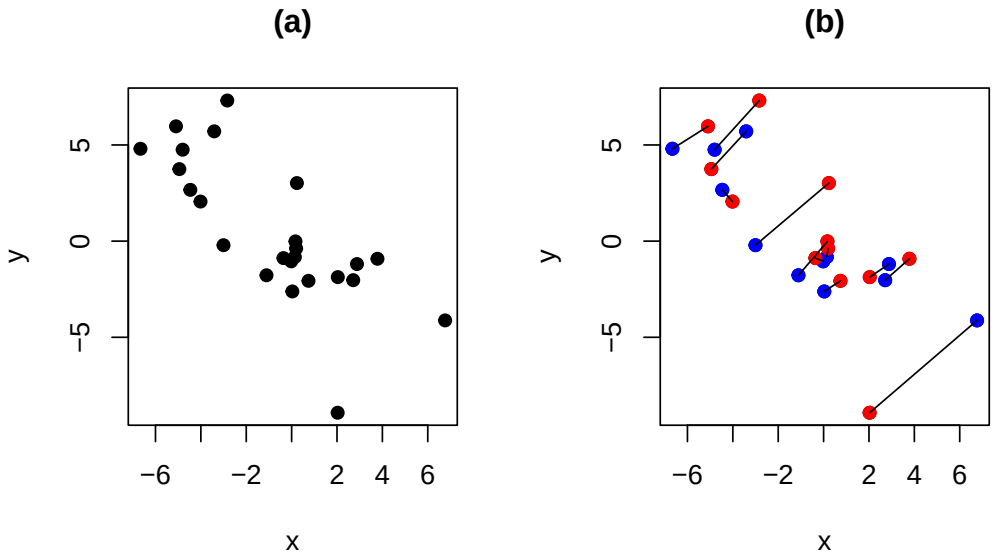


Figure 1.1: The data are the same for both panels. On Panel (b), blue dots are for $t = 1$, red dots are for $t = 2$. The lines relate the dots associated to the same entity i.

Figure 1.2 presents the same type of plot based on the Cigarette Consumption Panel dataset (`CigarettesSW` dataset, used in Stock and Watson (2003)). This dataset documents the average consumption of cigarettes in 48 continental US states for two dates (1985 and 1995).

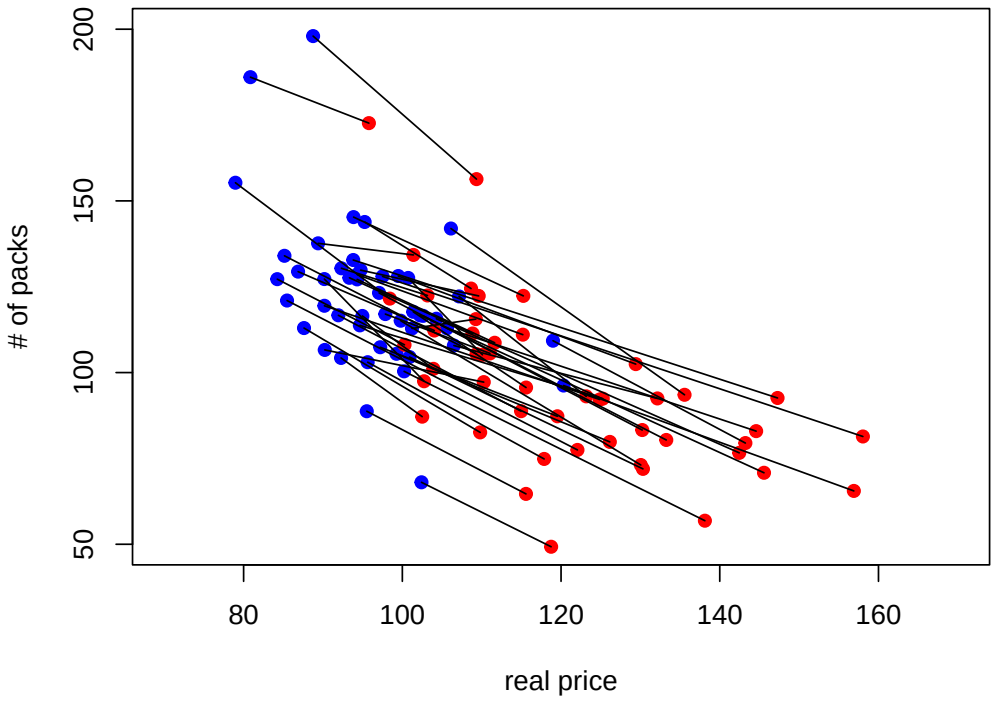


Figure 1.2: Cigarette consumption versus real price in the CigarettesSW panel dataset.

We will make use of the following notations:

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

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

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

where

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

1.2 Three standard cases

There are three typical situations:

- **Pooled regression**: $\mathbf{z}_i \equiv 1$. This case amounts to the case studied in Chapter **??**.
- **Fixed Effects** (Section 1.3): $\mathbf{z}_i$ is unobserved, but correlates with $\mathbf{x}_i \Rightarrow \mathbf{b}$ is biased and inconsistent in the OLS regression of $\mathbf{y}$ on $\mathbf{X}$ (omitted variable, see Section **??**).
- **Random Effects** (Section 1.4): $\mathbf{z}_i$ is unobserved, but uncorrelated with $\mathbf{x}_i$. The model writes:

$$y_{i,t} = \mathbf{x}_{i,t}'\beta + \alpha + \underbrace{u_i + \varepsilon_{i,t}}_{\text{compound error}},$$

where $\alpha = \mathbb{E}(\mathbf{z}_i'\alpha)$ and $u_i = \mathbf{z}_i'\alpha - \mathbb{E}(\mathbf{z}_i'\alpha) \perp \mathbf{x}_i$. In that case, the OLS is consistent, but not efficient. GLS can be used to gain efficiencies over OLS (see Section **??** for a presentation of the GLS approach).

1.3 Estimation of Fixed-Effects Models

Hypothesis 1.1 (Fixed-effect model)**.** We assume that:

i. There is no perfect multicollinearity among the regressors.
ii. $\mathbb{E}(\varepsilon_{i,t}|\mathbf{X}) = 0$, for all i, t.
iii. We have:

$$\mathbb{E}(\varepsilon_{i,t}\varepsilon_{j,s}|\mathbf{X}) = \begin{cases} \sigma^2 & \text{if } i = j \text{ and } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

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

(i) $\leftrightarrow$ Hyp. **??**, (ii) $\leftrightarrow$ Hyp. **??**, (iii) $\leftrightarrow$ Hyp. **??** + **??**.

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

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

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

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

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{D}] \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \varepsilon, \tag{1.2}$$

with:

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

The linear regression (Eq. (1.2)) —with the dummy variables— satisfies the Gauss-Markov conditions (Theorem **??**). Hence, in this context, the OLS estimator is the *best linear unbiased estimator* (BLUE).

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

$$\boxed{\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{y}.} \tag{1.3}$$

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

We have:

$$\boxed{\begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} \stackrel{d}{\to} \mathcal{N}\left(\begin{bmatrix} \beta \\ \alpha \end{bmatrix}, \sigma^2 \frac{Q^{-1}}{nT}\right),} \tag{1.4}$$

where

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

assuming the previous limit exists.

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

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

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

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

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

With these notations, using the Frisch-Waugh theorem (Theorem **??**), we get another expression for the estimator $\mathbf{b}$ appearing in Eq. (1.3):

$$\boxed{\mathbf{b} = [\mathbf{X}'\mathbf{M_D}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{M_D}\mathbf{y}.} \tag{1.5}$$

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

The estimate of α is given by:

$$\boxed{\mathbf{a} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'(\mathbf{y} - \mathbf{X}\mathbf{b}),} \tag{1.6}$$

which is obtained by developing the second row of

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

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

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

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

The LSDV approach (Eq. (1.2)) can still be resorted to. It suffices to extend the $\mathbf{Z}$ matrix with additional columns (then called *time dummies*):

$$\mathbf{y} = [\mathbf{X} \quad \mathbf{D} \quad \mathbf{C}] \begin{bmatrix} \beta \\ \alpha \\ \gamma \end{bmatrix} + \varepsilon, \tag{1.7}$$

with:

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

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

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

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

```
CigarettesSW$rincome <- with(CigarettesSW, income/population/cpi)
eq.pooled <- lm(log(packs)~log(rprice)+log(rincome),data=CigarettesSW)
eq.LSDV <- lm(log(packs)~log(rprice)+log(rincome)+state,
              data=CigarettesSW)
CigarettesSW$year <- as.factor(CigarettesSW$year)
eq.LSDV2 <- lm(log(packs)~log(rprice)+log(rincome)+state+year,
               data=CigarettesSW)
```

```
stargazer::stargazer(eq.pooled,eq.LSDV,eq.LSDV2,type="text",no.space = TRUE,
                     omit=c("state","year"),
                     add.lines=list(c('State FE','No','Yes','Yes'),
                                    c('Year FE','No','No','Yes')),
                     omit.stat=c("f","ser"))
```

```
##
## ==========================================
##                    Dependent variable:
##               ---------------------------
##                         log(packs)
##                  (1)       (2)       (3)
## ------------------------------------------
## log(rprice)  -1.334*** -1.210*** -1.056***
##                (0.135)   (0.114)   (0.149)
## log(rincome)  0.318**    0.121     0.497
##                (0.136)   (0.190)   (0.304)
## Constant     10.067*** 9.954***  8.360***
##                (0.516)   (0.264)   (1.049)
## ------------------------------------------
## State FE         No        Yes       Yes
## Year FE          No        No        Yes
## Observations     96        96        96
## R2              0.552     0.966     0.967
## Adjusted R2     0.542     0.929     0.931
## ==========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

Example 1.1 (Housing prices and interest rates)**.** In this example, we want to estimate the effect of short and long-term interest rate on housing prices. The data come from the Jordà et al. (2017) dataset (see this website).

```
library(AEC);library(sandwich)
data(JST); JST <- subset(JST,year>1950);N <- dim(JST)[1]
JST$hpreal <- JST$hpnom/JST$cpi # real house price index
JST$dhpreal <- 100*log(JST$hpreal/c(NaN,JST$hpreal[1:(N-1)]))
# Put NA's when change in country:
```

```
JST$dhpreal[c(0,JST$iso[2:N]!=JST$iso[1:(N-1)])] <- NaN
JST$dhpreal[abs(JST$dhpreal)>30] <- NaN # remove extreme price change
JST$YEAR <- as.factor(JST$year) # to have time fixed effects
eq1_noFE <- lm(dhpreal ~ stir + ltrate,data=JST)
eq1_FE   <- lm(dhpreal ~ stir + ltrate + iso + YEAR,data=JST)
eq2_noFE <- lm(dhpreal ~ I(ltrate-stir),data=JST)
eq2_FE <- lm(dhpreal ~ I(ltrate-stir) + iso + YEAR,data=JST)
vcov_cluster1_noFE <- vcovHC(eq1_noFE, cluster = JST[, c("iso","YEAR")])
vcov_cluster1_FE   <- vcovHC(eq1_FE, cluster = JST[, c("iso","YEAR")])
vcov_cluster2_noFE <- vcovHC(eq2_noFE, cluster = JST[, c("iso","YEAR")])
vcov_cluster2_FE   <- vcovHC(eq2_FE, cluster = JST[, c("iso","YEAR")])
robust_se_FE1_noFE <- sqrt(diag(vcov_cluster1_noFE))
robust_se_FE1_FE   <- sqrt(diag(vcov_cluster1_FE))
robust_se_FE2_noFE <- sqrt(diag(vcov_cluster2_noFE))
robust_se_FE2_FE   <- sqrt(diag(vcov_cluster2_FE))
stargazer::stargazer(eq1_noFE, eq1_FE, eq2_noFE, eq2_FE, type = "text",
                     column.labels = c("no FE", "with FE", "no FE","with FE"),
                     omit = c("iso","YEAR","Constant"),keep.stat = "n",
                     add.lines=list(c('Country FE','No','Yes','No','Yes'),
                                    c('Year FE','No','Yes','No','Yes')),
                     se = list(robust_se_FE1_noFE,robust_se_FE1_FE,
                               robust_se_FE2_noFE,robust_se_FE2_FE))
```

```
##
## ===================================================
##                           Dependent variable:
##                     ------------------------------
##                                 dhpreal
##                       no FE   with FE    no FE    with FE
##                         (1)      (2)       (3)       (4)
## ---------------------------------------------------
## stir               0.485***  0.532***
##                     (0.149)  (0.170)
##
## ltrate             -0.690*** -0.384**
##                     (0.164)  (0.182)
##
```

```
## I(ltrate - stir)                      -0.476*** -0.475***
##                                        (0.145)   (0.159)
##
## ------------------------------------------------------
## Country FE          No       Yes       No        Yes
## Year FE             No       Yes       No        Yes
## Observations       1,141    1,141     1,141     1,141
## ======================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

1.4 Estimation of random effects models

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

Random-effect models write:

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

with

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

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

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

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

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

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

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

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

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

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

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

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

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

$$y_{i,t} - \bar{y}_i = [\mathbf{x}_{i,t} - \bar{\mathbf{x}}_i]'\beta + (\varepsilon_{i,t} - \bar{\varepsilon}_i). \tag{1.8}$$

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

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

The $\varepsilon_{i,t}$'s are not observed but $\mathbf{b}$, the OLS estimator of β in Eq. (1.8), is a consistent estimator of β. Using an adjustment for the degrees of freedom, we can approximate their variance with:

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

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

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

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

Let us come back to Example 1.1 (relationship between changes in housing prices and interest rates). In the following, we use the random effect specification; and compare the results with those obtained with the pooled regression and with the fixed-effect model. For that, we use the function `plm` of the package of the same name. (Note that `eq.FE` is similar to `eq1` in Example 1.1.)

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

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

```
##                (0.127)     (0.020)       (0.145)
## Constant     4.103***    3.341***
##                (0.421)     (0.096)
## ------------------------------------------------
## State FE         No          Yes           Yes
## Year FE          No          Yes           Yes
## Observations   1,141       1,141         1,141
## R2             0.027       0.024         0.015
## Adjusted R2    0.025       0.022        -0.067
## ================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

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

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

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

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

Example 1.2 (Spatial data)**.** This example makes use of Airbnb prices (Zürich, 22 June 2017), collected from Tom Slee's website. The covariates are the number of bedrooms and the number of people that can be accommodated. We consider the use of district fixed effects. Figure 1.3 shows the price to explain (the size of the circles is proportional to the prices). The white lines delineate the 12 districts of the city.

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

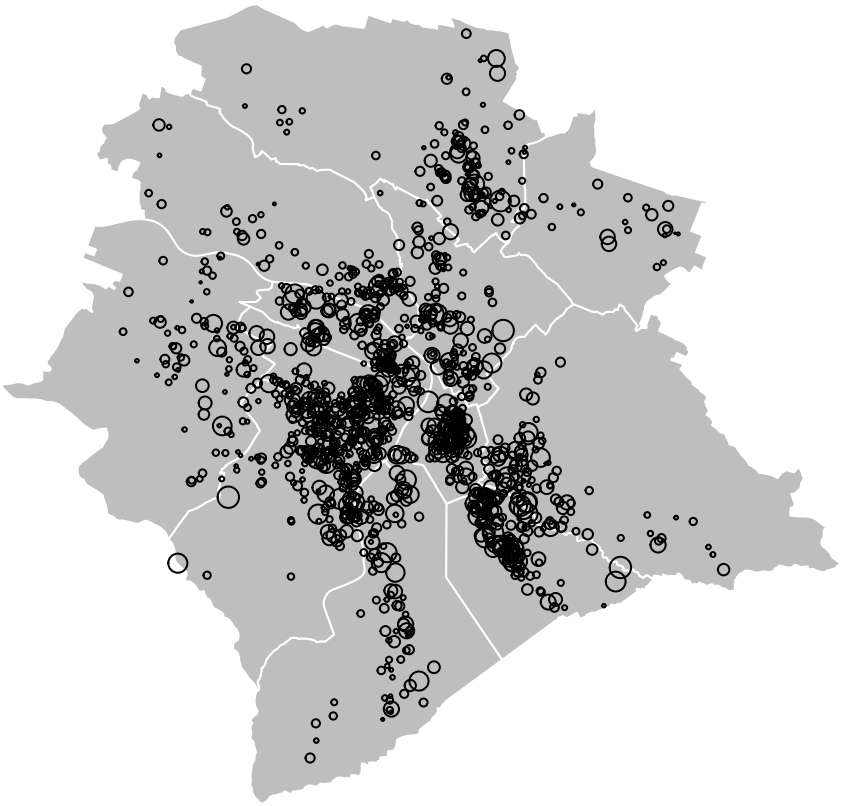


Figure 1.3: Airbnb prices for the Zurich area, 22 June 2017. The size of the circles is proportional to the prices. White lines delineate the 12 districts of the city.

```
eq_noFE <- lm(price~bedrooms+accommodates,data=airbnb)
eq_FE   <- lm(price~bedrooms+accommodates+neighborhood,data=airbnb)
# Adjust standard errors:
cov_FE          <- vcovHC(eq_FE, cluster = airbnb[, c("neighborhood")])
robust_se_FE    <- sqrt(diag(cov_FE))
cov_noFE        <- vcovHC(eq_noFE, cluster = airbnb[, c("neighborhood")])
robust_se_noFE  <- sqrt(diag(cov_noFE))
stargazer::stargazer(eq_FE, eq_noFE, eq_FE, eq_noFE, type = "text",
                     column.labels = c("FE (no HAC)", "No FE (no HAC)",
                                       "FE (with HAC)", "No FE (with HAC)"),
                     omit = c("neighborhood"),no.space = TRUE,
                     omit.labels = c("District FE"),keep.stat = "n",
                     se = list(NULL, NULL, robust_se_FE, robust_se_noFE))
```

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

Figure 1.4 compares the residuals with and without fixed effects. The sizes of the circles are proportional to the absolute values of the residuals, the color

indicates the sign (blue for positive).

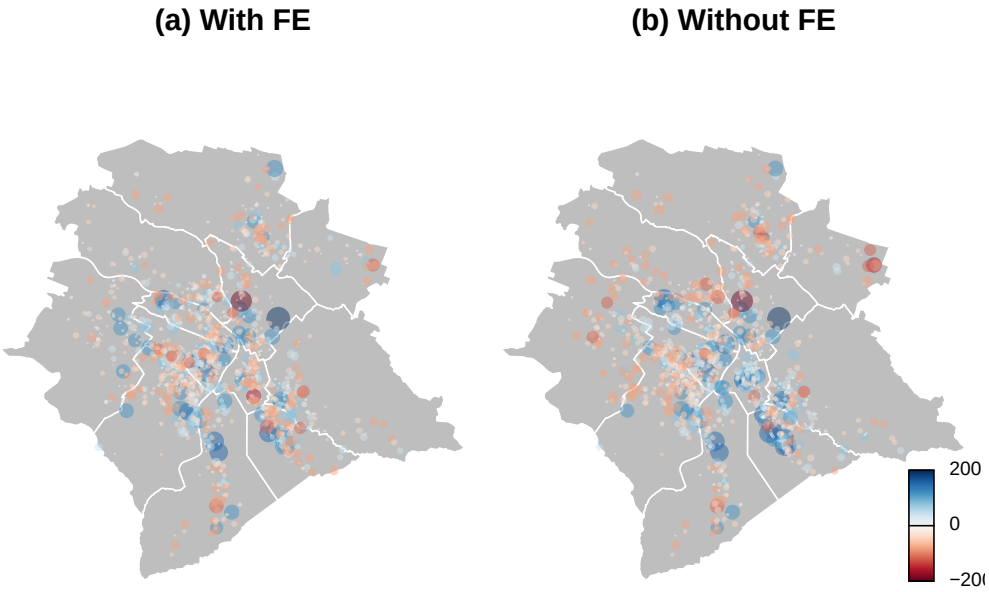


Figure 1.4: Regression residuals. The sizes of the circles are proportional to the absolute values of the residuals, the color indicates the sign (blue for negative).

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

1.5 Dynamic Panel Regressions

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

In several cases, auto-correlation in the variable of interest may stem from an auto-regressive specification. That is, Eq. (1.1) is then replaced by:

$$y_{i,t} = \rho y_{i,t-1} + \mathbf{x}_{i,t}' \underbrace{\beta}_{K\times 1} + \underbrace{\alpha_i}_{\text{Individual effects}} + \varepsilon_{i,t}. \tag{1.9}$$

In that case, even if the explanatory variables $\mathbf{x}_{i,t}$ are uncorrelated to the errors $\varepsilon_{i,t}$, we have that the additional *explanatory variable* $y_{i,t-1}$ correlates to the errors $\varepsilon_{i,t-1}, \varepsilon_{i,t-2}, \dots, \varepsilon_{i,1}$. As a result, the LSDV estimate of the model parameters $\{\rho, \beta\}$ may be biased, even if n is large. To see this, notice that the LSDV regression amounts to regressing $\tilde{\mathbf{y}}$ on $\widetilde{\mathbf{X}}$ (see Eq. (1.5)), where the elements of $\widetilde{\mathbf{X}}$ are the explanatory variables to which we subtract their within-sample means. In particular, we have:

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

which correlates to the corresponding error, that is:

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

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

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

```
n <- 400;T <- 10
rho <- 0.8;sigma <- .5
alpha <- rnorm(n)
y <- alpha /(1-rho) + sigma^2/(1 - rho^2) * rnorm(n)
all_y <- y
for(t in 2:T){
  y <- rho * y + alpha + sigma * rnorm(n)
  all_y <- rbind(all_y,y)
}
```

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

```
y   <- c(all_y[2:T,]);y_1 <- c(all_y[1:(T-1),])
D <- diag(n) %x% rep(1,T-1)
Z <- cbind(c(y_1),D)
b <- solve(t(Z) %*% Z) %*% t(Z) %*% y
a <- b[2:(n+1)]
plot(alpha,a)
lines(c(-10,10),c(-10,10),col="blue")
```

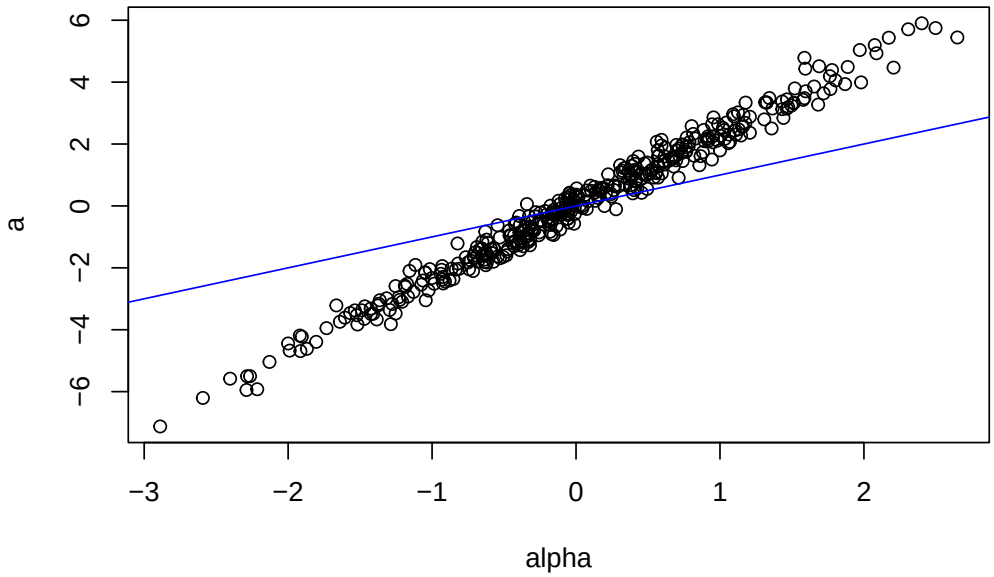


Figure 1.5: illustration of the bias pertianing to the LSDV estimation approach in the presence of auto-correlation of the depend variable.

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

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

$$\Delta y_{i,t} = \rho \Delta y_{i,t-1} + (\Delta \mathbf{x}_{i,t})' \beta + \Delta \varepsilon_{i,t}. \tag{1.10}$$

The OLS estimates of the parameters are biased because $\varepsilon_{i,t-1}$ —which is part of the error $\Delta\varepsilon_{i,t}$— is correlated to $y_{i,t-1}$ —which is part of the "explanatory variable", namely $\Delta y_{i,t-1}$. But consistent estimates can be obtained using 2SLS with instrumental variables that are correlated with $\Delta y_{i,t}$

but orthogonal to $\Delta\varepsilon_{i,t}$. One can for instance use $\{y_{i,t-2}, \mathbf{x}_{i,t-2}\}$ as instruments. Note that this approach can be implemented only if there are more than 3 time observations per entity i.

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

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

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

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

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

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

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

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

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

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

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

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

using the weighting matrix

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

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

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

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

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

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

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

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

```
library(pdynmc)
data(EmplUK, package = "plm")
dat <- EmplUK
dat[,c(4:7)]         <- log(dat[,c(4:7)])
m1 <- pdynmc(dat = dat, # name of the dataset
             varname.i = "firm", # name of the cross-section identifier
             varname.t = "year", # name of the time-series identifiers
             use.mc.diff = TRUE, # use moment conditions from equations in differenc
             use.mc.lev = FALSE, # use moment conditions from equations in levels? (
             use.mc.nonlin = FALSE, # use nonlinear (quadratic) moment conditions?
             include.y = TRUE, # instruments should be derived from the lags of the
             varname.y = "emp", # name of the dependent variable in the dataset
             lagTerms.y = 2, # number of lags of the dependent variable
             fur.con = TRUE, # further control variables (covariates) are included?
             fur.con.diff = TRUE, # include further control variables in equations f
             fur.con.lev = FALSE, # include further control variables in equations f
             varname.reg.fur = c("wage", "capital", "output"), # covariate(s) -in the
             lagTerms.reg.fur = c(1,2,2), # number of lags of the further controls
             include.dum = TRUE, # A logical variable indicating whether dummy varia
             dum.diff = TRUE, # A logical variable indicating whether dummy variable
             dum.lev = FALSE, # A logical variable indicating whether dummy variable
             varname.dum = "year",
             w.mat = "iid.err", # One of the character strings c('"iid.err"', '"iden
             std.err = "corrected",
             estimation = "onestep", # One of the character strings c('"onestep"', '
             opt.meth = "none" # numerical optimization procedure. When no nonlinear
)
summary(m1,digits=3)
```

```
##
## Dynamic linear panel estimation (onestep)
## Estimation steps: 1
##
## Coefficients:
##             Estimate Std.Err.rob z-value.rob Pr(>|z.rob|)
## L1.emp      0.686226    0.144594       4.746      < 2e-16 ***
## L2.emp     -0.085358    0.056016      -1.524      0.12751
```

```
## L0.wage    -0.607821    0.178205      -3.411      0.00065 ***
## L1.wage     0.392623    0.167993       2.337      0.01944 *
## L0.capital  0.356846    0.059020       6.046      < 2e-16 ***
## L1.capital -0.058001    0.073180      -0.793      0.42778
## L2.capital -0.019948    0.032713      -0.610      0.54186
## L0.output   0.608506    0.172531       3.527      0.00042 ***
## L1.output  -0.711164    0.231716      -3.069      0.00215 **
## L2.output   0.105798    0.141202       0.749      0.45386
## 1979        0.009554    0.010290       0.929      0.35289
## 1980        0.022015    0.017710       1.243      0.21387
## 1981       -0.011775    0.029508      -0.399      0.68989
## 1982       -0.027059    0.029275      -0.924      0.35549
## 1983       -0.021321    0.030460      -0.700      0.48393
## 1976       -0.007703    0.031411      -0.245      0.80646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  41 total instruments are employed to estimate 16 parameters
##  27 linear (DIF)
##  8 further controls (DIF)
##  6 time dummies (DIF)
##
## J-Test (overid restrictions):  70.82 with 25 DF, pvalue: <0.001
## F-Statistic (slope coeff):  528.06 with 10 DF, pvalue: <0.001
## F-Statistic (time dummies):  14.98 with 6 DF, pvalue: 0.0204
```

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

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

Arellano and Bond (1991) have proposed a specification test. If the model is correctly specified, then the errors of Eq. (1.10) —that is the first-difference

equation— should feature non-zero first-order auto-correlations, but zero higher-order autocorrelations.

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

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

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

One can also implement the Hansen J-test of the over-identifying restrictions (see Section 2.1.3):

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

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

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

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

1.6 Introduction to program evaluation

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

1.6.1 Presentation of the problem

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

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

$$Y_1 - Y_0,$$

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

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

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

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

Importantly, ATE is different from the following quantity:

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

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

$$\mathbb{E}(Y_1) = \mathbb{E}(Y_1|W=0)\mathbb{P}(W=0) + \mathbb{E}(Y_1|W=1)\mathbb{P}(W=1) \quad (1.11)$$
$$\mathbb{E}(Y_0) = \mathbb{E}(Y_0|W=0)\mathbb{P}(W=0) + \mathbb{E}(Y_0|W=1)\mathbb{P}(W=1). \quad (1.12)$$

1.6.2 Randomized controlled trials (RCTs)

In the context of Randomized controlled trials (RCTs), entities are randomly assigned to receive the treatment. As a result, we have $\mathbb{E}(Y_1) = \mathbb{E}(Y_1|W=0) = \mathbb{E}(Y_1|W=1)$ and $\mathbb{E}(Y_0) = \mathbb{E}(Y_0|W=0) = \mathbb{E}(Y_0|W=1)$. Using this into Eqs. (1.11) and (1.12) yields $ATE = \alpha$.

Therefore, in this context, estimating $\mathbb{E}(Y_1 - Y_0)$ amounts to computing the difference between two sample means, namely (a) the sample mean of the subset of Y_i's corresponding to the entities for which $W_i = 1$, and (b) the one for which $W_i = 0$.

More accurate estimates can be obtained through regressions. Assume that the model reads:

$$Y_i = W_i\beta_1 + X_i'\beta_z + \varepsilon_i,$$

where $\mathbb{E}(\varepsilon_i|X_i) = 0$ (and W_i is independent from X_i and ε_i). In this case, we obtain a consistent estimate of β_1 by regressing $\mathbf{y}$ on $\mathbf{Z} = [\mathbf{w}, \mathbf{X}]$.

1.6.3 Difference-in-Difference (DiD) approach

The DiD approach is a popular methodology implemented in cases where W cannot be considered as an independent variable. It exploits two dimensions: entities (i), and time (t). To simplify the exposition, we consider only two periods here ($t = 0$ and $t = 1$).

Consider the following model:

$$Y_{i,t} = W_{i,t}\beta_1 + \mu_i + \delta_t + \varepsilon_{i,t} \tag{1.13}$$

The parameter of interest is β_1, which is the treatment effect (recall that $W_{i,t} \in \{0,1\}$). Usually, for all entities i, we have $W_{i,t=0} = 0$. But only some of them are treated on date 1, i.e., $W_{i,1} \in \{0,1\}$.

The disturbance $\varepsilon_{i,t}$ affects the outcome, but we assume that it does not relate to the selection for treatment; therefore, $\mathbb{E}(\varepsilon_{i,t}|W_{i,t}) = 0$. By contrast, we do not exclude some correlation between $W_{i,t}$ (for $t = 1$) and μ_i; hence, μ_i may constitute a *confounder*. Finally, we suppose that the micro-variables W_i do not affect the time fixed effects δ_t, such that $\mathbb{E}(\delta_t|W_{i,t}) = \mathbb{E}(\delta_t)$.

We have:

$$\begin{array}{lcccccccc}
\mathbb{E}(Y_{i,1}|W_{i,1}=1) & = & \beta_1 & + & \mathbb{E}(\mu_i|W_{i,1}=1) & + & \mathbb{E}(\delta_1|W_{i,1}=1) & + & \mathbb{E}(\varepsilon_{i,1}) \\
\mathbb{E}(Y_{i,0}|W_{i,1}=1) & = & & & \mathbb{E}(\mu_i|W_{i,1}=1) & + & \mathbb{E}(\delta_0|W_{i,1}=1) & + & \mathbb{E}(\varepsilon_{i,0}) \\
\mathbb{E}(Y_{i,1}|W_{i,1}=0) & = & & & \mathbb{E}(\mu_i|W_{i,1}=0) & + & \mathbb{E}(\delta_1|W_{i,1}=0) & + & \mathbb{E}(\varepsilon_{i,1}) \\
\mathbb{E}(Y_{i,0}|W_{i,1}=0) & = & & & \mathbb{E}(\mu_i|W_{i,1}=0) & + & \mathbb{E}(\delta_0|W_{i,1}=0) & + & \mathbb{E}(\varepsilon_{i,0}).
\end{array}$$

and, under our assumptions, it can be checked that:

$$\beta_1 = \mathbb{E}(\Delta Y_{i,1}|W_{i,1}=1) - \mathbb{E}(\Delta Y_{i,1}|W_{i,1}=0),$$

where $\Delta Y_{i,1} = Y_{i,1} - Y_{i,0}$. Therefore, in this context, the treatment effect appears to be a difference (of two conditionnal expectations) of difference (of the outcome variable, through time).

This is illustrated by Figure 1.6, which represents the generic DiD framework.

In practice, implementing this approach consists in running a linear regression of the type of Eq. (1.13). These regressions also usually involve controls on top of the fixed effects μ_i. As illustrated in the next subsection, the parameter of interest (β_1) is often associated with an interaction term.

1.6.4 Application of the DiD approach

This example is based on the data used in Meyer et al. (1995). This dataset is part of the `wooldridge` package. This paper examines the effect of workers' compensation for injury on time out of work. It exploits a **natural**

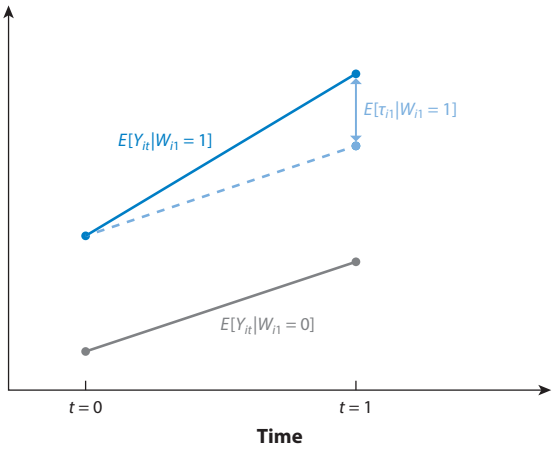


Figure 5

Identification in a difference-in-differences model. The dashed line represents the outcome that the treated units would have experienced in the absence of the treatment.

Figure 1.6: Source: Abadie et al., (1998).

experiment approach of comparing individuals injured before and after increases in the maximum weekly benefit amount. Specifically, in 1980, the cap on weekly earnings covered by worker's compensation was increased in Kentucky and Michigan. Let us check whether this new policy was followed by an increase in the amount of time workers spent unemployed (for example, higher compensation may reduce workers' incentives to avoid injury).

As shown in Figure 1.7, the measure has only affected high-earning workers. The idea exploited by Meyer et al. (1995) was to compare the increase in time out of work before-after 1980 for higher-earnings workers on the one hand (entities who received the treatment) and low-earnings workers on the other hand (control group).

The next lines of codes replicate some of their results. The dependent variable is the logarithm of the duration of benefits. For more information use `?injury`, after having loaded the `wooldridge` library.

In the table of results below, the parameter of interest is the one associated with the interaction term `afchnge:highearn`. Columns 2 and 3 correspond to the first two column of Table 6 in Meyer et al. (1995).

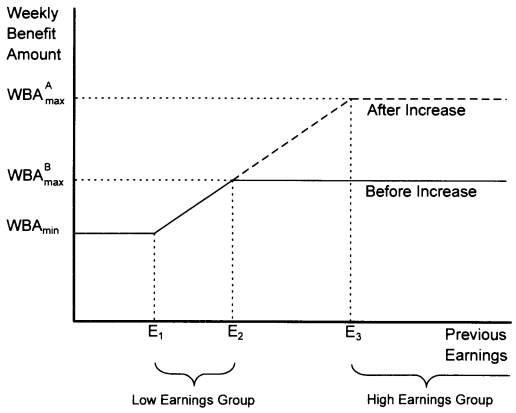


FIGURE 1. TEMPORARY TOTAL BENEFIT SCHEDULE BEFORE AND AFTER AN INCREASE IN THE MAXIMUM WEEKLY BENEFIT

Figure 1.7: Source: Meyer et al., (1995).

```
library(wooldridge)
data(injury)
injury <- subset(injury,ky==1)
injury$indust <- as.factor(injury$indust)
injury$injtype <- as.factor(injury$injtype)
#names(injury)
eq1 <- lm(log(durat) ~ afchnge + highearn + afchnge*highearn,data=injury)
eq2 <- lm(log(durat) ~ afchnge + highearn + afchnge*highearn +
            lprewage*highearn + male + married + lage + ltotmed + hosp +
            indust + injtype,data=injury)
eq3 <- lm(log(durat) ~ afchnge + highearn + afchnge*highearn +
            lprewage*highearn + male + married + lage + indust +
            injtype,data=injury)
stargazer::stargazer(eq1,eq2,eq3,type="text",
                     omit=c("indust","injtype","Constant"),no.space = TRUE,
                     add.lines = list(c("industry dummy","no","yes","yes"),
                                      c("injury dummy","no","yes","yes")),
                     order = c(1,2,18,3:17,19,20),omit.stat = c("f","ser"))
```

```
##
```

```
## ================================================
##                          Dependent variable:
##                     ----------------------------
##                               log(durat)
##                        (1)       (2)       (3)
## ------------------------------------------------
## afchnge               0.008    -0.004     0.016
##                      (0.045)   (0.038)   (0.045)
## highearn            0.256***   -0.595    -1.522
##                      (0.047)   (0.930)   (1.099)
## afchnge:highearn    0.191***  0.162***  0.215***
##                      (0.069)   (0.059)   (0.069)
## lprewage                       0.207**   0.258**
##                                (0.088)   (0.104)
## male                           -0.070*   -0.072
##                                (0.039)   (0.046)
## married                         0.055     0.051
##                                (0.035)   (0.041)
## lage                           0.244***  0.252***
##                                (0.044)   (0.052)
## ltotmed                        0.361***
##                                (0.011)
## hosp                           0.252***
##                                (0.044)
## highearn:lprewage               0.065     0.232
##                                (0.158)   (0.187)
## ------------------------------------------------
## industry dummy        no         yes       yes
## injury dummy          no         yes       yes
## Observations         5,626      5,347     5,347
## R2                   0.021      0.319     0.049
## Adjusted R2          0.020      0.316     0.046
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Chapter 2

Estimation Methods

This chapter presents three approaches to estimate parametric models: the General Method of Moments (GMM), the Maximum Likelihood approach (ML), and the Bayesian approach. The general context is the following: You observe a sample $\mathbf{y} = \{y_1, ..., y_n\}$, you assume that these data have been generated by a model parameterized by $\theta \in \mathbb{R}^K$, and you want to estimate this vector θ_0.

2.1 Generalized Method of Moments (GMM)

2.1.1 Definition of the GMM estimator

We denote by y_i a $p \times 1$ vector of variables; by θ an $K \times 1$ vector of parameters, and by $h(y_i; \theta)$ a continuous $r \times 1$ vector-valued function.

We denote by θ_0 the true value of θ and we assume that θ_0 satisfies:

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

We denote by $\underline{y_i}$ the information contained in the current and past observations of y_i, that is: $\underline{y_i} = \{y_i, y_{i-1}, ..., y_1\}$. We denote by $g(\underline{y_n}; \theta)$ the sample average of the $h(y_i; \theta)$ vectors, i.e.:

$$g(\underline{y_n}; \theta) = \frac{1}{n} \sum_{i=1}^{n} h(y_i; \theta).$$

The intuition behind the GMM estimator is the following: choose θ so as to make the sample moment as close as possible to their population values, that is 0.

Definition 2.1. A GMM estimator of θ_0 is given by:

$$\hat{\theta}_n = \text{argmin}_\theta \quad g(\underline{y_n};\theta)' \, W_n \, g(\underline{y_n};\theta),$$

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

In the specific case where $K = r$ (the dimension of θ is the same as that of $h(y_i;\theta)$ —or of $g(\underline{y_n};\theta)$— then $\hat{\theta}_n$ satisfies:

$$g(\underline{y_n};\hat{\theta}_n) = \mathbf{0}.$$

Under regularity and identification conditions, this estimator is consistent, that is $\hat{\theta}_n$ converges towards θ_0 in probability, which we denote by:

$$\text{plim}_n \, \hat{\theta}_n = \theta_0, \quad \text{or} \quad \hat{\theta}_n \xrightarrow{p} \theta_0, \tag{2.1}$$

i.e. $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta_0| > \varepsilon) = 0$ (this is Definition 4.16).

Definition 2.1 involves a positive definite matrix W_n. While one can take any positive definite matrix to have consistency (Eq. (2.1)), it can be shown that he GMM estimator achieves the minimum asymptotic variance when W_n is the inverse of matrix S, the latter being defined by:

$$S = Asy.\mathbb{V}ar\left(\sqrt{n}g(\underline{y_n};\hat{\theta}_n)\right).$$

In this case, W_n is said to be the *optimal weighting matrix*.

The intuition behind this result is the same that underlies Generalized Least Squares (see Section **??**), that is: it is beneficial to use a criterion in which the weights are inversely proportional to the variances of the moments.

If $h(x_i;\theta_0)$ is not correlated to $h(x_j;\theta_0)$, for $i \neq j$, then we have:

$$S = \mathbb{V}ar(h(x_i;\theta_0)),$$

which can be approximated by

$$\hat{\Gamma}_{0,n} = \frac{1}{n}\sum_{i=1}^{n} h(x_i;\hat{\theta}_n)h(x_i;\hat{\theta}_n)'.$$

In a time series context, we often have correlation between x_i and x_{i+k}, especially for small k's. In this case, and if the time series $\{y_i\}$ is covariance stationary (see Def. **??**), then we have:

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

where $\Gamma_\nu := \mathbb{E}[h(x_i;\theta_0)h(x_{i-\nu};\theta_0)']$. Matrix S is called the **long-run variance** of process $\{y_i\}$ (see Def. **??**).

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

$$\hat{\Gamma}_{\nu,n} = \frac{1}{n} \sum_{i=\nu+1}^{n} h(x_i;\hat{\theta}_n)h(x_{i-\nu};\hat{\theta}_n)',$$

then S can be approximated by the Newey and West (1987) formula (similar to Eq. (**??**)):

$$\hat{\Gamma}_{0,n} + \sum_{\nu=1}^{q} \left[1 - \frac{\nu}{q+1}\right] (\hat{\Gamma}_{\nu,n} + \hat{\Gamma}'_{\nu,n}). \tag{2.2}$$

2.1.2 Asymptotic distribution of the GMM estimator

We have:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, V),} \tag{2.3}$$

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

$$D := \mathbb{E}\left(\frac{\partial h(x_i;\theta)}{\partial \theta'}\right)\Bigg|_{\theta=\theta_0}.$$

Matrix V can be approximated by

$$\hat{V}_n = (\hat{D}_n \hat{S}_n^{-1} \hat{D}'_n)^{-1}, \tag{2.4}$$

where $\hat{S}_n$ is given by Eq. (2.2) and

$$\hat{D}'_n := \frac{\partial g(\underline{y_n};\theta)}{\partial \theta'}\Bigg|_{\theta=\hat{\theta}_n}.$$

In practice, the previous matrix is computed numerically.

2.1.3 Testing hypotheses in the GMM framework

A first important test is the one concerning the validity of the moment restrictions (Sargan-Hansen test; Sargan (1958) and Hansen (1982)). Assume that the number of restrictions imposed is larger than the number of parameters to estimate ($r > K$). In this case, the restrictions are said to be over-identifiying.

Under correct specification, we asymptotically have:

$$\sqrt{n}g(\underline{y_n};\theta_0) \sim \mathcal{N}(0,S).$$

As a result, it comes that:

$$J_n = \left(\sqrt{n}g(\underline{y_n};\theta_0)\right)' S^{-1} \left(\sqrt{n}g(\underline{y_n};\theta_0)\right) \tag{2.5}$$

asymptotically follows a χ^2 distribution. The number of degrees of freedom is equal to $r-K$. (Note that, for $r = K$, we have, as expected, $J = 0$.) That is, asymptotically:

$$J_n \sim \chi^2(r-K).$$

The GMM framework also allows to easily test linear restrictions on the parameters. First, given Eq. (2.3), Wald tests (see Eq. (**??**) in Section **??**) are readily available. Second, one can also resort to a test equivalent to the *likelihood ratio tests* (see Definition 2.8). More precisely, consider an unconstrained model and a constrained version of this model, the number of restrictions being equal to k. If the two models are estimated by considering the same moment constraints, and the same weighting matrix —using Eq. (2.4), based on the unrestricted model—, then we have that:

$$n\left[(g(\underline{y_n};\widehat{\theta}_n^*) - g(\underline{y_n};\widehat{\theta}_n)\right] \sim \chi^2(k),$$

where $\widehat{\theta}_n^*$ is the constrained estimate of θ_0.

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

Under the no-arbitrage assumption, there exists a random variable $\mathcal{M}_{t,t+1}$ (a s.d.f.) such that

$$\mathbb{E}_t(\mathcal{M}_{t,t+1}R_{t+1}) = 1$$

for any (gross) asset return R_t. In the following, R_t denotes a n_r-dimensional vector of gross returns.

We consider the following specification of the s.d.f.:

$$\mathcal{M}_{t,t+1} = 1 - \mathbf{b}_M'(F_{t+1} - \mathbb{E}_t(F_{t+1})), \tag{2.6}$$

where F_t is a vector of factors. Eq. (2.6) then reads:

$$\mathbb{E}_t([1 - \mathbf{b}_M'(F_{t+1} - \mathbb{E}_t(F_{t+1}))]R_{t+1}) = 1.$$

Assume that the date-t information set is $\mathcal{J}_t = \{\mathbf{z}_t, \mathcal{J}_{t-1}\}$, where $\mathbf{z}_t$ is a vector of variables observed on date t. (We then have $\mathbb{E}_t(\bullet) \equiv \mathbb{E}(\bullet|\mathcal{J}_t)$.)

We can use $\mathbf{z}_t$ as an instrument. Indeed, we have:

$$\begin{aligned} & \mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]) \\ = \; & \mathbb{E}(\mathbb{E}_t\{z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]\}) \\ = \; & \mathbb{E}(z_{i,t} \underbrace{\mathbb{E}_t\{\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1\}}_{1-\mathbb{E}_t(\mathcal{M}_{t,t+1}R_{t+1})=0}) = 0. \end{aligned} \tag{2.7}$$

We have then converted a conditional moment condition into a unconditional one (which we need to implement the GMM approach described above). However, at that stage, we can still not directly use the GMM formulas because of the conditional expectation $\mathbb{E}_t(F_{t+1})$ that appears in $\mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbb{E}_t(F_{t+1})\}R_{t+1} - R_{t+1} + 1]) = 0$.

To go further, let us assume that:

$$\mathbb{E}_t(F_{t+1}) = \mathbf{b}_F\mathbf{z}_t.$$

We can then easily estimate matrix $\mathbf{b}_F$ (of dimension $n_F \times n_z$) by OLS. Note here that these OLS can be seen as a special GMM case. Indeed, as was done in Eq. (2.7), we can show that, for the j^{th} component of F_t, we have:

$$\mathbb{E}([F_{j,t+1} - \mathbf{b}_{F,j}\mathbf{z}_t]\mathbf{z}_t) = 0,$$

where $\mathbf{b}_{F,j}$ denotes the j^{th} row of $\mathbf{b}_F$. This yields the OLS formula.

Equipped with $\mathbf{b}_F$, we rely on the following moment restrictions to estimate $\mathbf{b}_M$:

$$\mathbb{E}(z_{i,t}[\mathbf{b}_M'\{F_{t+1} - \mathbf{b}_F\mathbf{z}_t\}R_{t+1} - R_{t+1} + 1]) = 0.$$

Specifically, the number of restrictions is $n_R \times n_z$. Let us implement this approach in the U.S. context, using data extracted from the FRED database. In factor F_t, we use the changes in the VIX and in the personal consumption expenditures. The returns (R_t) are based on the Wilshire 5000 Price Index (a stock price index) and on the ICE BofA BBB US Corporate Index Total Return Index (a bond return index).

```
library(fredr)
fredr_set_key("df65e14c054697a52b4511e77fcfa1f3")
start_date <- as.Date("1990-01-01"); end_date <- as.Date("2022-01-01")
f <- function(ticker){
  fredr(series_id = ticker,
        observation_start = start_date,observation_end = end_date,
        frequency = "m",aggregation_method = "avg")
}
vix <- f("VIXCLS") # VIX
pce <- f("PCE") # Personal consumption expenditures
sto <- f("WILL5000PRFC") # Wilshire 5000 Full Cap Price Index
bdr <- f("BAMLCC0A4BBBTRIV") # ICE BofA BBB US Corp. Index Tot. Return
T <- dim(vix)[1]
dvix <- c(vix$value[3:T]/vix$value[2:(T-1)]) # change in VIX t+1
dpce <- c(pce$value[3:T]/pce$value[2:(T-1)]) # change in PCE t+1
dsto <- c(sto$value[3:T]/sto$value[2:(T-1)]) # return t+1
dbdr <- c(bdr$value[3:T]/bdr$value[2:(T-1)]) # return t+1
dvix_1 <- c(vix$value[2:(T-1)]/vix$value[1:(T-2)]) # change in VIX t
dpce_1 <- c(pce$value[2:(T-1)]/pce$value[1:(T-2)]) # change in PCE t
dsto_1 <- c(sto$value[2:(T-1)]/sto$value[1:(T-2)]) # return t
dbdr_1 <- c(bdr$value[2:(T-1)]/bdr$value[1:(T-2)]) # return t
```

Define the matrices containing the F_{t+1}, $\mathbf{z}_t$, and R_{t+1} vectors:

```
F_tp1 <- cbind(dvix,dpce)
Z     <- cbind(1,dvix_1,dpce_1,dsto_1,dbdr_1)
b_F <- t(solve(t(Z) %*% Z) %*% t(Z) %*% F_tp1)
F_innov <- F_tp1 - Z %*% t(b_F)
R_tp1 <- cbind(dsto,dbdr)
n_F <- dim(F_tp1)[2]; n_R <- dim(R_tp1)[2]; n_z <- dim(Z)[2]
```

Function `f_aux` compute the $h(x_t;\theta)$ and the $g(\underline{y_T};\theta)$; function `f2beMin` is the function to be minimized.

```
f_aux <- function(theta){
  b_M <- matrix(theta[1:n_F],ncol=1)
  R_aux <- matrix(F_innov %*% b_M,T-2,n_R) * R_tp1 - R_tp1 + 1
  H <- (R_aux %x% matrix(1,1,n_z)) * (matrix(1,1,n_R) %x% Z)
  g <- matrix(apply(H,2,mean),ncol=1)
  return(list(g=g,H=H))
}
f2beMin <- function(theta,W){# function to be minimized
  res <- f_aux(theta)
  return(t(res$g) %*% W %*% res$g)
}
```

Now, let's minimize this function, using use the BFGS numerical algorithm (part of the `optim` wrapper). We run 5 iterations (where W is updated).

```
library(AEC)
theta <- c(rep(0,n_F)) # inital value
for(i in 1:10){# recursion on W
  res <- f_aux(theta)
  W <-  solve(NW.LongRunVariance(res$H,q=6))
  res.optim <- optim(theta,f2beMin,W=W,
                     method="BFGS", # could be "Nelder-Mead"
                     control=list(trace=FALSE,maxit=200),hessian=TRUE)
  theta <- res.optim$par
}
```

Finally, let's compute the standard deviation of the parameter estimates, using Eq. (2.4):

```
eps <- .0001
g0 <- f_aux(theta)$g
D <- NULL
for(i in 1:length(theta)){
  theta.i <- theta
```

```
  theta.i[i] <- theta.i[i] + eps
  gi <- f_aux(theta.i)$g
  D <- cbind(D,(gi-g0)/eps)
}
V <- 1/T * solve(t(D) %*% W %*% D)
std.dev <- sqrt(diag(V));t.stud <- theta/std.dev
cbind(theta,std.dev,t.stud)
```

```
##            theta    std.dev     t.stud
## [1,]  -0.7180716  0.4646617 -1.5453642
## [2,] -11.2042452 17.1039449 -0.6550679
```

The Hansen statistic can be used to test the model (see Eq. (2.5)). If the model is correct, we have:

$$T g(\underline{y_T};\theta)' \, S^{-1} \, g(\underline{y_T};\theta) \sim \; i.i.d. \, \chi^2(J-K),$$

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

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

```
##           [,1]
## [1,] 0.8789782
```

2.2 Maximum Likelihood Estimation

2.2.1 Intuition

Intuitively, the *Maximum Likelihood Estimation (MLE)* consists in looking for the value of θ that is such that the probability of having observed $\mathbf{y}$ (the sample at hand) is the highest possible.

To set an example, assume that the time periods between the arrivals of two customers in a shop, denoted by y_i, are i.i.d. and follow an exponential distribution, i.e. $y_i \sim i.i.d.\,\mathcal{E}(\lambda)$. You have observed these arrivals for some time, thereby constituting a sample $\mathbf{y} = \{y_1, ..., y_n\}$. You want to estimate λ (i.e. in that case, the vector of parameters is simply $\theta = \lambda$).

The density of Y (one observation) is $f(y;\lambda) = \frac{1}{\lambda}\exp(-y/\lambda)$. Fig. 2.1 represents such density functions for different values of λ.

Your 200 observations are reported at the bottom of Fig. 2.1 (red bars). You build the histogram and display it on the same chart.

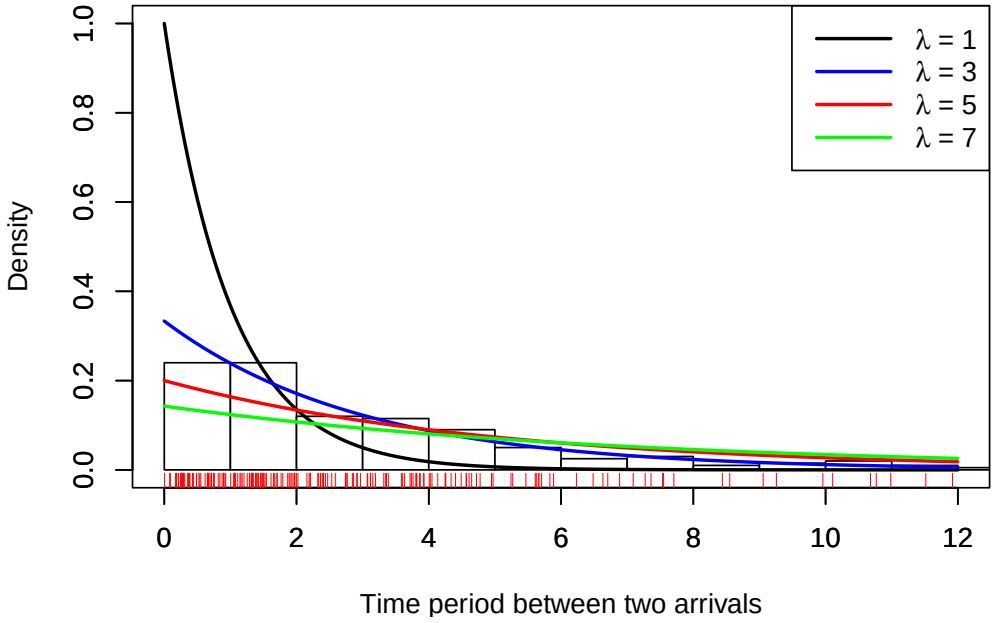


Figure 2.1: The red ticks, at the bottom, indicate observations (there are 200 of them). The historgram is based on these 200 observations

What is your estimate of λ? Intuitively, one is led to take the λ for which the (theoretical) distribution is the closest to the histogram (that can be seen as an "empirical distribution"). This approach is consistent with the idea of picking the λ for which the probability of observing the values included in $\mathbf{y}$ is the highest.

Let us be more formal. Assume that you have only four observations: $y_1 =$

1.1, $y_2 = 2.2$, $y_3 = 0.7$ and $y_4 = 5.0$. What was the probability of jointly observing:

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

Because the y_i's are i.i.d., this probability is $\prod_{i=1}^{4}(2\varepsilon f(y_i, \lambda))$. The next plot shows the probability (divided by $16\varepsilon^4$, which does not depend on λ) as a function of λ.

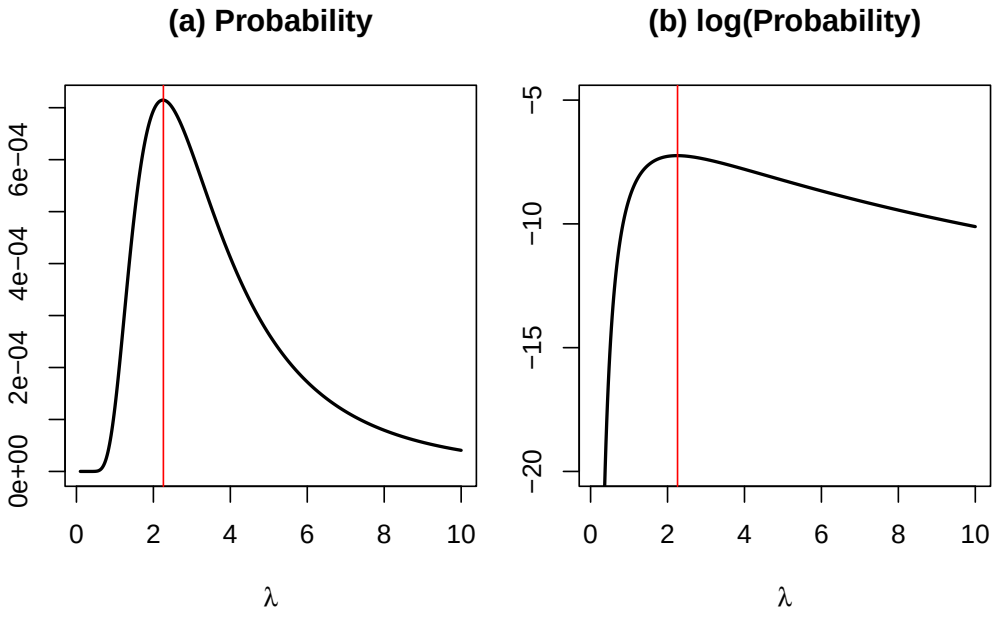


Figure 2.2: Proba. that $y_i - \varepsilon \leq Y_i < y_i + \varepsilon$, $i \in \{1, 2, 3, 4\}$. The vertical red line indicates the maximum of the function.

The value of λ that maximizes the probability is 2.26.

Let us come back to the example with 200 observations:

In that case, the value of λ that maxmimizes the probability is 3.42.

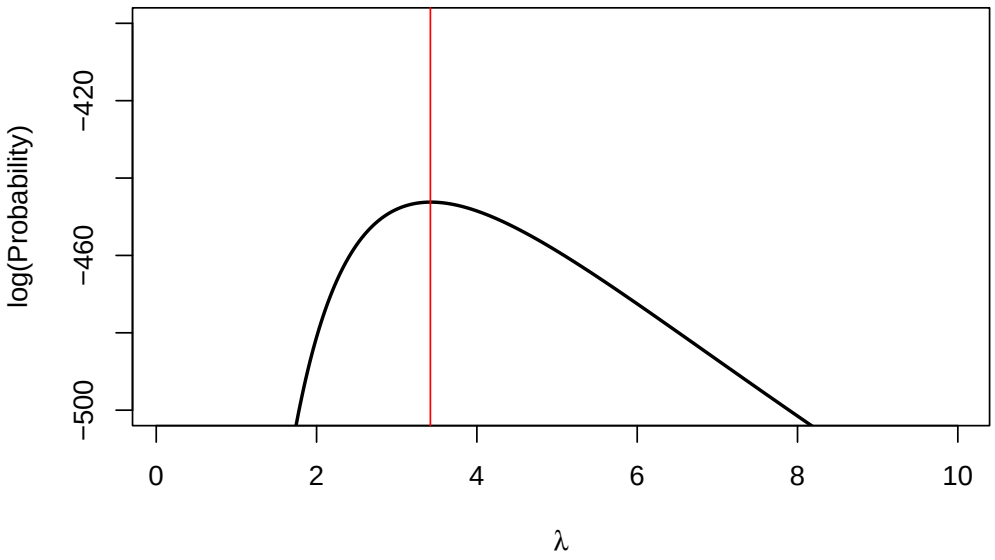


Figure 2.3: Log-likelihood function associated with the 200 i.i.d. observations. The vertical red line indicates the maximum of the function.

2.2.2 Definition and properties

$f(y;\theta)$ denotes the probability density function (p.d.f.) of a random variable Y which depends on a set of parameters θ. The density of n independent and identically distributed (i.i.d.) observations of Y is given by:

$$f(\mathbf{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta),$$

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

Definition 2.2 (Likelihood function)**.** The likelihood function is:

$$\mathcal{L} : \theta \to \mathcal{L}(\theta;\mathbf{y}) = f(\mathbf{y};\theta) = f(y_1, \dots, y_n;\theta).$$

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

Example 2.1 (Gaussian distribution)**.** If $y_i \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\log \mathcal{L}(\theta;\mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\log \sigma^2 + \log 2\pi + \frac{(y_i-\mu)^2}{\sigma^2}\right).$$

Definition 2.3 (Score)**.** The score $S(y;\theta)$ is given by $\frac{\partial \log f(y;\theta)}{\partial \theta}$.

If $y_i \sim \mathcal{N}(\mu, \sigma^2)$ (Example 2.1), then

$$\frac{\partial \log f(y;\theta)}{\partial \theta} = \begin{bmatrix} \dfrac{\partial \log f(y;\theta)}{\partial \mu} \\ \dfrac{\partial \log f(y;\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{y-\mu}{\sigma^2} \\ \frac{1}{2\sigma^2}\left(\frac{(y-\mu)^2}{\sigma^2} - 1\right) \end{bmatrix}.$$

Proposition 2.1 (Score expectation)**.** *The expectation of the score is zero.*

Proof. We have:

$$\begin{aligned} \mathbb{E}\left(\frac{\partial \log f(Y;\theta)}{\partial \theta}\right) &= \int \frac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta) dy \\ &= \int \frac{\partial f(y;\theta)/\partial \theta}{f(y;\theta)} f(y;\theta) dy = \frac{\partial}{\partial \theta} \int f(y;\theta) dy \\ &= \partial 1/\partial \theta = 0, \end{aligned}$$

which gives the result. □

Definition 2.4 (Fisher information matrix)**.** The information matrix is (minus) the the expectation of the second derivatives of the log-likelihood function:

$$\mathcal{I}_Y(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log f(Y;\theta)}{\partial \theta \partial \theta'}\right).$$

Proposition 2.2. *We have*

$$\mathcal{I}_Y(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(Y;\theta)}{\partial \theta}\right)\left(\frac{\partial \log f(Y;\theta)}{\partial \theta}\right)'\right] = \mathbb{V}ar[S(Y;\theta)].$$

Proof. We have $\frac{\partial^2 \log f(Y;\theta)}{\partial \theta \partial \theta'} = \frac{\partial^2 f(Y;\theta)}{\partial \theta \partial \theta'} \frac{1}{f(Y;\theta)} - \frac{\partial \log f(Y;\theta)}{\partial \theta} \frac{\partial \log f(Y;\theta)}{\partial \theta'}$. The expectation of the first right-hand side term is $\partial^2 1/(\partial \theta \partial \theta') = \mathbf{0}$, which gives the result. □

Example 2.2. If $y_i \sim$ *i.i.d.* $\mathcal{N}(\mu, \sigma^2)$, let $\theta = [\mu, \sigma^2]'$ then

$$\frac{\partial \log f(y;\theta)}{\partial \theta} = \left[\frac{y-\mu}{\sigma^2} \quad \frac{1}{2\sigma^2}\left(\frac{(y-\mu)^2}{\sigma^2} - 1\right)\right]',$$

and

$$\mathcal{I}_Y(\theta) = \mathbb{E}\left(\frac{1}{\sigma^4}\begin{bmatrix} \sigma^2 & y-\mu \\ y-\mu & \frac{(y-\mu)^2}{\sigma^2} - \frac{1}{2} \end{bmatrix}\right) = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}.$$

Proposition 2.3 (Additive property of the Information matrix)**.** *The information matrix resulting from two independent experiments is the sum of the information matrices:*

$$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta).$$

Proof. Directly deduced from the definition of the information matrix (Def. 2.4), using that the epxectation of a product of independent variables is the product of the expectations. □

Theorem 2.1 (Frechet-Darmois-Cramer-Rao bound)**.** *Consider an unbiased estimator of θ denoted by $\widehat{\theta}(Y)$. The variance of the random variable $\omega'\widehat{\theta}$ (which is a linear combination of the components of $\widehat{\theta}$) is larger than:*

$$(\omega'\omega)^2/(\omega'\mathcal{I}_Y(\theta)\omega).$$

Proof. The Cauchy-Schwarz inequality implies that $\sqrt{\mathbb{V}ar(\omega'\widehat{\theta}(Y))\mathbb{V}ar(\omega' S(Y;\theta))} \geq |\omega'\mathbb{C}ov[\widehat{\theta}(Y), S(Y;\theta)]\omega|$. Now, $\mathbb{C}ov[\widehat{\theta}(Y), S(Y;\theta)] = \int_y \widehat{\theta}(y)\frac{\partial \log f(y;\theta)}{\partial\theta} f(y;\theta)dy = \frac{\partial}{\partial\theta}\int_y \widehat{\theta}(y)f(y;\theta)dy = \mathbf{I}$ because $\widehat{\theta}$ is unbiased. Therefore $\mathbb{V}ar(\omega'\widehat{\theta}(Y)) \geq \mathbb{V}ar(\omega' S(Y;\theta))^{-1}(\omega'\omega)^2$. Prop. 2.2 leads to the result. □

Definition 2.5 (Identifiability)**.** The vector of parameters θ is identifiable if, for any other vector θ^*:

$$\theta^* \neq \theta \Rightarrow \mathcal{L}(\theta^*; \mathbf{y}) \neq \mathcal{L}(\theta; \mathbf{y}).$$

Definition 2.6 (Maximum Likelihood Estimator (MLE))**.** The maximum likelihood estimator (MLE) is the vector θ that maximizes the likelihood function. Formally:

$$\theta_{MLE} = \arg\max_\theta \mathcal{L}(\theta; \mathbf{y}) = \arg\max_\theta \log \mathcal{L}(\theta; \mathbf{y}). \tag{2.8}$$

Definition 2.7 (Likelihood equation)**.** A necessary condition for maximizing the likelihood function (under regularity assumption, see Hypotheses 2.1) is:

$$\frac{\partial \log \mathcal{L}(\theta; \mathbf{y})}{\partial\theta} = \mathbf{0}. \tag{2.9}$$

Hypothesis 2.1 (Regularity assumptions)**.** We have:

i. $\theta \in \Theta$ where Θ is compact.
ii. θ_0 is identified.
iii. The log-likelihood function is continuous in θ.
iv. $\mathbb{E}_{\theta_0}(\log f(Y;\theta))$ exists.
v. The log-likelihood function is such that $(1/n)\log \mathcal{L}(\theta;\mathbf{y})$ converges almost surely to $\mathbb{E}_{\theta_0}(\log f(Y;\theta))$, uniformly in $\theta \in \Theta$.
vi. The log-likelihood function is twice continuously differentiable in an open neighborood of θ_0.
vii. The matrix $\mathbf{I}(\theta_0) = -\mathbb{E}_0\left(\frac{\partial^2 \log \mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'}\right)$ —the Fisher Information matrix— exists and is nonsingular.

Proposition 2.4 (Properties of MLE)**.** *Under regularity conditions (Assumptions 2.1), the MLE is:*

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

$$\boxed{\sqrt{n}(\theta_{MLE} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_Y(\theta_0)^{-1}).} \tag{2.10}$$

c. ***Asymptotically efficient****:* θ_{MLE} *is asymptotically efficient and achieves the Freechet-Darmois-Cramer-Rao lower bound for consistent estimators.*
d. ***Invariant****: The MLE of* $g(\theta_0)$ *is* $g(\theta_{MLE})$ *if* g *is a continuous and continuously differentiable function.*

Proof. See Appendix 4.5. □

Since $\mathcal{I}_Y(\theta_0) = \frac{1}{n}\mathbf{I}(\theta_0)$, the asymptotic covariance matrix of the MLE is $[\mathbf{I}(\theta_0)]^{-1}$, that is:

$$[\mathbf{I}(\theta_0)]^{-1} = \left[-\mathbb{E}_0\left(\frac{\partial^2 \log \mathcal{L}(\theta;\mathbf{y})}{\partial\theta\partial\theta'}\right)\right]^{-1}.$$

A direct (analytical) evaluation of this expectation is often out of reach. It can however be estimated by, either:

$$\hat{\mathbf{I}}_1^{-1} = \left(-\frac{\partial^2 \log \mathcal{L}(\theta_{MLE}; \mathbf{y})}{\partial\theta\partial\theta'} \right)^{-1}, \tag{2.11}$$

$$\hat{\mathbf{I}}_2^{-1} = \left(\sum_{i=1}^{n} \frac{\partial \log \mathcal{L}(\theta_{MLE}; y_i)}{\partial\theta} \frac{\partial \log \mathcal{L}(\theta_{MLE}; y_i)}{\partial\theta'} \right)^{-1}. \tag{2.12}$$

Asymptotically, we have $(\hat{\mathbf{I}}_1^{-1})\hat{\mathbf{I}}_2 = Id$, that is, the two formulas provide the same result.

In case of (suspected) misspecification, one can use the so-called *sandwich estimator* of the covariance matrix.[1] This covariance matrix is given by:

$$\hat{\mathbf{I}}_3^{-1} = \hat{\mathbf{I}}_2^{-1}\hat{\mathbf{I}}_1\hat{\mathbf{I}}_2^{-1}. \tag{2.13}$$

2.2.3 To sum up – MLE in practice

To implement MLE, we need:

- A parametric model (depending on the vector of parameters θ whose "true" value is θ_0) is specified.
- i.i.d. sources of randomness are identified.
- The density associated to one observation y_i is computed analytically (as a function of θ): $f(y; \theta)$.
- The log-likelihood is $\log \mathcal{L}(\theta; \mathbf{y}) = \sum_i \log f(y_i; \theta)$.
- The MLE estimator results from the optimization problem (this is Eq. (2.8)):

$$\theta_{MLE} = \arg\max_{\theta} \log \mathcal{L}(\theta; \mathbf{y}). \tag{2.14}$$

- We have: $\theta_{MLE} \sim \mathcal{N}(\theta_0, \mathbf{I}(\theta_0)^{-1})$, where $\mathbf{I}(\theta_0)^{-1}$ is estimated by means of Eq. (2.11), Eq. (2.12), or Eq. (2.13). Most of the time, this computation is numerical.

[1] For more details, see, e.g., Charles Geyer's lectures notes.

2.2.4 Example: MLE estimation of a mixture of Gaussian distribution

Consider the returns of the Swiss Market Index (SMI). Assume that these returns are independently drawn from a mixture of Gaussian distributions. The p.d.f. $f(x;\theta)$, with $\theta = [\mu_1, \sigma_1, \mu_2, \sigma_2, p]'$, is given by:

$$p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

(See p.d.f. of mixtures of Gaussian distributions.)

```
library(AEC);data(smi)
T <- dim(smi)[1]
h <- 5 # holding period (one week)
smi$r <- c(rep(NaN,h),
           100*c(log(smi$Close[(1+h):T]/smi$Close[1:(T-h)])))
indic.dates <- seq(1,T,by=5)  # weekly returns
smi <- smi[indic.dates,]
smi <- smi[complete.cases(smi),]
par(mfrow=c(1,1));par(plt=c(.15,.95,.1,.95))
plot(smi$Date,smi$r,type="l",xlab="",ylab="in percent")
abline(h=0,col="blue")
abline(h=mean(smi$r,na.rm = TRUE)+2*sd(smi$r,na.rm = TRUE),lty=3,col="blue")
abline(h=mean(smi$r,na.rm = TRUE)-2*sd(smi$r,na.rm = TRUE),lty=3,col="blue")
```

Build the log-likelihood function (fucntion `log.f`), and use the numerical BFGS algorithm to maximize it (using the `optim` wrapper):

```
f <- function(theta,y){ # Likelihood function
  mu.1 <- theta[1]; mu.2 <- theta[2]
  sigma.1 <- theta[3]; sigma.2 <- theta[4]
  p <- exp(theta[5])/(1+exp(theta[5]))
  res <- p*1/sqrt(2*pi*sigma.1^2)*exp(-(y-mu.1)^2/(2*sigma.1^2)) +
    (1-p)*1/sqrt(2*pi*sigma.2^2)*exp(-(y-mu.2)^2/(2*sigma.2^2))
  return(res)
}
```

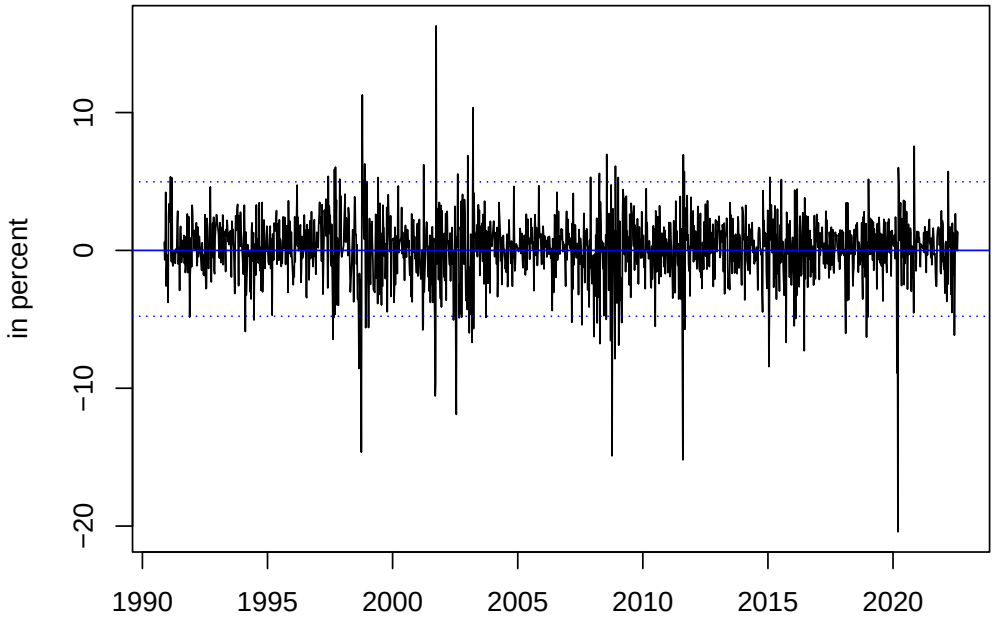


Figure 2.4: Time series of SMI weekly returns (source: Yahoo Finance).

```
log.f <- function(theta,y){ #log-Likelihood function
  return(-sum(log(f(theta,y))))
}
res.optim <- optim(c(0,0,0.5,1.5,.5),
                   log.f,
                   y=smi$r,
                   method="BFGS", # could be "Nelder-Mead"
                   control=list(trace=FALSE,maxit=100),hessian=TRUE)
theta <- res.optim$par
theta
```

```
## [1]  0.3012379 -1.3167476  1.7715072  4.8197596  1.9454889
```

Next, compute estimates of the covariance matrix of the MLE (using Eqs. (2.11), (2.12), and (2.13)), and compare the three sets of resulting standard deviations for the five estimated paramters:

```
# Hessian approach:
I.1 <- solve(res.optim$hessian)
# Outer-product of gradient approach:
log.f.0 <- log(f(theta,smi$r))
epsilon <- .00000001
d.log.f <- NULL
for(i in 1:length(theta)){
  theta.i <- theta
  theta.i[i] <- theta.i[i] + epsilon
  log.f.i <- log(f(theta.i,smi$r))
  d.log.f <- cbind(d.log.f,
                   (log.f.i - log.f.0)/epsilon)
}
I.2 <- solve(t(d.log.f) %*% d.log.f)
# Misspecification-robust approach (sandwich formula):
I.3 <- I.1 %*% solve(I.2) %*% I.1
cbind(diag(I.1),diag(I.2),diag(I.3))
```

```
##             [,1]        [,2]       [,3]
## [1,] 0.003683422 0.003199481 0.00586160
## [2,] 0.226892824 0.194283391 0.38653389
## [3,] 0.005764271 0.002769579 0.01712255
## [4,] 0.194081311 0.047466419 0.83130838
## [5,] 0.092114437 0.040366005 0.31347858
```

According to the first (respectively third) type of estimate for the covariance matrix, a 95% confidence interval for μ_1 is [0.182, 0.42] (resp. [0.151, 0.451]).

Note that we have not directly estimated parameter p but $\nu = \log(p/(1-p))$ (in such a way that $p = \exp(\nu)/(1+\exp(\nu))$). In order to get an estimate of the standard deviation of our esitmate of p, we can implement the **Delta method**. This method is based on the fact that, for a function g that is continuous in the neighborhood of θ_0 and for large n, we have:

$$\mathbb{V}ar(g(\hat{\theta}_n)) \approx \frac{\partial g(\hat{\theta}_n)}{\partial \theta'} \mathbb{V}ar(\hat{\theta}_n) \frac{\partial g(\hat{\theta}_n)'}{\partial \theta}. \tag{2.15}$$

```
g <- function(theta){
  mu.1 <- theta[1]; mu.2 <- theta[2]
  sigma.1 <- theta[3]; sigma.2 <- theta[4]
  p <- exp(theta[5])/(1+exp(theta[5]))
  return(c(mu.1,mu.2,sigma.1,sigma.2,p))
}
# Computation of g's gradient around estimated theta:
eps <- .00001
g.theta <- g(theta)
g.gradient <- NULL
for(i in 1:5){
  theta.perturb <- theta
  theta.perturb[i] <- theta[i] + eps
  g.gradient <- cbind(g.gradient,(g(theta.perturb)-g.theta)/eps)
}
Var <- g.gradient %*% I.3 %*% t(g.gradient)
stdv.g.theta <- sqrt(diag(Var))
stdv.theta <- sqrt(diag(I.3))
cbind(theta,stdv.theta,g.theta,stdv.g.theta)
```

```
##           theta stdv.theta    g.theta stdv.g.theta
## [1,]  0.3012379 0.07656108  0.3012379   0.07656108
## [2,] -1.3167476 0.62171850 -1.3167476   0.62171850
## [3,]  1.7715072 0.13085316  1.7715072   0.13085316
## [4,]  4.8197596 0.91176114  4.8197596   0.91176114
## [5,]  1.9454889 0.55989158  0.8749539   0.06125726
```

The previous results show that the MLE estimate of p is 0.8749539, and its standard deviation is approximately equal to 0.0612573.

To finish with, let us draw the estimated parametric p.d.f. (the mixture of Gaussian distribution), and compare it to a non-parametric (kernel-based) estimate of this p.d.f. (using function `density`):

```
x <- seq(-5,5,by=.01)
par(plt=c(.1,.95,.1,.95))
plot(x,f(theta,x),type="l",lwd=2,xlab="returns, in percent",ylab="",
```

```
     ylim=c(0,1.4*max(f(theta,x))))
lines(density(smi$r),type="l",lwd=2,lty=3)
lines(x,dnorm(x,mean=mean(smi$r),sd = sd(smi$r)),col="red",lty=2,lwd=2)
rug(smi$r,col="blue")
legend("topleft",
       c("Kernel estimate (non-parametric)",
         "Estimated mixture of Gaussian distr. (MLE, parametric)",
         "Normal distribution"),
       lty=c(3,1,2),lwd=c(2), # line width
       col=c("black","black","red"),pt.bg=c(1),pt.cex = c(1),
       bg="white",seg.len = 4)
```

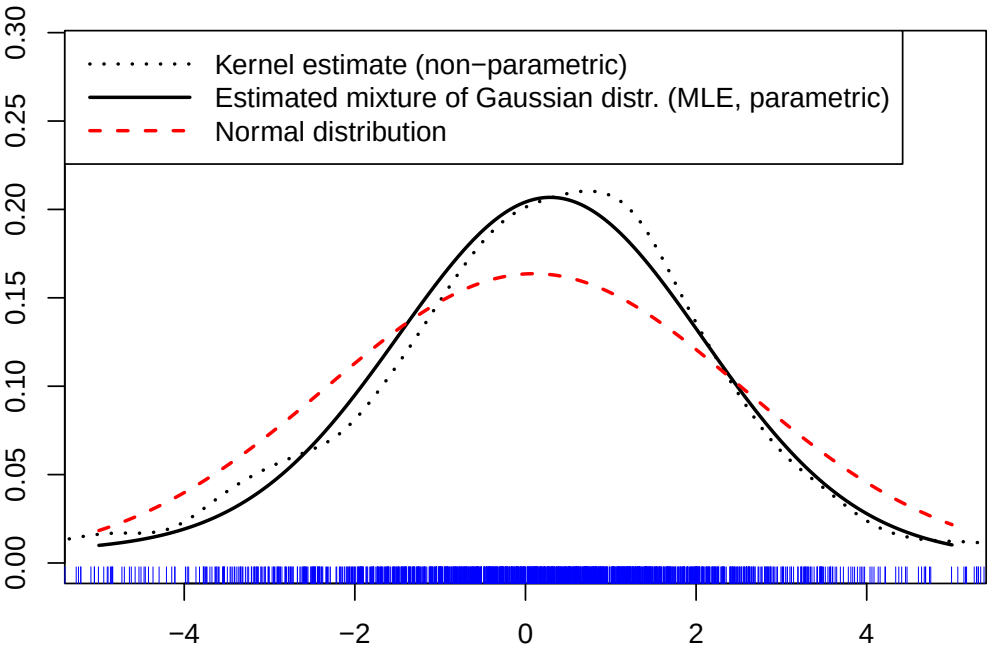


Figure 2.5: Comparison of different estimates of the distribution of returns.

2.2.5 Test procedures

Suppose we want to test the following parameter restrictions:

$$\boxed{H_0 : \underbrace{h(\theta)}_{r\times 1} = 0.} \tag{2.16}$$

In the context of MLE, three tests are largely used:

- Likelihood Ratio (LR) test,
- Wald (W) test,
- Lagrange Multiplier (LM) test.

Here is the rationale behind these three tests:[2]

- LR: If $h(\theta) = 0$, then imposing this restriction during the estimation (restricted estimator) should not result in a large decrease in the likelihood function (w.r.t the unrestricted estimation).
- Wald: If $h(\theta) = 0$, then $h(\widehat{\theta})$ should not be far from 0 (even if these restrictions are not imposed during the MLE).
- LM: If $h(\theta) = 0$, then the gradient of the likelihood function should be small when evaluated at the restricted estimator.

In terms of implementation, while the LR necessitates to estimate both restricted and unrestricted models, the Wald test requires the estimation of the unrestricted model only, and the LM tests requires the estimation of the restricted model only.

As shown below, the three test statistics associated with these three tests coincide asymptotically. (Therefore, they naturally have the same asymptotic distribution, that are χ^2.)

Proposition 2.5 (Asymptotic distribution of the Wald statistic)**.** *Under regularity conditions (Assumptions 2.1) and under $H_0 : h(\theta) = 0$, the Wald statistic, defined by:*

$$\boxed{\xi^W = h(\widehat{\theta})' \mathbb{V}ar[h(\widehat{\theta})]^{-1} h(\widehat{\theta}),}$$

[2] An interesting graphical presentation of the tests is proposed in Buse (1982).

where

$$\mathbb{V}ar[h(\widehat{\theta})] = \left(\frac{\partial h(\widehat{\theta})}{\partial \theta'}\right) \mathbb{V}ar[\widehat{\theta}] \left(\frac{\partial h(\widehat{\theta})'}{\partial \theta}\right), \tag{2.17}$$

is asymptotically $\chi^2(r)$*, where the number of degrees of freedom* r *corresponds to the dimension of* $h(\theta)$*. (Note that Eq.* (2.17) *is the same as the one used in the Delta method, see Eq.* (2.15)*.)*

The Wald test, defined by the critical region

$$\{\xi^W \geq \chi^2_{1-\alpha}(r)\},$$

where $\chi^2_{1-\alpha}(r)$ *denotes the quantile of level* $1-\alpha$ *of the* $\chi^2(r)$ *distribution, has asymptotic level* α *and is consistent.*[3]

Proof. See Appendix 4.5. □

In practice, in Eq. (2.17), $\mathbb{V}ar[\widehat{\theta}]$ is replaced by an estimate given, e.g., by Eq. (2.11), Eq. (2.12), or Eq. (2.13).

Proposition 2.6 (Asymptotic distribution of the LM test statistic)**.** *Under regularity conditions (Assumptions 2.1) and under* $H_0 : h(\theta) = 0$*, the LM statistic*

$$\boxed{\xi^{LM} = \left(\left.\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta'}\right|_{\theta=\widehat{\theta}^0}\right) [\mathbf{I}(\widehat{\theta}^0)]^{-1} \left(\left.\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}^0}\right),} \tag{2.18}$$

(where $\widehat{\theta}^0$ *is the restricted MLE estimator) is* $\chi^2(r)$*.*

The test defined by the critical region:

$$\{\xi^{LM} \geq \chi^2_{1-\alpha}(r)\}$$

has asymptotic level α *and is consistent (see Defs. 4.7 and 4.8). This test is called* Score *or* Lagrange Multiplier (LM) *test.*

Proof. See Appendix 4.5. □

[3] See Defs. 4.7 and 4.8 for definitions of the asymptotic levels and consistency of tests.

Definition 2.8 (Likelihood Ratio test statistics)**.** The likelihood ratio associated to a restriction of the form $H_0 : h(\theta) = 0$ is given by:

$$LR = \frac{\mathcal{L}_R(\theta; \mathbf{y})}{\mathcal{L}_U(\theta; \mathbf{y})} \quad (\in [0, 1]),$$

where $\mathcal{L}_R$ (respectively $\mathcal{L}_U$) is the likelihood function that imposes (resp. that does not impose) the restriction. The likelihood ratio test statistic is given by $-2\log(LR)$, that is:

$$\boxed{\xi^{LR} = 2(\log \mathcal{L}_U(\theta; \mathbf{y}) - \log \mathcal{L}_R(\theta; \mathbf{y})).}$$

Proposition 2.7 (Asymptotic equivalence of LR, LM, and Wald tests)**.** *Under the null hypothesis H_0, we have, asymptotically:*

$$\xi^{LM} = \xi^{LR} = \xi^{W}.$$

Proof. See Appendix 4.5. □

2.3 Bayesian approach

2.3.1 Introduction

An excellent introduction to Bayesian methods is proposed by Martin Haugh, 2017.

As suggested by the name of this approach, the starting point is the Bayes formula:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\&B)}{\mathbb{P}(B)},$$

where A and B are two "events". For instance, A may be: parameter α (conceived as something stochastic) lies in interval $[a, b]$. Assume that you are interested in the probability of occurrence of A. Without any specific information (or "unconditionally"), this probability if $\mathbb{P}(A)$. Your evaluation of this probability can only be better if you are provided with any additional form of information. Typically, if the event B tends to occur simultaneously with A, then knowledge of B can be useful. The Bayes formula says how

this additional information (on B) can be used to "update" the probability of event A.

In our case, this intuition will work as follows: assume that you know the form of the data-generating process (DGP). That is, you know the structure of the model used to draw some stochastic data; you also know the type of distributions used to generate these data. However, you do not know the numerical values of all the parameters characterizing the DGP. Let us denote by θ the vector of unknown parameters. While these parameters are not known exactly, assume that we have –even without having observed any data– some **priors** on their distribution. Then, as was the case in the example above (with A and B), the observation of data generated by the model can only reduce the uncertainty associated with θ. Loosely speaking, combining the priors and the observations of data generated by the model should result in "thinner" distributions for the components of θ. The latter distributions are called the **posterior distributions**.[4]

Let us formalize this intuition. Define the prior by $f_\theta(\theta)$ and the model realizations (the "data") by vector $\mathbf{y}$. The joint distribution of $(\mathbf{y}, \theta)$ is given by:

$$f_{Y,\theta}(\mathbf{y}, \theta) = f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta),$$

and, symmetrically, by

$$f_{Y,\theta}(\mathbf{y}, \theta) = f_{\theta|Y}(\theta, \mathbf{y}) f_Y(\mathbf{y}),$$

where $f_{\theta|Y}(\cdot, \mathbf{y})$, the distribution of the parameters conditional on the observations, is the **posterior** distribution.

The last two equations imply that:

$$f_{\theta|Y}(\theta, \mathbf{y}) = \frac{f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta)}{f_Y(\mathbf{y})}. \tag{2.19}$$

Note that f_Y is the marginal (or unconditional) distribution of $\mathbf{y}$, that can be written:

$$f_Y(\mathbf{y}) = \int f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta) d\theta. \tag{2.20}$$

[4]The output of the Bayesian approach will be the (posterior) distribution of the vector of parameters (θ). When we speak about the *distributions of the components of* θ, we mean the marginal distributions of each component of the vector.

Eq. (2.19) is sometimes rewritten as follows:

$$f_{\theta|Y}(\theta, \mathbf{y}) \propto f_{\theta,Y}(\theta, \mathbf{y}) := f_{Y|\theta}(\mathbf{y}, \theta) f_{\theta}(\theta), \tag{2.21}$$

where $\propto$ means, loosely speaking, "*proportional to*". In rare instances, starting from given priors, one can analytically compute the posterior distribution $f_{\theta}(\theta, \mathbf{y})$. However, in most cases, this is out of reach. One then has to resort to numerical approaches to compute the posterior distribution. Monte Carlo Markov Chains (MCMC) is one of them.

According to the Bernstein-von Mises Theorem, Bayesian and MLE estimators have the same large sample properties. (In particular, the Bayesian approach also achieve the FDCR bound, see Theorem 2.1.) The intuition behind this result is that the influence of the prior diminishes with increasing sample sizes.

2.3.2 Monte-Carlo Markov Chains

MCMC techniques aim at using simulations to approach a distribution whose distribution is difficult to obtain analytically. Indeed, in some circumstances, one can draw in a distribution even if we do not know its analytical expression.

Definition 2.9 (Markov Chain)**.** The sequence $\{z_i\}$ is said to be a (first-order) Markovian process is it satisfies:

$$f(z_i|z_{i-1}, z_{i-2}, \dots) = f(z_i|z_{i-1}).$$

The Metropolis-Hastings (MH) algorithm is a specific MCMC approach that allows to generate samples of θ's whose distribution approximately corresponds to the posterior distribution of Eq. (2.19).

The MH algorithm is a recursive algorithm. That is, one can draw the i^{th} value of θ, denoted by θ_i, if one has already drawn θ_{i-1}. Assume we have θ_{i-1}. We obtain a value for θ_i by implementing the following steps:

1. Draw $\tilde{\theta}_i$ from the conditional distribution $Q_{\tilde{\theta}|\theta}(\cdot, \theta_{i-1})$, called **proposal distribution**.
2. Draw u in a uniform distribution on $[0, 1]$.

3. Compute

$$\alpha(\tilde{\theta}_i, \theta_{i-1}) := \min\left(\frac{f_{\theta,Y}(\tilde{\theta}_i, \mathbf{y})}{f_{\theta,Y}(\theta_{i-1}, \mathbf{y})} \times \frac{Q_{\tilde{\theta}|\theta}(\theta_{i-1}, \tilde{\theta}_i)}{Q_{\tilde{\theta}|\theta}(\tilde{\theta}_i, \theta_{i-1})}, 1\right), \tag{2.22}$$

   where $f_{\theta,Y}$ is given in Eq. (2.21).
4. If $u < \alpha(\tilde{\theta}_i, \theta_{i-1})$, then take $\theta_i = \tilde{\theta}_i$, otherwise we leave θ_i equal to θ_{i-1}.

It can be shown that, the distribution of the draws converges to the posterior distribution. That is, after a sufficiently large number of iterations, the draws can be considered to be drawn from the posterior distribution.[5]

To get some insights into the algorithm, consider the case of a **symmetric proposal distribution**, that is:

$$Q_{\tilde{\theta}|\theta}(\tilde{\theta}_i, \theta_{i-1}) = Q_{\tilde{\theta}|\theta}(\theta_{i-1}, \tilde{\theta}_i). \tag{2.23}$$

We then have:

$$\alpha(\tilde{\theta}, \theta_{i-1}) = \min\left(\frac{q(\tilde{\theta}, y)}{q(\theta_{i-1}, y)}, 1\right). \tag{2.24}$$

Remember that, up to the marginal distribution of the data ($f_Y(\mathbf{y})$), $f_{\theta,Y}(\tilde{\theta}, \mathbf{y})$ is the probability of observing $\mathbf{y}$ conditional on having a model parameterized by $\tilde{\theta}$. Then, under Eq. (2.24), it appears that if this probability is larger for $\tilde{\theta}$ than for θ_{i-1} (in which case $\tilde{\theta}$ seems "more consistent with the observations $\mathbf{y}$" than θ_{i-1}), we accept θ_i. By contrast, if $f_{\theta,Y}(\tilde{\theta}, \mathbf{y}) < f_{\theta,Y}(\theta_{i-1}, \mathbf{y})$, then we do not necessarily accept the proposed value $\tilde{\theta}$, especially if $f_{\theta,Y}(\tilde{\theta}, \mathbf{y}) \ll f_{\theta,Y}(\theta_{i-1}, \mathbf{y})$ (in which case $\tilde{\theta}$ seems far less consistent with the observations $\mathbf{y}$ than θ_{i-1}, and, accordingly, the acceptance probability, namely $\alpha(\tilde{\theta}, \theta_{i-1})$, is small).

The choice of the **proposal distribution** $Q_{\tilde{\theta}|\theta}$ is crucial to get a rapid convergence of the algorithm. Looking at Eq. (2.22), it is easily seen that the optimal choice would be $Q_{\tilde{\theta}|\theta}(\cdot, \theta_i) = f_{\theta|Y}(\cdot, \mathbf{y})$. In that case, we would have

[5] The proof of this claim is based on the fact that, if θ_{i-1} is drawn from the posterior distribution, then it is also the case for θ_i.

$\alpha(\tilde{\theta}_i, \theta_{i-1}) \equiv 1$ (see Eq. (2.22)). We would then accept all draws from the proposal distribution, as this distribution would directly be the posterior distribution. Of course, this situation is not realistoc as the objective of the algorithm is precisely to approximate the posterior distribution.

A common choice for Q is a multivariate normal distribution. If θ is of dimension K, we can for instance use:

$$Q(\tilde{\theta}, \theta) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^K} \exp\left(-\frac{1}{2}\sum_{j=1}^{K}\frac{(\tilde{\theta}_j - \theta_j)^2}{\sigma^2}\right),$$

which is an example of symmetric proposal distribution (see Eq. (2.23)). Equivalently, we then have:

$$\tilde{\theta} = \theta + \varepsilon,$$

where ε is a K-dimensional vector of independent zero-mean normal disturbances of variance σ^2.[6] One then has to determine an appropriate value for σ. If it is too low, then α will be close to 1 (as $\tilde{\theta}_i$ will be close to θ_{i-1}), and we will accept very often the proposed value ($\tilde{\theta}_i$). This seems to be a favourable situation. But it may not be. Indeed, it means that it will take a large number of iterations to explore the whole distribution of θ. What if σ is very large? In this case, it is likely that the porposed values ($\tilde{\theta}_i$) will often result in poor likelihoods; The probability of acceptance will then be low and the Markov chain may be blocked at its initial value. Therefore, intermediate values of σ^2 have to be determined. The acceptance rate (i.e., the average value of $\alpha(\tilde{\theta}, \theta_{i-1})$) can be used as a guide for that. Indeed, a literature explores the optimal values for such acceptance rate (in order to obtain the best possible fit of the posterior for a minimum number of algorithm iterations). In particular, following Roberts et al. (1997), people often target acceptance rate of the order of magnitude of 20%.

It is important to note that, to implement this approach, one only has to be able to compute the joint p.d.f. $q(\theta, \mathbf{y}) = f_{Y|\theta}(\mathbf{y}, \theta) f_\theta(\theta)$ (Eq. (2.21)). That is, as soon as one can evaluate the likelihood ($f_{Y|\theta}(\mathbf{y}, \theta)$) and the prior ($f_\theta(\theta)$), we can employ this methodology.

[6] We could also have different variances for the different components of θ. However, this may lead to complicated settings. A useful practice consists in looking for model (re)parametrization –based, e.g., on exponential and/or logistic functions– that are such that the components of θ are all expected to be of the order of magnitude of the unity.

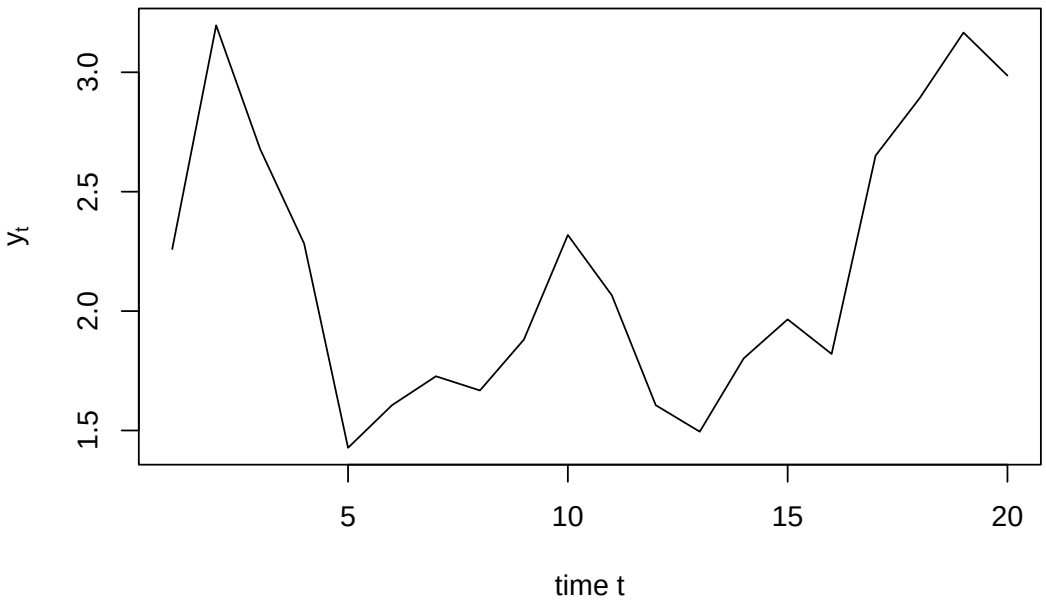
2.3.3 Example: AR(1) specification

In the following example, we employ MCMC in order to estimate the posterior distributions of the three parameters defining an AR(1) model (see Section **??**). The specification is as follows:

$$y_t = \mu + \rho y_{t-1} + \sigma\varepsilon_t, \quad \varepsilon_t \sim i.i.d.\,\mathcal{N}(0,1).$$

Hence, we have $\theta = [\mu, \rho, \sigma]$. Let us first simulate the process on T periods:

```
mu <- .6; rho <- .8; sigma <- .5 # true model specification
T <- 20 # number of observations
y0 <- mu/(1-rho)
Y <- NULL
for(t in 1:T){
  if(t==1){y <- y0}
  y <- mu + rho*y + sigma * rnorm(1)
  Y <- c(Y,y)}
plot(Y,type="l",xlab="time t",ylab=expression(y[t]))
```



Next, let us write the likelihood function, i.e. $f_{Y|\theta}(\mathbf{y}, \theta)$. For ρ, which is expected to be between 0 and 1, we use a logistic transformation. For σ, that is expected to be positive, we use an exponential transformation.

```
likelihood <- function(param,Y){
  mu  <- param[1]
  rho <- exp(param[2])/(1+exp(param[2]))
  sigma <- exp(param[3])
  MU <- mu/(1-rho)
  SIGMA2 <- sigma^2/(1-rho^2)
  L <- 1/sqrt(2*pi*SIGMA2)*exp(-(Y[1]-MU)^2/(2*SIGMA2))
  Y1 <- Y[2:length(Y)]
  Y0 <- Y[1:(length(Y)-1)]
  aux <- 1/sqrt(2*pi*sigma^2)*exp(-(Y1-mu-rho*Y0)^2/(2*sigma^2))
  L <- L * prod(aux)
  return(L)
}
```

Next define function `rQ` that draws from the (Gaussian) proposal distribution, as well as function `Q`, that computes $Q_{\tilde{\theta}|\theta}(\tilde{\theta}, \theta)$:

```
rQ <- function(x,a){
  n <- length(x)
  y <- x + a * rnorm(n)
  return(y)}
Q <- function(y,x,a){
  q <- 1/sqrt(2*pi*a^2)*exp(-(y - x)^2/(2*a^2))
  return(prod(q))}
```

We consider Gaussian priors:

```
prior <- function(param,means_prior,stdv_prior){
  f <- 1/sqrt(2*pi*stdv_prior^2)*exp(-(param -
                                       means_prior)^2/(2*stdv_prior^2))
  return(prod(f))}
```

Function `p_tilde` corresponds to $f_{\theta,Y}$:

```
p_tilde <- function(param,Y,means_prior,stdv_prior){
  p <- likelihood(param,Y) * prior(param,means_prior,stdv_prior)
  return(p)}
```

We can now define function α (Eq. (2.22)):

```
alpha <- function(y,x,means_prior,stdv_prior,a){
  aux <- p_tilde(y,Y,means_prior,stdv_prior)/
    p_tilde(x,Y,means_prior,stdv_prior) * Q(y,x,a)/Q(x,y,a)
  alpha_proba <- min(aux,1)
  return(alpha_proba)}
```

Now, all is set for us to write the MCMC function:

```
MCMC <- function(Y,means_prior,stdv_prior,a,N){
  x <- means_prior
  all_theta <- NULL
  count_accept <- 0
  for(i in 1:N){
    y <- rQ(x,a)
    alph <- alpha(y,x,means_prior,stdv_prior,a)
    #print(alph)
    u <- runif(1)
    if(u < alph){
      count_accept <- count_accept + 1
      x <- y}
    all_theta <- rbind(all_theta,x)}
  print(paste("Acceptance rate:",toString(round(count_accept/N,3))))
  return(all_theta)}
```

Specify the Gaussian priors:

```
true_values <- c(mu,log(rho/(1-rho)),log(sigma))
means_prior <- c(1,0,0) # as if we did not know the true values
stdv_prior <- rep(2,3)
resultMCMC <- MCMC(Y,means_prior,stdv_prior,a=.45,N=20000)
```

```
## [1] "Acceptance rate: 0.098"
```

```
par(mfrow=c(2,3))
for(i in 1:length(means_prior)){
  m <- means_prior[i]
  s <- stdv_prior[i]
  x <- seq(m-3*s,m+3*s,length.out = 100)
  par(mfg=c(1,i))
  aux <- density(resultMCMC[,i])
  par(plt=c(.15,.95,.15,.85))
  plot(x,dnorm(x,m,s),type="l",xlab="",ylab="",main=paste("Parameter",i),
       ylim=c(0,max(aux$y)))
  lines(aux$x,aux$y,col="red",lwd=2)
  abline(v=true_values[i],lty=2,col="blue")
  par(mfg=c(2,i))
  plot(resultMCMC[,i],1:length(resultMCMC[,i]),xlim=c(min(x),max(x)),
       type="l",xlab="",ylab="")}
```

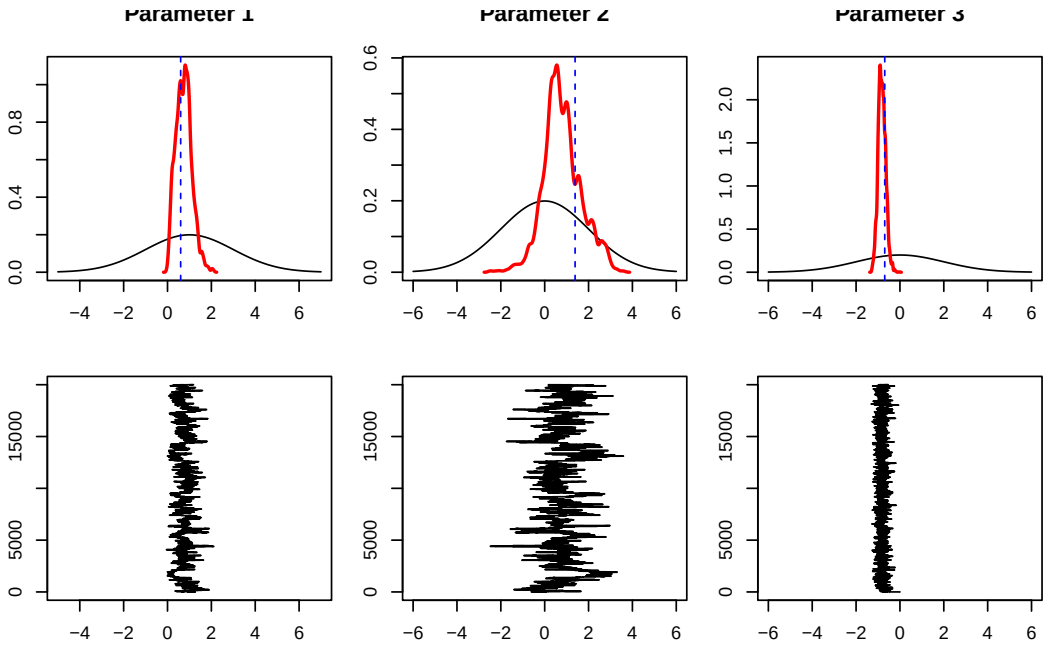


Figure 2.6: The upper line of plot compares prior (black) and posterior (red) distributions. The vertical dashed blue lines indicate the true values of the parameters. The second row of plots show the sequence of θ_i's generated by the MCMC algorithm. These sequences are the ones used to produce the posterior distributions (red lines) in the upper plots.

Chapter 3

Microeconometrics

In microeconometric models, the variables of interest often feature restricted distributions —for instance with discontinuous support—, which necessitates specific models. Typical examples are discrete-choice models (binary, multinomial, ordered outcomes), sample selection models (censored or truncated outcomes), and count-data models (integer outcomes). This chapter describes the estimation and interpretation of these models. It also shows how the discrete-choice models can emerge from (structural) random-utility frameworks.

3.1 Binary-choice models

In many instances, the variables to be explained (the y_i's) have only two possible values (0 and 1, say). That is, they are binary variables. The probability for them to be equal to either 0 or 1 may depend on independent variables, gathered in vectors $\mathbf{x}_i$ $(K \times 1)$.

The spectrum of applications is wide:

- Binary decisions (e.g. in referendums, being owner or renter, living in the city or in the countryside, in/out of the labour force,...),
- Contamination (disease or default),
- Success/failure (exams).

Without loss of generality, the model reads:

$$y_i|\mathbf{X} \sim \mathcal{B}(g(\mathbf{x}_i;\theta)), \tag{3.1}$$

where $g(\mathbf{x}_i;\theta)$ is the parameter of the Bernoulli distribution. In other words, conditionally on $\mathbf{X}$:

$$y_i = \begin{cases} 1 & \text{with probability } g(\mathbf{x}_i;\theta) \\ 0 & \text{with probability } 1 - g(\mathbf{x}_i;\theta), \end{cases} \tag{3.2}$$

where θ is a vector of parameters to be estimated.

An estimation strategy is to assume that $g(\mathbf{x}_i;\theta)$ can be proxied by $\tilde{\theta}'\mathbf{x}_i$ and to run a linear regression to estimate $\tilde{\theta}$ (a situation called **Linear Probability Model, LPM**):

$$y_i = \tilde{\theta}'\mathbf{x}_i + \varepsilon_i.$$

Notwithstanding the fact that this specification does not exclude negative probabilities or probabilities greater than one, it could be compatible with the *assumption of zero conditional mean* (Hypothesis **??**) and with the *assumption of non-correlated residuals* (Hypothesis **??**), but more difficultly with the *homoskedasticity assumption* (Hypothesis **??**). Moreover, the ε_i's cannot be Gaussian (because $y_i \in \{0,1\}$). Hence, using a linear regression to study the relationship between $\mathbf{x}_i$ and y_i can be consistent but it is inefficient.

Figure 3.1 illustrates the fit resulting from an application of the LPM model to binary (dependent) variables.

Except for its last row (LPM case), Table 3.1 provides examples of functions g valued in $[0,1]$, and that can therefore used in models of the type: $\mathbb{P}(y_i = 1|\mathbf{x}_i;\theta) = g(\theta'\mathbf{x}_i)$ (see Eq. (3.2)). The "linear" case is given for comparison, but note that it does not satisfy $g(\theta'\mathbf{x}_i) \in [0,1]$ for any value of $\theta'\mathbf{x}_i$.

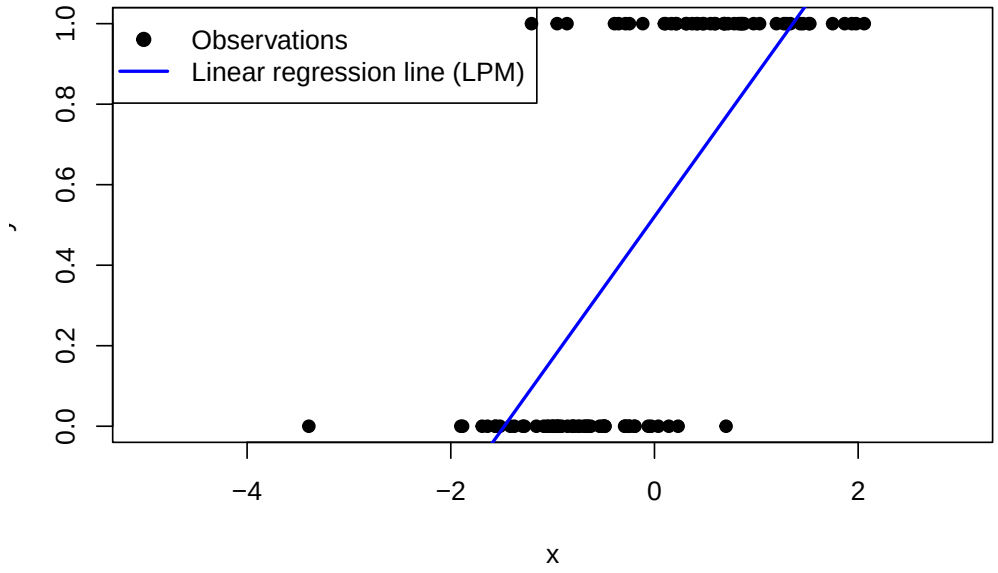
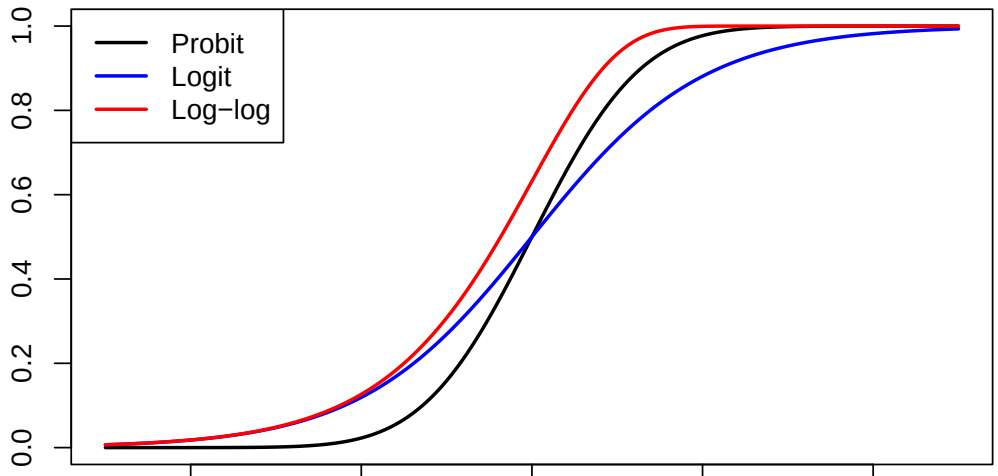


Figure 3.1: Fitting a binary variable with a linear model (Linear Probability Model, LPM). The model is $\mathbb{P}(y_i = 1|x_i) = \Phi(0.5 + 2x_i)$, where Φ is the c.d.f. of the normal distribution and where $x_i \sim i.i.d.\, \mathcal{N}(0, 1)$.

Table 3.1: This table provides examples of function g, s.t. $\mathbb{P}(y_i = 1|\mathbf{x}_i; heta) = g(\theta'\mathbf{x}_i)$. The LPM case (last row) is given for comparison but, again, it does not satisfy $g(\theta'\mathbf{x}_i) \in [0, 1]$ for any value of $\theta'\mathbf{x}_i$.

Model	Function g	Derivative
Probit	Φ	ϕ
Logit	$\dfrac{\exp(x)}{1+\exp(x)}$	$\dfrac{\exp(x)}{(1+\exp(x))^2}$
log-log	$1-\exp(-\exp(x))$	$\exp(-\exp(x))\exp(x)$
linear (LPM)	x	1

Figure 3.2 displays the first three g functions appearing in Table 3.1.



where Φ is the c.d.f. of the normal distribution. And for the logit model:

$$g(z) = \frac{1}{1+\exp(-z)}. \tag{3.4}$$

Figure 3.3 shows the conditional probabilities associated with the (probit) model that had been used to generate the data of Figure 3.1.

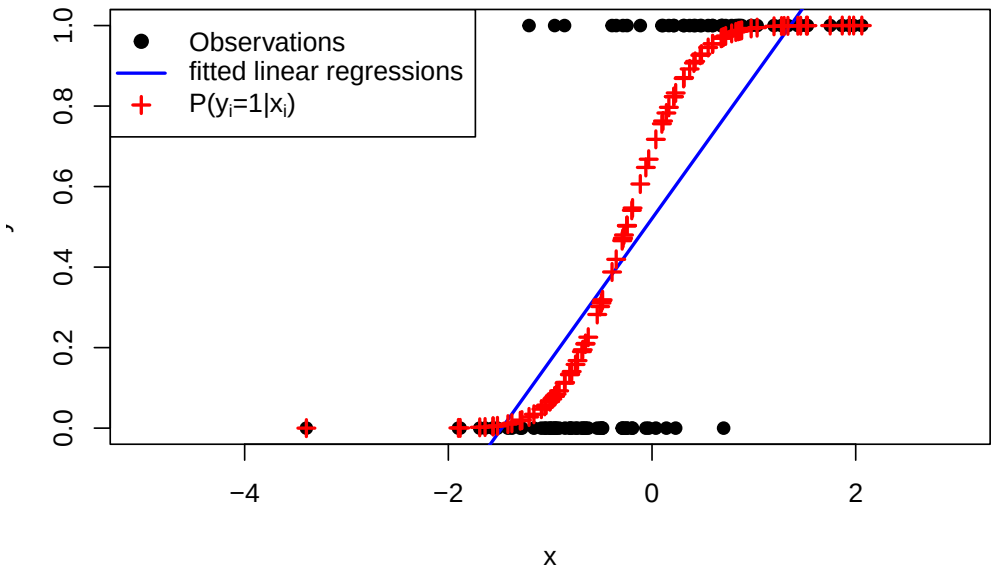


Figure 3.3: The model is $\mathbb{P}(y_i = 1|x_i) = \Phi(0.5 + 2x_i)$, where Φ is the c.d.f. of the normal distribution and where $x_i \sim i.i.d.\,\mathcal{N}(0,1)$. Crosses give the model-implied probabilities of having $y_i = 1$ (conditional on x_i).

3.1.1 Interpretation in terms of latent variable, and utility-based models

The probit model has an interpretation in terms of latent variables, which, in turn, is often exploited in structural models, called **Random Utility Models (RUM)**. In such structural models, it is assumed that the agents that have to take the decision do so by selecting the outcome that provides them with the larger utility (for agent i, two possible outcomes: $y_i = 0$ or $y_i = 1$). Part of this utility is observed by the econometrician —it depends on the covariates $\mathbf{x}_i$— and part of it is latent.

In the probit model, we have:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i;\theta) = \Phi(\theta'\mathbf{x}_i) = \mathbb{P}(-\varepsilon_i < \theta'\mathbf{x}_i),$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$. That is:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i;\theta) = \mathbb{P}(0 < y_i^*),$$

where $y_i^* = \theta'\mathbf{x}_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0,1)$. Variable y_i^* can be interpreted as a (latent) variable that determines y_i (since $y_i = \mathbb{I}_{\{y_i^*>0\}}$).

Figure 3.4 illustrates this situation.

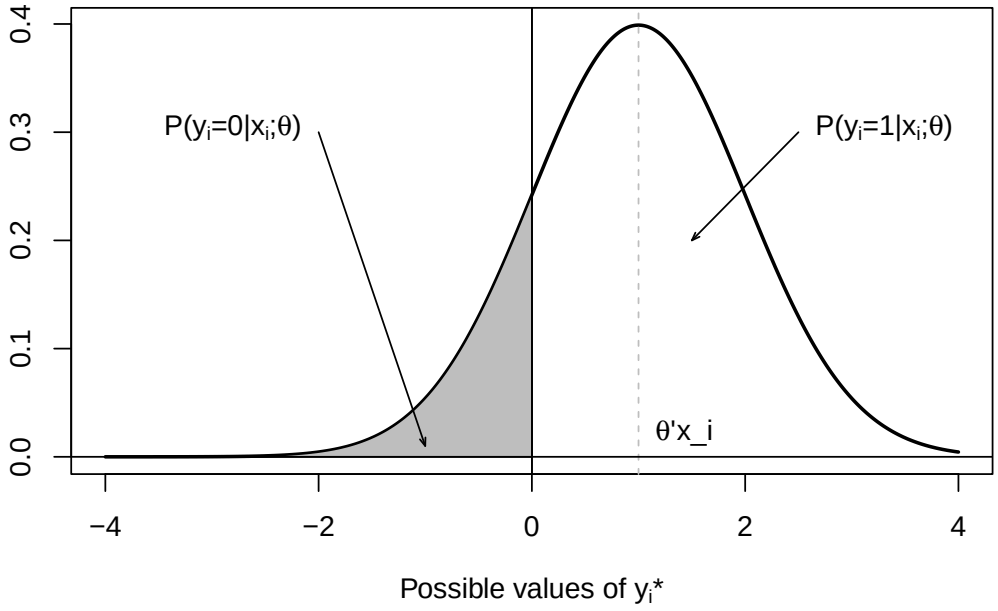


Figure 3.4: Distribution of y_i^* conditional on $\mathbf{x}_i$.

Assume that agent (i) chooses $y_i = 1$ if the utility associated with this choice $(U_{i,1})$ is higher than the one associated with $y_i = 0$ (that is $U_{i,0}$). Assume further that the utility of agent i, if she chooses outcome j ($\in \{0,1\}$), is given by

$$U_{i,j} = V_{i,j} + \varepsilon_{i,j},$$

where $V_{i,j}$ is the deterministic component of the utility associated with choice and where $\varepsilon_{i,j}$ is a random (agent-specific) component. Moreover, posit

$V_{i,j} = \theta_j' \mathbf{x}_i$. We then have:

$$\begin{aligned} \mathbb{P}(y_i = 1|\mathbf{x}_i;\theta) &= \mathbb{P}(\theta_1'\mathbf{x}_i + \varepsilon_{i,1} > \theta_0'\mathbf{x}_i + \varepsilon_{i,0}) \\ &= F(\theta_1'\mathbf{x}_i - \theta_0'\mathbf{x}_i) = F([\theta_1 - \theta_0]'\mathbf{x}_i), \end{aligned} \tag{3.5}$$

where F is the c.d.f. of $\varepsilon_{i,0} - \varepsilon_{i,1}$.

Note that only the difference $\theta_1 - \theta_0$ is identifiable (as opposed to θ_1 *and* θ_0). Indeed, replacing U with aU ($a > 0$) gives the same model. This *scaling* issue can be solved by fixing the variance of $\varepsilon_{i,0} - \varepsilon_{i,1}$.

Example 3.1 (Migration and income)**.** The RUM approach has been used by Nakosteen and Zimmer (1980) to study migration choices. Their model is based on the comparison of marginal costs and benefits associated with migration. The main ingredients of their approach are as follows:

- Wage that can be earned at the present location: $y_p^* = \theta_p'\mathbf{x}_p + \varepsilon_p$.
- Migration cost: $C^* = \theta_c'\mathbf{x}_c + \varepsilon_c$.
- Wage earned elsewhere: $y_m^* = \theta_m'\mathbf{x}_m + \varepsilon_m$.

In this context, agents decision to migrate if $y_m^* > y_p^* + C^*$, i.e. if

$$y^* = y_m^* - y_p^* - C^* = \theta'\mathbf{x} + \underbrace{\varepsilon}_{=\varepsilon_m - \varepsilon_c - \varepsilon_p} > 0,$$

where $\mathbf{x}$ is the union of the $\mathbf{x}_i$s, for $i \in \{p, m, c\}$.

3.1.2 Alternative-Varying Regressors

In some cases, regressors may depend on the considered alternative (0 or 1). For instance:

- When modeling the decision to participate in the labour force (or not), the wage depends on the alternative. Typically, it is zero if the considered agent has decided not to work (and strictly positive otherwise).
- In the context of the choice of transportation mode, "time cost" depends on the considered transportation mode.

In terms of utility, we then have:

$$V_{i,j} = \theta_j^{(u)'} \mathbf{u}_{i,j} + \theta_j^{(v)'} \mathbf{v}_i,$$

where the $\mathbf{u}_{i,j}$'s are regressors associated with agent i, but taking different values for the different choices ($j = 0$ or $j = 1$). In that case, Eq. (3.5) becomes:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = F\left(\theta_1^{(u)'} \mathbf{u}_{i,1} - \theta_0^{(u)'} \mathbf{u}_{i,0} + [\theta_1^{(v)} - \theta_0^{(v)}]' \mathbf{v}_i\right), \tag{3.6}$$

and, if $\theta_1^{(u)} = \theta_0^{(u)} = \theta^{(u)}$ —as is customary— we get:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = F\left(\theta_1^{(u)'} (\mathbf{u}_{i,1} - \mathbf{u}_{i,0}) + [\theta_1^{(v)} - \theta_0^{(v)}]' \mathbf{v}_i\right). \tag{3.7}$$

Example 3.2 (Fishing-mode dataset)**.** The fishing-mode dataset used in Cameron and Trivedi (2005) (Chapters 14 and 15) contains alternative-specific variables. Specifically, for each individual, the price and catch rate depend on the fishing model. In the table reported below, lines `price` and `catch` correspond to the prices and catch rates associated with the chosen alternative.

```
library(mlogit)
data("Fishing",package="mlogit")
stargazer::stargazer(Fishing,type="text")
```

```
##
## ==========================================================
## Statistic        N     Mean    St. Dev.    Min      Max
## ----------------------------------------------------------
## price.beach   1,182  103.422   103.641   1.290   843.186
## price.pier    1,182  103.422   103.641   1.290   843.186
## price.boat    1,182  55.257    62.713    2.290   666.110
## price.charter 1,182  84.379    63.545   27.290   691.110
## catch.beach   1,182   0.241     0.191    0.068    0.533
## catch.pier    1,182   0.162     0.160    0.001    0.452
## catch.boat    1,182   0.171     0.210   0.0002    0.737
## catch.charter 1,182   0.629     0.706    0.002    2.310
## income        1,182 4,099.337 2,461.964 416.667 12,500.000
## ----------------------------------------------------------
```

3.1.3 Estimation

These models can be estimated by Maximum Likelihood approaches (see Section 2.2).

To simplify the exposition, we consider the $\mathbf{x}_i$ vectors of covariates to be deterministic. Moreover, we assume that the r.v. are independent across entities i. How to write the likelihood in that case? It is easily checked that:

$$f(y_i|\mathbf{x}_i;\theta) = g(\theta'\mathbf{x}_i)^{y_i}(1-g(\theta'\mathbf{x}_i))^{1-y_i}.$$

Therefore, if the observations $(\mathbf{x}_i, y_i)$ are independent across entities i, we obtain:

$$\log \mathcal{L}(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n} y_i \log[g(\theta'\mathbf{x}_i)] + (1-y_i)\log[1-g(\theta'\mathbf{x}_i)].$$

The likelihood equation reads (FOC of the optimization program, see Def. 2.7):

$$\frac{\partial \log \mathcal{L}(\theta;\mathbf{y},\mathbf{X})}{\partial \theta} = \mathbf{0},$$

that is:

$$\sum_{i=1}^{n} y_i \mathbf{x}_i \frac{g'(\theta'\mathbf{x}_i)}{g(\theta'\mathbf{x}_i)} - (1-y_i)\mathbf{x}_i \frac{g'(\theta'\mathbf{x}_i)}{1-g(\theta'\mathbf{x}_i)} = \mathbf{0}.$$

This is a nonlinear (multivariate) equation that can be solved numerically. Under regularity conditions (Hypotheses 2.1), we approximately have (Prop. 2.4):

$$\theta_{MLE} \sim \mathcal{N}(\theta_0, \mathbf{I}(\theta_0)^{-1}),$$

where

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

For finite samples, we can e.g. approximate $\mathbf{I}(\theta_0)^{-1}$ by Eq. (2.11):

$$\mathbf{I}(\theta_0)^{-1} \approx -\left(\frac{\partial^2 \log \mathcal{L}(\theta_{MLE};\mathbf{y},\mathbf{X})}{\partial\theta\partial\theta'}\right)^{-1}.$$

In the Probit case (see Table 3.1), it can be shown that we have:

$$\frac{\partial^2 \log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X})}{\partial \theta \partial \theta'} = -\sum_{i=1}^{n} g'(\theta' \mathbf{x}_i)[\mathbf{x}_i \mathbf{x}_i'] \times$$
$$\left[y_i \frac{g'(\theta' \mathbf{x}_i) + \theta' \mathbf{x}_i g(\theta' \mathbf{x}_i)}{g(\theta' \mathbf{x}_i)^2} + (1 - y_i) \frac{g'(\theta' \mathbf{x}_i) - \theta' \mathbf{x}_i (1 - g(\theta' \mathbf{x}_i))}{(1 - g(\theta' \mathbf{x}_i))^2} \right].$$

In the Logit case (see Table 3.1), it can be shown that we have:

$$\frac{\partial^2 \log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X})}{\partial \theta \partial \theta'} = -\sum_{i=1}^{n} g'(\theta' \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i',$$

where $g'(x) = \dfrac{\exp(-x)}{(1 + \exp(-x))^2}$.

Remark that, since $g'(x) > 0$, $-\partial^2 \log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X}) / \partial \theta \partial \theta'$ is positive definite.

3.1.4 Marginal effects

How to measure marginal effects, i.e. the effect on the probability that $y_i = 1$ of a marginal increase of $x_{i,k}$? This object is given by:

$$\frac{\partial \mathbb{P}(y_i = 1 | \mathbf{x}_i; \theta)}{\partial x_{i,k}} = \underbrace{g'(\theta' \mathbf{x}_i)}_{>0} \theta_k,$$

which is of the same sign as θ_k if function g is monotonously increasing.

For agent i, this marginal effect is consistently estimated by $g'(\theta'_{MLE} \mathbf{x}_i) \theta_{MLE,k}$. It is important to see that the marginal effect depends on $\mathbf{x}_i$: respective increases by 1 unit of $x_{i,k}$ (entity i) and of $x_{j,k}$ (entity j) do not necessarily have the same effect on $\mathbb{P}(y_i = 1 | \mathbf{x}_i; \theta)$ as on $\mathbb{P}(y_j = 1 | \mathbf{x}_j; \theta)$. To address this issue, one can compute some measures of "average" marginal effect. There are two main solutions. For each explanatory variable k:

i. Denoting by $\hat{\mathbf{x}}$ the sample average of the $\mathbf{x}_i$s, compute $g'(\theta'_{MLE} \hat{\mathbf{x}}) \theta_{MLE,k}$.
ii. Compute the average (across i) of $g'(\theta'_{MLE} \mathbf{x}_i) \theta_{MLE,k}$.

3.1.5 Goodness of fit

There is no obvious version of "R^2" for binary-choice models. Existing measures are called **pseudo-R^2 measures**.

Denoting by $\log \mathcal{L}_0(\mathbf{y})$ the (maximum) log-likelihood that would be obtained for a model containing only a constant term (i.e. with $\mathbf{x}_i = 1$ for all i), the McFadden's pseudo-R^2 is given by:

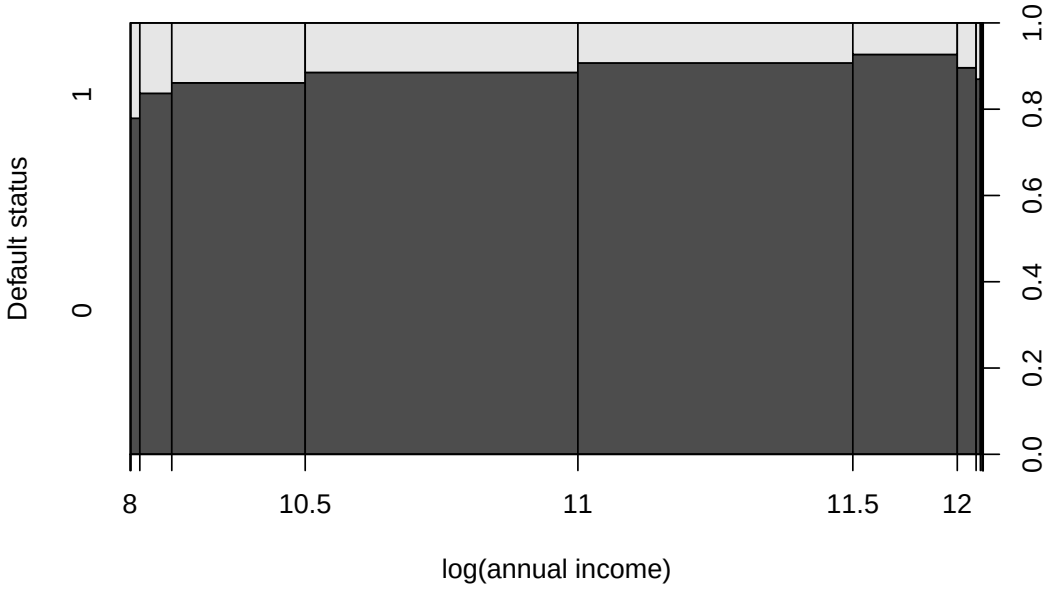
$$R^2_{MF} = 1 - \frac{\log \mathcal{L}(\theta; \mathbf{y})}{\log \mathcal{L}_0(\mathbf{y})}.$$

Intuitively, $R^2_{MF} = 0$ if the explanatory variables do not convey any information on the outcome y. Indeed, in this case, the model is not better than the reference model, that simply captures the fraction of y_i's that are equal to 1.

Example 3.3 (Credit and defaults (Lending-club dataset))**.** This example makes use of the `credit` data of package `AEC`. The objective is to model the default probabilities of borrowers.

Let us first represent the relationship between the fraction of households that have defaulted on their loan and their annual income:

```
library(AEC)
credit$Default <- 0
credit$Default[credit$loan_status == "Charged Off"] <- 1
credit$Default[credit$loan_status ==
                 "Does not meet the credit policy. Status:Charged Off"] <- 1
credit$amt2income <- credit$loan_amnt/credit$annual_inc
plot(as.factor(credit$Default)~log(credit$annual_inc),
     ylevels=2:1,ylab="Default status",xlab="log(annual income)")
```



The previous figure suggests that the effect of annual income on the probability of default is non-monotonous. We will therefore include a quadratic term in one of our specification (namely `eq1` below).

We consider three specifications. The first one (`eq0`), with no explanatory variables, is trivial. It will just be used to compute the pseudo-R^2. In the second (`eq1`), we consider a few covariates (loan amount, the ratio between the amount and annual income, The number of more-than-30 days past-due incidences of delinquency in the borrower's credit file for the past 2 years, and a quadratic function of annual income). In the third model (`eq2`), we add a credit rating.

```
eq0 <- glm(Default ~ 1,data=credit,family=binomial(link="probit"))
eq1 <- glm(Default ~ log(loan_amnt) + amt2income + delinq_2yrs +
             log(annual_inc)+ I(log(annual_inc)^2),
           data=credit,family=binomial(link="probit"))
eq2 <- glm(Default ~ grade + log(loan_amnt) + amt2income + delinq_2yrs +
             log(annual_inc)+ I(log(annual_inc)^2),
           data=credit,family=binomial(link="probit"))
stargazer::stargazer(eq0,eq1,eq2,type="text",no.space = TRUE)
```

```
##
```

```
## ================================================
##                          Dependent variable:
##                     ----------------------------
##                                 Default
##                        (1)        (2)        (3)
## ------------------------------------------------
## gradeB                                     0.400***
##                                            (0.055)
## gradeC                                     0.587***
##                                            (0.057)
## gradeD                                     0.820***
##                                            (0.061)
## gradeE                                     0.874***
##                                            (0.091)
## gradeF                                     1.230***
##                                            (0.147)
## gradeG                                     1.439***
##                                            (0.227)
## log(loan_amnt)                  -0.149**  -0.194***
##                                 (0.060)    (0.061)
## amt2income                      1.266***   1.222***
##                                 (0.383)    (0.393)
## delinq_2yrs                     0.096***    0.009
##                                 (0.034)    (0.035)
## log(annual_inc)                 -1.444**    -0.874
##                                 (0.569)    (0.586)
## I(log(annual_inc)2)             0.064**     0.038
##                                 (0.025)    (0.026)
## Constant            -1.231***   7.937***    4.749
##                      (0.017)    (3.060)    (3.154)
## ------------------------------------------------
## Observations          9,156      9,156      9,156
## Log Likelihood      -3,157.696 -3,120.625 -2,981.343
## Akaike Inf. Crit.   6,317.392  6,253.250  5,986.686
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Let us compute the pseudo R2 for the last two models:

```
logL0 <- logLik(eq0);logL1 <- logLik(eq1);logL2 <- logLik(eq2)
pseudoR2_eq1 <- 1 - logL1/logL0 # pseudo R2
pseudoR2_eq2 <- 1 - logL2/logL0 # pseudo R2
c(pseudoR2_eq1,pseudoR2_eq2)
```

```
## [1] 0.01173993 0.05584870
```

Let us now compute the (average) marginal effects, using method ii of Section 3.1.4:

```
mean(dnorm(predict(eq2)),na.rm=TRUE)*eq2$coefficients
```

```
##          (Intercept)                gradeB                gradeC
##          0.840731198          0.070747353          0.103944305
##                gradeD                gradeE                gradeF
##          0.145089219          0.154773742          0.217702041
##                gradeG       log(loan_amnt)           amt2income
##          0.254722161         -0.034289921          0.216251992
##          delinq_2yrs      log(annual_inc) I(log(annual_inc)^2)
##          0.001574178         -0.154701321          0.006813694
```

There is an issue for the `annual_inc` variable. Indeed, the previous computation does not realize that this variable appears twice among the explanatory variables (through `log(annual_inc)` and `I(log(annual_inc)^2)`). To address this, one can proceed as follows: (1) we construct a new counterfactual dataset where annual incomes are increased by 1%, (2) we use the model to compute model-implied probabilities of default on this new dataset and (3), we subtract the probabilities resulting from the original dataset from these counterfactual probabilities:

```
new_credit <- credit
new_credit$annual_inc <- 1.01 * new_credit$annual_inc
bas_predict_eq2  <- predict(eq2, newdata = credit, type = "response")
# This is equivalent to pnorm(predict(eq2, newdata = credit))
new_predict_eq2  <- predict(eq2, newdata = new_credit, type = "response")
mean(new_predict_eq2 - bas_predict_eq2)
```

```
## [1] -6.562126e-05
```

The negative sign means that, on average across the entities considered in the analysis, a 1% increase in annual income results in a decrease in the default probability. This average effect is however pretty low. To get an economic sense of the size of this effect, let us compute the average effect associated with a unit increase in the number of delinquencies:

```
new_credit <- credit
new_credit$delinq_2yrs <- credit$delinq_2yrs + 1
new_predict_eq2  <- predict(eq2, newdata = new_credit, type = "response")
mean(new_predict_eq2 - bas_predict_eq2)
```

```
## [1] 0.001582332
```

We can employ a likelihood ratio test (see Def. 2.8) to see if the two variables associated with annual income are jointly statistically significant (in the context of `eq1`):

```
eq1restr <- glm(Default ~ log(loan_amnt) + amt2income + delinq_2yrs,
                data=credit,family=binomial(link="probit"))
LRstat <- 2*(logL1 - logLik(eq1restr))
pvalue <- 1 - c(pchisq(LRstat,df=2))
```

The computation gives a p-value of 0.0436.

Example 3.4 (Replicating Table 14.2 of Cameron and Trivedi (2005))**.** The following lines of codes replicate Table 14.2 of Cameron and Trivedi (2005) (see Example 3.2).

```
data.reduced <- subset(Fishing,mode %in% c("charter","pier"))
data.reduced$lnrelp <- log(data.reduced$price.charter/data.reduced$price.pier)
data.reduced$y <- 1*(data.reduced$mode=="charter")
# check first line of Table 14.1:
price.charter.y0 <- mean(data.reduced$pcharter[data.reduced$y==0])
price.charter.y1 <- mean(data.reduced$pcharter[data.reduced$y==1])
```

```
price.charter    <- mean(data.reduced$pcharter)
# Run probit regression:
reg.probit <- glm(y ~ lnrelp,
                  data=data.reduced,
                  family=binomial(link="probit"))
# Run Logit regression:
reg.logit <- glm(y ~ lnrelp,
                 data=data.reduced,
                 family=binomial(link="logit"))
# Run OLS regression:
reg.OLS <- lm(y ~ lnrelp,
              data=data.reduced)
# Replicates Table 14.2 of Cameron and Trivedi:
stargazer::stargazer(reg.logit, reg.probit, reg.OLS,no.space = TRUE,
                     type="text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                         y
##                     logistic   probit             OLS
##                        (1)       (2)               (3)
## ---------------------------------------------------------------
## lnrelp              -1.823*** -1.056***        -0.243***
##                      (0.145)   (0.075)          (0.010)
## Constant            2.053***  1.194***          0.784***
##                      (0.169)   (0.088)          (0.013)
## ---------------------------------------------------------------
## Observations           630       630               630
## R2                                                0.463
## Adjusted R2                                       0.462
## Log Likelihood      -206.827  -204.411
## Akaike Inf. Crit.    417.654   412.822
## Residual Std. Error                          0.330 (df = 628)
## F Statistic                               542.123*** (df = 1; 628)
## ===============================================================
```

```
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

3.1.6 Predictions and ROC curves

How to compute model-implied predicted outcomes? As is the case for y_i, predicted outcomes $\widehat{y}_i$ need to be valued in $\{0,1\}$. A natural choice consists in considering that $\widehat{y}_i = 1$ if $\mathbb{P}(y_i = 1|\mathbf{x}_i;\theta) > 0.5$, i.e., in taking a cutoff of $c = 0.5$. There exist, though, situations where doing so is not relevant. For instance, we may have some models where all predicted probabilities are small, but some less than others. In this context, a model-implied probability of 10% (say) could characterize a "high-risk" entity. However, using a cutoff of 50% would not identify this level of riskiness.

The **receiver operating characteristics (ROC)** curve consitutes a more general approach. The idea is to remain agnostic and to consider all possible values of the cutoff c. It works as follows. For each potential cutoff $c \in [0,1]$, compute (and plot):

- The fraction of $y = 1$ values correctly classified (*True Positive Rate*) against
- The fraction of $y = 0$ values incorrectly specified (*False Positive Rate*).

Such a curve mechanically starts at (0,0) —which corresponds to $c = 1$— and terminates at (1,1) –situation when $c = 0$.

In the case of no predictive ability (worst situation), the ROC curve is a straight line between (0,0) and (1,1).

Example 3.5 (ROC with the fishing-mode dataset)**.** Figure 3.5 shows the ROC curve associated with the probit model estimated in Example 3.4.

```
library(pROC)
predict_model <- predict.glm(reg.probit,type = "response")
roc(data.reduced$y, predict_model, percent=T,
    boot.n=1000, ci.alpha=0.9, stratified=T, plot=TRUE, grid=TRUE,
    show.thres=TRUE, legacy.axes = TRUE, reuse.auc = TRUE,
    print.auc = TRUE, print.thres.col = "blue", ci=TRUE,
    ci.type="bars", print.thres.cex = 0.7, col = 'red',
    main = paste("ROC curve using","(N = ",nrow(data.reduced),")") )
```

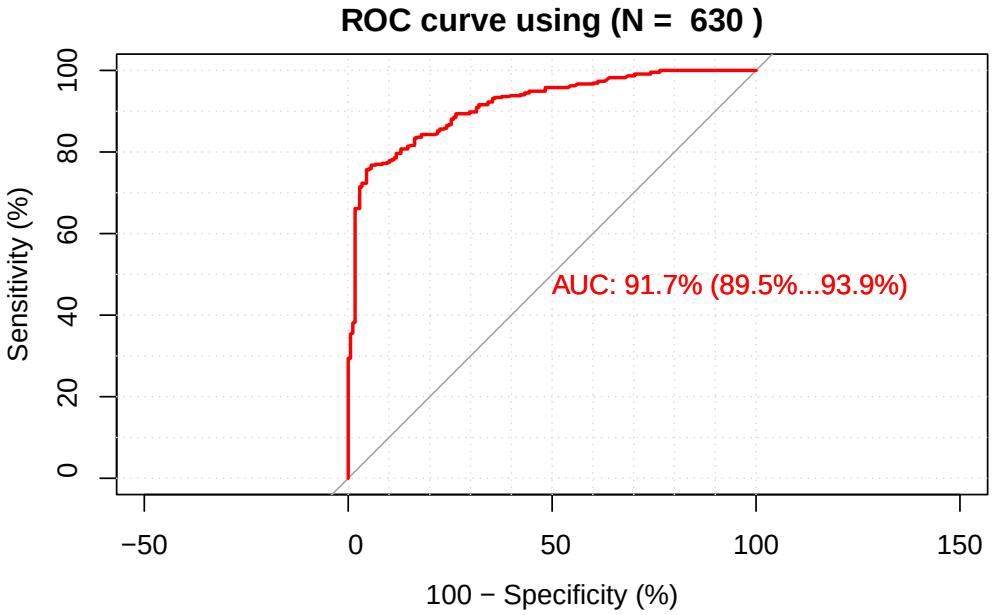


Figure 3.5: Application of the ROC methodology on the fishing-mode dataset.

3.2 Multiple Choice Models

We will now consider cases where the number of possible outcomes (or alternatives) is larger than two. Let us denote by J this number. We have $y_j \in \{1, \ldots, J\}$. This situation arise for instance when the outcome variable reflects:

- Opinions: strongly opposed / opposed / neutral / support (ranked choices),
- Occupational field: lawyer / farmer / engineer / doctor / ...,
- Alternative shopping areas,
- Transportation types.

In a few cases, the values associated with the choices will themselves be meaningful, for example, number of accidents per day: $y = 0, 1, 2, \ldots$ (count data). In most cases, the values are meaningless.

We assume the existence of covariates, gathered in vector $\mathbf{x}_i$ $(K \times 1)$, that are suspected to influence for the probabilities of obtaining the different outcomes $(y_i = j,\, j \in \{1, \dots, J\})$.

In what follows, we will assume that the y_i's are assumed to be independently distributed, with:

$$y_i = \begin{cases} 1 & \text{with probability } g_1(\mathbf{x}_i; \theta) \\ \vdots & \\ J & \text{with probability } g_J(\mathbf{x}_i; \theta). \end{cases} \tag{3.8}$$

(Of course, for all entities (i), we must have $\sum_{j=1}^{J} g_j(\mathbf{x}_i; \theta) = 1$.) Our objective is to estimate the vector of population parameters θ given functional forms for the g_j's.

3.2.1 Ordered case

Sometimes, there exists a natural order for the different alternatives. This is typically the case where respondents have to choose a level of agreement to a statement, e.g.: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree. Another standard case is that of ratings (from A to F, say).

The ordered probit model consists in extending the binary case, considering the latent-variable view of the latter (see Section 3.1.1). Formally, the model is as follows:

$$\mathbb{P}(y_i = j|\mathbf{x}_i) = \mathbb{P}(\alpha_{j-1} < y_i^* < \alpha_j|\mathbf{x}_i), \tag{3.9}$$

where

$$y_i^* = \theta'\mathbf{x}_i + \varepsilon_i,$$

with $\varepsilon_i \sim i.i.d.\, \mathcal{N}(0,1)$. The α_j's, $j \in \{1, \dots, J-1\}$, are (new) parameters that have to be estimated, on top of θ. Naturally, we have $\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}$. Moreover α_0 is $-\infty$ and α_J is $+\infty$, so that Eq. (3.9) is valid for any $j \in \{1, \dots, J\}$ (including 1 and J).

We have:

$$\begin{aligned} g_j(\mathbf{x}_i; \theta, \alpha) = \mathbb{P}(y_i = j|\mathbf{x}_i) &= \mathbb{P}(\alpha_{j-1} < y_i^* < \alpha_j|\mathbf{x}_i) \\ &= \mathbb{P}(\alpha_{j-1} - \theta'\mathbf{x}_i < \varepsilon_i < \alpha_j - \theta'\mathbf{x}_i) \\ &= \Phi(\alpha_j - \theta'\mathbf{x}_i) - \Phi(\alpha_{j-1} - \theta'\mathbf{x}_i), \end{aligned}$$

where Φ is the c.d.f. of $\mathcal{N}(0,1)$.

If, for all i, one of the components of $\mathbf{x}_i$ is equal to 1 (which is what is done in linear regression to introduce an intercept in the specification), then one of the α_j ($j \in \{1, \dots, J-1\}$) is not identified. One can then arbitrarily set $\alpha_1 = 0$. This is what is done in the binary logit/probit cases.

This model can be estimated by maximizing the likelihood function (see Section 2.2). This function is given by:

$$\log \mathcal{L}(\theta, \alpha; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \mathbb{I}_{\{y_i = j\}} \log \left(g_j(\mathbf{x}_i; \theta, \alpha) \right). \tag{3.10}$$

Let us stress that we have two types of parameters to estimate: those included in vector θ, and the α_j's, gathered in vector α.

The estimated values of the θ_j's are slightly more complicated to interpret (at least in term of sign) than in the binary case. Indeed, we have:

$$\mathbb{P}(y_i \leq j | \mathbf{x}_i) = \Phi(\alpha_j - \theta' \mathbf{x}_i) \Rightarrow \frac{\partial \mathbb{P}(y_i \leq j | \mathbf{x}_i)}{\mathbf{x}_i} = - \underbrace{\Phi'(\alpha_j - \theta' \mathbf{x}_i)}_{>0} \theta.$$

Hence the sign of θ_k indicates whether $\mathbb{P}(y_i \leq j | \mathbf{x}_i)$ increases or decreases w.r.t. $x_{i,k}$ (the k^{th} component of $\mathbf{x}_i$). By contrast:

$$\frac{\partial \mathbb{P}(y_i = j | \mathbf{x}_i)}{\mathbf{x}_i} = \underbrace{\left(-F'(\alpha_j + \theta' \mathbf{x}_i) + F'(\alpha_{j-1} + \theta' \mathbf{x}_i) \right)}_{A} \theta.$$

Therefore the signs of the components of θ are not necessarily those of the marginal effects. (For the sign of A is a priori unknown.)

Example 3.6 (Predicting credit ratings (Lending-club dataset))**.** Let us use credit dataset again (see Example 3.3), and let use try and model the ratings attributed by the lending-club:

```
library(AEC)
library(MASS)
credit$emp_length_low5y   <- credit$emp_length %in%
  c("< 1 year","1 year","2 years","3 years","4 years")
```

```
credit$emp_length_high10y <- credit$emp_length=="10+ years"
credit$annual_inc <- credit$annual_inc/1000
credit$loan_amnt  <- credit$loan_amnt/1000
credit$income2loan <- credit$annual_inc/credit$loan_amnt
training <- credit[1:20000,] # sample is reduced
training <- subset(training,grade!=c("E","F","G"))
training <- droplevels(training)
training$grade.ordered <- factor(training$grade,ordered=TRUE,
                                 levels = c("D","C","B","A"))
model1 <- polr(grade.ordered ~ log(loan_amnt) + log(income2loan) + delinq_2yrs
               data=training, Hess=TRUE, method="probit")
model2 <- polr(grade.ordered ~ log(loan_amnt) + log(income2loan) + delinq_2yrs
                 emp_length_low5y + emp_length_high10y,
               data=training, Hess=TRUE, method="probit")
stargazer::stargazer(model1,model2,ord.intercepts = TRUE,type="text",
                     no.space = TRUE)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                            grade.ordered
##                           (1)            (2)
## ------------------------------------------------
## log(loan_amnt)          -0.014         -0.040*
##                        (0.022)        (0.022)
## log(income2loan)       0.115***       0.092***
##                        (0.022)        (0.022)
## delinq_2yrs           -0.399***      -0.404***
##                        (0.025)        (0.025)
## emp_length_low5y                     -0.096***
##                                       (0.027)
## emp_length_high10y                    0.088**
##                                       (0.035)
## D|C                   -0.937***      -1.073***
##                        (0.082)        (0.086)
## C|B                    -0.160**       -0.295***
```

```
##                          (0.082)         (0.085)
## B| A                    0.696***        0.564***
##                          (0.082)         (0.086)
## ------------------------------------------------
## Observations              8,695           8,695
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Predicted ratings (and probabilties of being given a given rating) can be computed as follows:

```
pred.grade <- predict(model1,newdata = training)
# pred.grade = predicted grade, defined as the most likely according model
pred.proba <- predict(model1,newdata = training, type="probs")
```

3.2.2 General multinomial logit model

This section introduces the general multinomial logit model, which is the natural extension of the binary logit model (see Table 3.1). Its general formulation is as follows:

$$g_j(\mathbf{x}_i;\theta) = \frac{\exp(\theta_j'\mathbf{x}_i)}{\sum_{k=1}^{J}\exp(\theta_k'\mathbf{x}_i)}. \tag{3.11}$$

Note that, by construction, $g_j(\mathbf{x}_i;\theta) \in [0,1]$ and $\sum_j g_j(\mathbf{x}_i;\theta) = 1$.

The components of $\mathbf{x}_i$ (regressors, or covariates) may be *alternative-specific* or *alternative invariant* (see also Section 3.1.2). We may, e.g., organize $\mathbf{x}_i$ as follows:

$$\mathbf{x}_i = [\mathbf{u}_{i,1}', \dots, \mathbf{u}_{i,J}', \mathbf{v}_i']', \tag{3.12}$$

where the notations are as in Section 3.1.2, that is:

- $\mathbf{u}_{i,j}$ ($j \in \{1,\dots,J\}$): vector of variables associated with agent i and alternative j (alternative-specific regressors). Examples: Travel time per type of transportation (transportation choice), wage per type of work, cost per type of car.

- $\mathbf{v}_i$: vector of variables associated with agent i but alternative-invariant. Examples: age or gender of agent i,

When $\mathbf{x}_i$ is as in Eq. (3.12), with obvious notations, θ_j is of the form:

$$\theta_j = [\theta_{1,j}^{(u)'}, ..., \theta_{J,j}^{(u)'}, \theta_j^{(v)'}]', \tag{3.13}$$

and $\theta = [\theta_1', ..., \theta_J']'$.

The literature has considered different specific cases of the general multinomial logit model:[1]

- **Conditional logit (CL)** with alternative-varying regressors:

$$\theta_j = [\mathbf{0}', ..., \mathbf{0}', \underbrace{\beta'}_{\text{j}^{th} \text{ position}}, \mathbf{0}', ...]', \tag{3.14}$$

  i.e., we have $\beta = \theta_{1,1}^{(u)} = \cdots = \theta_{J,J}^{(u)}$ and $\theta_{i,j}^{(u)} = \mathbf{0}$ for $i \neq j$.
- **Multinomial logit (MNL)** with alternative-invariant regressors:

$$\theta_j = \left[\mathbf{0}', ..., \mathbf{0}', \theta_j^{(v)'}\right]'. \tag{3.15}$$

- **Mixed logit:**

$$\theta_j = \left[\mathbf{0}', ..., \mathbf{0}', \beta', \mathbf{0}', ..., \mathbf{0}', \theta_j^{(v)'}\right]'. \tag{3.16}$$

Example 3.7 (CL and MNL with the fishing-mode dataset)**.** The following lines replicate Table 15.2 in Cameron and Trivedi (2005) (see also Examples 3.2 and 3.4):

```
# Specify data organization:
library(mlogit)
library(stargazer)
data("Fishing",package="mlogit")
```

[1] The labelling "CL" and "MNL" —used in the literature— are relatively *ad hoc* (see 15.4.1 in Cameron and Trivedi (2005)).

```
Fish <- mlogit.data(Fishing,
                    varying = c(2:9),
                    choice = "mode",
                    shape = "wide")
MNL1 <- mlogit(mode ~ price + catch, data = Fish)
MNL2 <- mlogit(mode ~ price + catch - 1, data = Fish)
MNL3 <- mlogit(mode ~ 0 | income, data = Fish)
MNL4 <- mlogit(mode ~ price + catch | income, data = Fish)
stargazer(MNL1,MNL2,MNL3,MNL4,type="text",no.space = TRUE,
          omit.stat = c("lr"))
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                        mode
##                         (1)        (2)        (3)        (4)
## ---------------------------------------------------------------
## (Intercept):boat      0.871***                0.739***   0.527**
##                       (0.114)                 (0.197)    (0.223)
## (Intercept):charter   1.499***                1.341***   1.694***
##                       (0.133)                 (0.195)    (0.224)
## (Intercept):pier      0.307***                0.814***   0.778***
##                       (0.115)                 (0.229)    (0.220)
## price                -0.025***  -0.020***                -0.025***
##                       (0.002)    (0.001)                 (0.002)
## catch                 0.377***   0.953***                0.358***
##                       (0.110)    (0.089)                 (0.110)
## income:boat                                   0.0001**   0.0001*
##                                              (0.00004)   (0.0001)
## income:charter                                -0.00003   -0.00003
##                                              (0.00004)   (0.0001)
## income:pier                                  -0.0001*** -0.0001**
##                                               (0.0001)   (0.0001)
## ---------------------------------------------------------------
## Observations           1,182      1,182      1,182      1,182
## R2                     0.178      0.014      0.189
```

```
## Log Likelihood      -1,230.784 -1,311.980 -1,477.151 -1,215.138
## ============================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

ML estimation

General multinomial logit models can be estimated by Maximum Likelihood techniques (see Section 2.2). Consider the general model described in Eq. (3.8). It can be noted that:

$$f(y_i|\mathbf{x}_i;\theta) = \prod_{j=1}^{J} g_j(\mathbf{x}_i;\theta)^{\mathbb{I}_{\{y_i=j\}}},$$

which leads to

$$\log f(y_i|\mathbf{x}_i;\theta) = \sum_{j=1}^{J} \mathbb{I}_{\{y_i=j\}} \log\left(g_j(\mathbf{x}_i;\theta)\right).$$

The log-likelihood function is therefore given by:

$$\log \mathcal{L}(\theta;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{n}\sum_{j=1}^{J} \mathbb{I}_{\{y_i=j\}} \log\left(g_j(\mathbf{x}_i;\theta)\right). \tag{3.17}$$

Numerical methods have to be employed in order to find the maximum-likelihood estimate of θ. (Standard packages contain fast algorithms.)

Marginal Effects

Let us consider the computation of marginal effects in the general multinomial logit model (Eq. (3.11)). Using the notation $p_{i,j} \equiv \mathbb{P}(y_i = j|\mathbf{x}_i;\theta)$, we

have:

$$\begin{aligned}\frac{\partial p_{i,j}}{\partial x_{i,s}} &= \frac{\theta_{j,s}\exp(\theta_j'\mathbf{x}_i)\sum_{k=1}^J\exp(\theta_k'\mathbf{x}_i)}{(\sum_{k=1}^J\exp(\theta_k'\mathbf{x}_i))^2} \\ &\quad -\frac{\exp(\theta_j'\mathbf{x}_i)\sum_{k=1}^J\theta_{k,s}\exp(\theta_k'\mathbf{x}_i)}{(\sum_{k=1}^J\exp(\theta_k'\mathbf{x}_i))^2} \\ &= \theta_{j,s}p_{i,j} - \sum_{k=1}^J \theta_{k,s}p_{i,j}p_{i,k} \\ &= p_{i,j}\times\Big(\theta_{j,s} - \underbrace{\sum_{k=1}^J\theta_{k,s}p_{i,k}}_{=\overline{\theta}_s^{(i)}}\Big),\end{aligned}$$

where $\overline{\theta}_s^{(i)}$ does not depend on j. Note that the sign of the marginal effect is not necessarily that of $\theta_{j,s}$.

Random Utility models

The general multinomial logit model may arise as the natural specification arising in structural contexts where agents compare (random) utilities associated with J potential outcomes (see Section 3.1.1 for the binary situation).

Let's drop the i subscript for simplicity and assume that the utility derived form choosing j is given by $U_j = V_j + \varepsilon_j$, where V_j is deterministic (may depend on observed covariates) and ε_j is stochastic. We have (with obvious notations):

$$\begin{aligned}\mathbb{P}(y=j) &= \mathbb{P}(U_j > U_k,\ \forall k\neq j) \\ \mathbb{P}(y=j) &= \mathbb{P}(U_k - U_j < 0,\ \forall k\neq j) \\ \mathbb{P}(y=j) &= \mathbb{P}(\underbrace{\varepsilon_k-\varepsilon_j}_{=:\tilde{\varepsilon}_{k,j}} < \underbrace{V_j - V_k}_{=:-\tilde{V}_{k,j}},\ \forall k\neq j).\end{aligned}$$

The last expression is an $(J-1)$-variate integral. While it has, in general, no analytical solution, Prop. 3.1 shows that it is the case when employing Gumbel distributions (see Def. 3.1).

Definition 3.1 (Gumbel distribution)**.** The c.d.f. of the Gumbel distribution ($\mathcal{W}$) is:

$$F(u) = \exp(-\exp(-u)), \qquad f(u) = \exp(-u - \exp(u)).$$

Remark: if $X \sim \mathcal{W}$, then $\mathbb{E}(X) = 0.577$ (Euler constant)[2] and $\mathbb{V}ar(X) = \pi^2/6$.

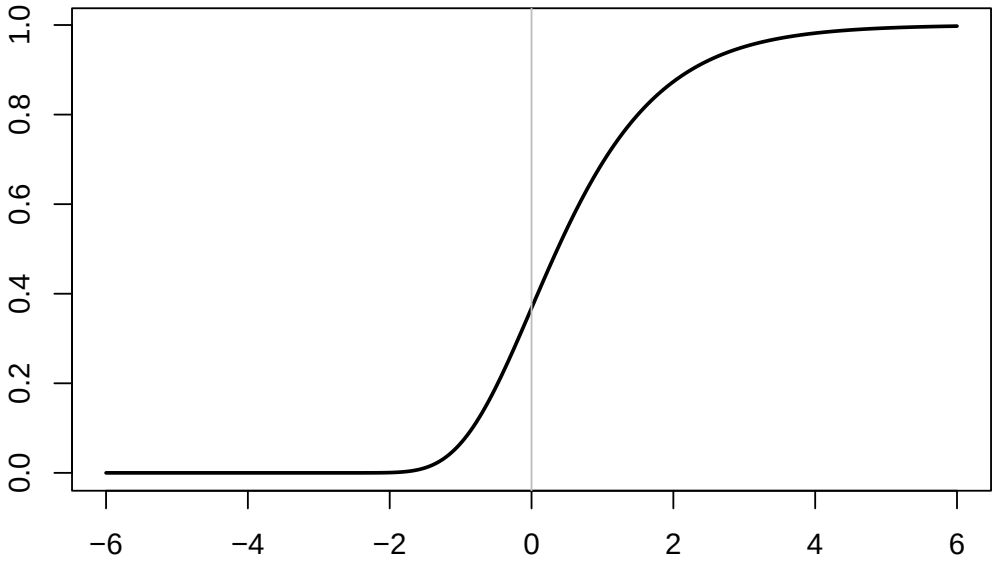


Figure 3.6: C.d.f. of the Gumbel distribution ($F(x) = \exp(-\exp(-x))$).

Proposition 3.1 (Weibull)**.** *In the context of the utility model described above, if $\varepsilon_j \sim$ i.i.d. $\mathcal{W}$, then*

$$\mathbb{P}(y = j) = \frac{\exp(V_j)}{\sum_{k=1}^{J} \exp(V_k)}.$$

Proof. We have:

$$\begin{aligned} \mathbb{P}(y = j) &= \mathbb{P}(\forall\, k \neq j,\, U_k < U_j) = \mathbb{P}(\forall\, k \neq j,\, \varepsilon_k < V_j - V_k + \varepsilon_j) \\ &= \int \prod_{k \neq j} F(V_j - V_k + \varepsilon) f(\varepsilon) d\varepsilon. \end{aligned}$$

[2]The Euler constant γ satisfies $\gamma = \lim_{n\to\infty} \left(-\ln(n) + \sum_{k=1}^{n} \frac{1}{k}\right)$.

After computation, it comes that

$$\prod_{k\neq j} F(V_j - V_k + \varepsilon) f(\varepsilon) = \exp\left[-\varepsilon - \exp(-\varepsilon + \lambda_j)\right],$$

where $\lambda_j = \log\left(1 + \frac{\sum_{k\neq j}\exp(V_k)}{\exp(V_j)}\right)$. We then have:

$$\begin{aligned}\mathbb{P}(y=j) &= \int \exp\left[-\varepsilon - \exp(-\varepsilon + \lambda_j)\right] d\varepsilon \\ &= \int \exp\left[-t - \lambda_j - \exp(-t)\right] d\varepsilon = \exp(-\lambda_j),\end{aligned}$$

which leads to the result. □

Some remarks on identification (see Def. 2.5) are in order.

1. We have:

$$\mathbb{P}(y=j) = \frac{\exp(V_j)}{\sum_{k=1}^{J}\exp(V_k)} = \frac{\exp(V_j^*)}{1+\sum_{k=2}^{J}\exp(V_k^*)},$$

   where $V_j^* = V_j - V_1$. We can therefore always assume that $V_1 = 0$. In the case where $V_{i,j} = \theta_j'\mathbf{x}_i = \beta'\mathbf{u}_{i,j} + \theta_j^{(v)\prime}\mathbf{v}_i$ (see Eqs. (3.12) and (3.16)), we can for instance assume that:

$$\begin{aligned}&(A) \quad \mathbf{u}_{i,1} = 0,\\ &(B) \quad \theta_1^{(v)} = 0.\end{aligned}$$

   If (A) does not hold, we can replace $\mathbf{u}_{i,j}$ with $\mathbf{u}_{i,j} - \mathbf{u}_{i,1}$.
2. If $J = 2$ and $j \in \{0,1\}$ (shift by one unit), we have $\mathbb{P}(y=1|\mathbf{x}) = \frac{\exp(\theta'\mathbf{x})}{1+\exp(\theta'\mathbf{x})}$, this is the logit model (Table 3.1).

Limitations of logit models

In a Logit model, we have:

$$\mathbb{P}(y=j|y\in\{k,j\}) = \frac{\exp(\theta_j'\mathbf{x})}{\exp(\theta_j'\mathbf{x})+\exp(\theta_k'\mathbf{x})}. \tag{3.18}$$

This conditional probability does not depend on other alternatives (i.e., it does not depend on θ_m, $m \neq j, k$). In particular, if $\mathbf{x} = [\mathbf{u}_1', ..., \mathbf{u}_J', \mathbf{v}']'$, then changes in $\mathbf{u}_m$ ($m \neq j$, k) have no impact on the object shown in Eq. (3.18).

That is, a Multinomial Logit can be seen as a series of pairwise comparisons that are unaffected by the characteristics of alternatives. Such a model is said to satisfy the **independence from irrelevant alternatives (IIA)** property. That is, in these models, for any individual, the ratio of probabilities of choosing two alternatives is independent of the availability or attributes of any other alternatives. While this may not sound alarming, there are situations where you would like it not to be the case, this is for instance the case when you want to extrapolate the results of your estimated model to a situation where there is a novel outcome that is highly susbstitutable to one of the previous ones. This can be illustrated with the famous "red-blue bus" example:

Example 3.8 (Red-blue bus and IIA)**.** Assume one has a logit model capturing the decision to travel using either a car ($y = 1$) or a (red) bus ($y = 2$). Assume you want to augment this model to allow for a third choice ($y = 3$): travel with a blue bus. If a blue bus ($y = 3$) is exactly as a red bus, except for the color, then one would expect to have:

$$\mathbb{P}(y = 3 | y \in \{2, 3\}) = 0.5,$$

i.e. $\theta_2 = \theta_3$.

Assume we had $V_1 = V_2$. We expect to have $V_2 = V_3$ (hence $p_2 = p_3$). A multinomial logit model would then imply $p_1 = p_2 = p_3 = 0.33$. It would however seem more reasonable to have $p_1 = p_2 + p_3 = 0.5$ and $p_2 = p_3 = 0.25$.

3.2.3 Nested logits

Nested Logits are natural extensions of logit models when choices feature a nesting structure. This approach is relevant when it makes sense to group some choices into the same *nest*, also called *limbs*. Intuitively, this framework is consistent with the idea according to which, for each agent, there exist unobserved nest-specific variables.

The setup is as follows: we consider J *limbs*. For each limb j, we have K_j *branches*. Let us denotes by y_1 the limb choice (i.e., $y_1 \in \{1, ..., J\}$) and by

y_2 the branch choice (with $y_2 \in \{1, \dots, K_j\}$). The utility associated with the pair of choices (j, k) is given by

$$U_{j,k} = V_{j,k} + \varepsilon_{j,k}.$$

We have:

$$\mathbb{P}[(y_1, y_2) = (j, k)|\mathbf{x}] = \mathbb{P}(U_{j,k} > U_{l,m},\ (l, m) \neq (j, k)|\mathbf{x}).$$

One usually make the following two assumptions:

i. The deterministic part of the utility is given by $V_{j,k} = \mathbf{u}_j'\alpha + \mathbf{v}_{j,k}'\beta_j$, where α is common to all nests and the β_j's are nest-specific.

ii. The disturbances ε follow the Generalized Extreme Value (GEV) distribution (see Def. 4.15).

The following figure displays simulations of pairs $(\varepsilon_1, \varepsilon_2)$ drawn from GEV distributions for different values of ρ. The simulation approach is based on Bhat. The code used to produce this chart is provided in Appendix 4.6.1.

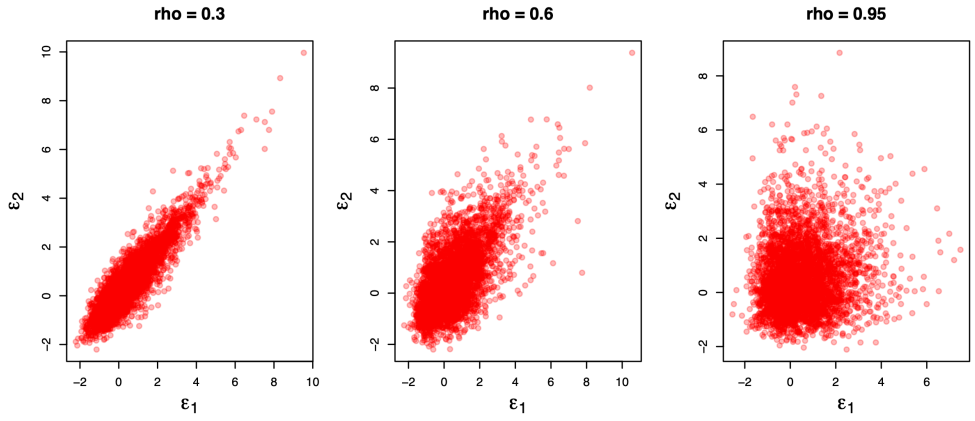


Figure 3.7: GEV simulations.

Under (i) and (ii), we have:

$$\begin{aligned}\mathbb{P}[(y_1,y_2)=(j,k)|\mathbf{x}] &= \underbrace{\frac{\exp(\mathbf{u}_j'\alpha+\rho_j I_j)}{\sum_{m=1}^{J}\exp(\mathbf{u}_m'\alpha+\rho_m I_m)}}_{=\mathbb{P}[y_1=j|\mathbf{x}]}\times \\ &\quad \underbrace{\frac{\exp(\mathbf{v}_{j,k}'\beta_j/\rho_j)}{\sum_{l=1}^{K_j}\exp(\mathbf{v}_{j,l}'\beta_j/\rho_j)}}_{=\mathbb{P}[y_2=k|y_1=j,\mathbf{x}]}, \end{aligned} \tag{3.19}$$

where I_j's are called inclusive values (or log sums), given by:

$$I_j=\log\left(\sum_{l=1}^{K_j}\exp(\mathbf{v}_{j,l}'\beta_j/\rho_j)\right).$$

Some remarks are in order:

a. It can be shown that $\rho_j=\sqrt{1-\mathbb{C}or(\varepsilon_{j,k},\varepsilon_{j,l})}$, for $k\neq l$.
b. $\rho_j=1$ implies that $\varepsilon_{j,k}$ and $\varepsilon_{j,l}$ are uncorrelated (we are then back to the multinomial logit case).
c. When $J=1$:

$$F([\varepsilon_1,...,\varepsilon_K]',\rho)=\exp\left(-\left(\sum_{k=1}^{K}\exp(-\varepsilon_k/\rho)\right)^{\rho}\right).$$

d. We have:

$$I_j=\mathbb{E}(\max_k(U_{j,k})) = \mathbb{E}(\max_k(V_{j,k}+\varepsilon_{j,k})),$$

The inclusive values can therefore be seen as measures of the relative attractiveness of a nest.

This approach allows for some level of correlation across the $\varepsilon_{j,k}$ (for a given j). This can be interpreted as the existence of an (unobserved) *common error component* for the alternatives of a same nest. This component contributes to making the alternatives of a given nest more similar. In other words, this

approach can accommodate a higher sensitivity (cross-elasticity) between the alternatives of a given nest.

Note that if the common component is reduced to zero (i.e. $\rho_i = 1$), the model boils down to the multinomial logit model with no covariance of error terms among the alternatives.

Contrary to the general multinmial model, nested logits can solve the Red-Blue problem described in Section 3.2.2 (see Example 3.8). Assume you have estimated a model specifying $U_1 = V_1 + \varepsilon_1$ (car choice) and $U_2 = V_2 + \varepsilon_2$ (red bus choice). You can then assume that the blue-bus utility is of the form $U_3 = V_2 + \varepsilon_3$ where ε_3 is perfectly correlated to ε_2. This is done by redefining the set of choices as follows:

$$\begin{aligned} j = 1 &\Leftrightarrow (j' = 1, k = 1) \\ j = 2 &\Leftrightarrow (j' = 2, k = 1) \\ j = 3 &\Leftrightarrow (j' = 2, k = 2), \end{aligned}$$

and by setting $\rho_2 \to 0$.

IIA holds within a nest, but not when considering alternatives in different nests. Indeed, using Eq. (3.19):

$$\frac{\mathbb{P}[y_1 = j, y_2 = k_A|\mathbf{x}]}{\mathbb{P}[y_1 = j, y_2 = k_B|\mathbf{x}]} = \frac{\exp(\mathbf{v}'_{j,k_A}\beta_j/\rho_j)}{\exp(\mathbf{v}'_{j,k_B}\beta_j/\rho_j)},$$

i.e. we have IIA in nest j.

By contrast:

$$\begin{aligned} \frac{\mathbb{P}[y_1 = j_A, y_2 = k_A|\mathbf{x}]}{\mathbb{P}[y_1 = j_B, y_2 = k_B|\mathbf{x}]} &= \frac{\exp(\mathbf{u}'_{j_A}\alpha + \rho_{j_A} I_{j_A}) \exp(\mathbf{v}'_{j_A,k_A}\beta_{j_A}/\rho_{j_A})}{\exp(\mathbf{u}'_{j_B}\alpha + \rho_{j_B} I_{j_B}) \exp(\mathbf{v}'_{j_B,k_B}\beta_{j_B}/\rho_{j_B})} \times \\ &\quad \frac{\sum_{l=1}^{K_{j_B}} \exp(\mathbf{v}'_{j_B,l}\beta_{j_B}/\rho_{j_B})}{\sum_{l=1}^{K_{j_A}} \exp(\mathbf{v}'_{j_A,l}\beta_{J_A}/\rho_{j_A})}, \end{aligned}$$

which depends on the expected utilities of all alternatives in nest j_A and j_B. So the IIA does not hold.

Example 3.9 (Travel-mode dataset)**.** Let us illustrate nested logits on the travel-mode dataset used, e.g., by Hensher and Greene (2002) (see also Heiss (2002)).

```
library(mlogit)
library(stargazer)
data("TravelMode", package = "AER")
Prepared.TravelMode <- mlogit.data(TravelMode,chid.var = "individual",
                                   alt.var = "mode",choice = "choice",
                                   shape = "long")
# Fit a multinomial model:
hl <- mlogit(choice ~ wait + travel + vcost, Prepared.TravelMode,
             method = "bfgs", heterosc = TRUE, tol = 10)
## Fit a nested logit model:
TravelMode$avincome <- with(TravelMode, income * (mode == "air"))
TravelMode$time <- with(TravelMode, travel + wait)/60
TravelMode$timeair <- with(TravelMode, time * I(mode == "air"))
TravelMode$income <- with(TravelMode, income / 10)
# Hensher and Greene (2002), table 1 p.8-9 model 5
TravelMode$incomeother <- with(TravelMode,
                               ifelse(mode %in% c('air', 'car'), income, 0))
nl1 <- mlogit(choice ~ gcost + wait + incomeother, TravelMode,
              shape='long', # Indicates how the dataset is organized
              alt.var='mode', # variable that defines the alternative choices
              nests=list(public=c('train', 'bus'),
                         car='car',air='air'), # defines the "limbs".
              un.nest.el = TRUE)
nl2 <- mlogit(choice ~ gcost + wait + time, TravelMode,
              shape='long', # Inidcates how the dataset is organized
              alt.var='mode', # variable that defines the alternative choices
              nests=list(public=c('train', 'bus'),
                         car='car',air='air'), # defines the "limbs".
              un.nest.el = TRUE)
stargazer(nl1,nl2,type="text",no.space = TRUE)
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                choice
##                           (1)            (2)
```

```
## ------------------------------------------------
## (Intercept):train      -0.211         -0.284
##                        (0.562)        (0.551)
## (Intercept):bus        -0.824         -0.712
##                        (0.708)        (0.690)
## (Intercept):car       -5.237***      -3.845***
##                        (0.785)        (0.844)
## gcost                 -0.013***       -0.004
##                        (0.004)        (0.006)
## wait                  -0.088***      -0.089***
##                        (0.011)        (0.011)
## incomeother            0.430***
##                        (0.113)
## time                                 -0.202***
##                                       (0.060)
## iv                     0.835***       0.877***
##                        (0.192)        (0.198)
## ------------------------------------------------
## Observations             210            210
## R2                      0.328          0.313
## Log Likelihood         -190.779       -194.841
## LR Test (df = 7)      185.959***     177.836***
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

3.3 Tobit models

In some situations, the dependent variable is incompletely observed, which may result in a non-representative sample. Typically, in some cases, observations of the dependent variable can have a lower and/or an upper limit, while the "true", underlying, dependent variable has not. In this case, OLS regression may lead to inconsistent parameter estimates.

Tobit models have been designed to address some of these situations. This approach has been named after James Tobin, who developed this model in the late 50s (see Tobin (1956)).

Figure 3.8 illustrates the situation. The dots (white and black) represent

the "true" observations. Now, assume that only the black are observed. If ones uses these observations in an OLS regression to estimate the relatonship between x and y, then one gets the red line. It is clear that the sensitivity of y to x is then underestimated. The blue line is the line one would obtain if white dots were also observed; the grey line represents the model used to genrate the data ($y_i = x_i + \varepsilon_i$).

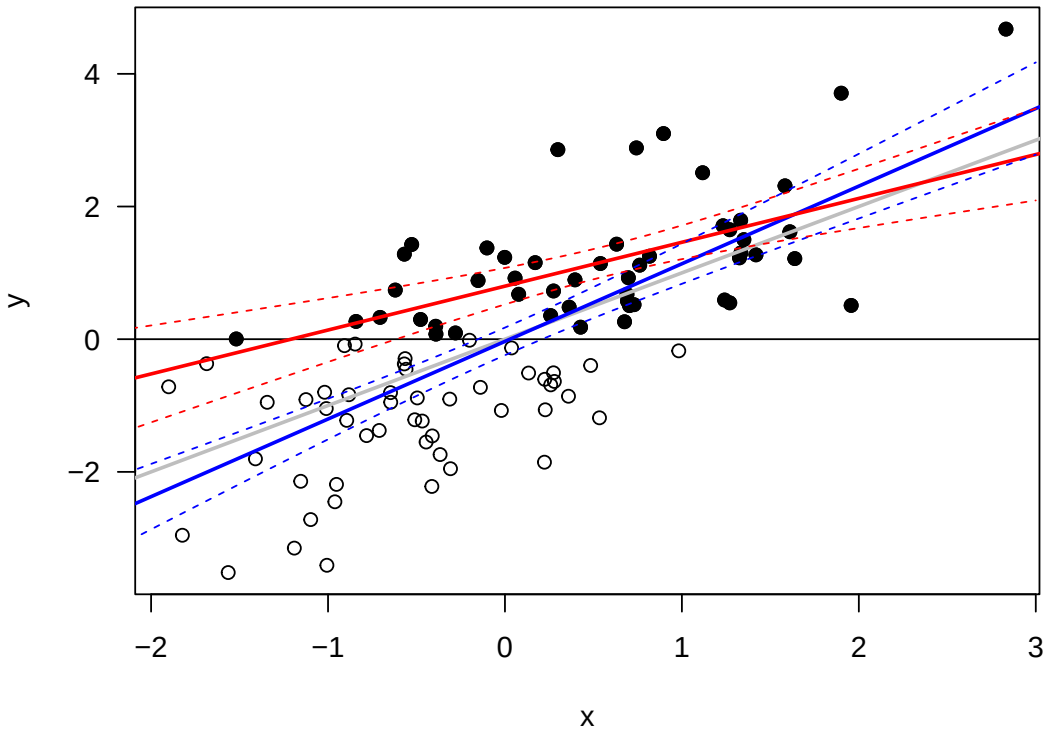


Figure 3.8: Bias in the case of sample selection. The grey line represents the population regression line. The model is $y_i = x_i + \varepsilon_i$, with $\varepsilon_{i,t} \sim \mathcal{N}(0, 1)$. The red line is the OLS regression line based on black dots only.

Assume that the (partially) observed dependent variable follows:

$$y^* = \beta' \mathbf{x} + \varepsilon,$$

with ε is drawn from a distribution characterized by a p.d.f. denoted by f^*_γ and a c.d.f. denoted by F^*_γ; these functions depend on a vector of parameters γ.

The observed dependent variable is:

$$\text{Censored case:} \qquad y = \begin{cases} y^* & if \quad y^* > L \\ L & if \quad y^* \leq L, \end{cases}$$

$$\text{Truncated case:} \qquad y = \begin{cases} y^* & if \quad y^* > L \\ - & if \quad y^* \leq L, \end{cases}$$

where "$-$" stands for missing observations.

This formulation is easily extended to censoring from above ($L \to U$), or censoring from both below and above.

The model parameters are gathered in vector $\theta = [\beta', \gamma']'$. Let us write the conditional p.d.f. of the observed variable:

$$\text{Censored case:} \qquad f(y|\mathbf{x};\theta) = \begin{cases} f^*_\gamma(y - \beta'\mathbf{x}) & if \quad y > L \\ F^*_\gamma(L - \beta'\mathbf{x}) & if \quad y = L, \end{cases}$$

$$\text{Truncated case:} \qquad f(y|\mathbf{x};\theta) = \frac{f^*_\gamma(y - \beta'\mathbf{x})}{1 - F^*_\gamma(L - \beta'\mathbf{x})} \quad \text{with} \quad y > L.$$

The (conditional) log-likelihood function is then given by:

$$\log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\theta).$$

In the censored case, we have:

$$\begin{aligned} \log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X}) \quad = \quad & \sum_{i=1}^{n} \Big\{ \mathbb{I}_{\{y_i = L\}} \log \left[F^*_\gamma(L - \beta'\mathbf{x}_i) \right] + \\ & \mathbb{I}_{\{y_i > 0\}} \log \left[f^*_\gamma(y_i - \beta'\mathbf{x}_i) \right] \Big\}. \end{aligned}$$

The Tobit, or censored/truncated normal regression model, corresponds to the case described above, but with Gaussian errors ε. Specifically:

$$y^* = \beta'\mathbf{x} + \varepsilon,$$

with $\varepsilon \sim \ i.i.d. \ \mathcal{N}(0, \sigma^2)$ ($\Rightarrow \gamma = \sigma^2$).

Without loss of generality, we can assume that $L = 0$. (One can shift observed data if necessary.)

- The censored density (with $L = 0$) is given by:

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \beta'\mathbf{x})^2\right)\right]^{\mathbb{I}_{\{y>0\}}} \left[1 - \Phi\left(\frac{\beta'\mathbf{x}}{\sigma}\right)\right]^{\mathbb{I}_{\{y=0\}}}.$$

- The truncated density (with $L = 0$) is given by:

$$f(y) = \frac{1}{\Phi(\beta'\mathbf{x})} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \beta'\mathbf{x})^2\right).$$

Results usually heavily rely on distributional assumptions (more than in uncensored/untruncated case). The framework is easy to extend to an heteroskedastic case, for instance by setting $\sigma_i^2 = \exp(\alpha'\mathbf{x}_i)$. Such a situation is illustrated by Figure 3.9.

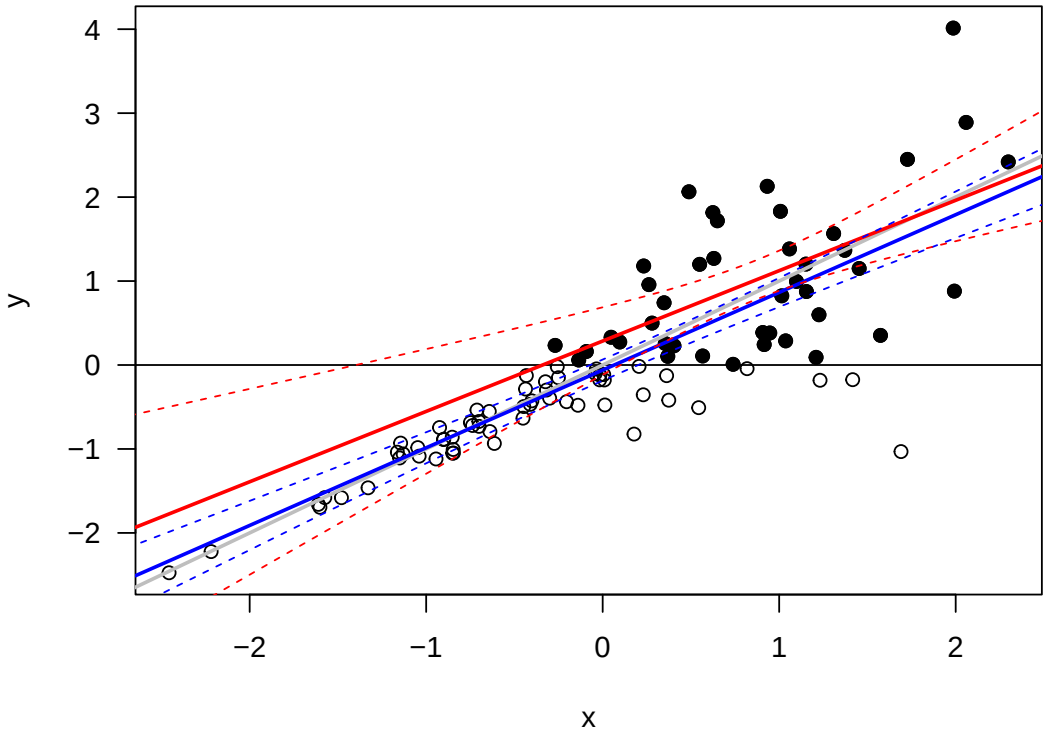


Figure 3.9: Censored dataset with heteroskedasticitiy. The model is $y_i = x_i + \varepsilon_i$, with $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$ where $\sigma_i = \exp(-1 + x_i)$.

Let us consider the conditional means of y in the general case, i.e., for any ε distribution. Assume $\mathbf{x}$ is observed, such that expectations are conditional on $\mathbf{x}$.

- For data that are left-truncated at 0, we have:

$$\mathbb{E}(y) = \mathbb{E}(y^*|y^*>0) = \underbrace{\beta'\mathbf{x}}_{=\mathbb{E}(y^*)} + \underbrace{\mathbb{E}(\varepsilon|\varepsilon>-\beta'\mathbf{x})}_{>0} > \mathbb{E}(y^*).$$

- Consider data that are left-censored at 0. By Bayes, we have:

$$f_{y^*|y^*>0}(u) = \frac{f_{y^*}(u)}{\mathbb{P}(y^*>0)}\mathbb{I}_{\{u>0\}}.$$

  Therefore:

$$\begin{aligned}\mathbb{E}(y^*|y^*>0) &= \frac{1}{\mathbb{P}(y^*>0)}\int_{-\infty}^{\infty} u\, f_{y^*}(u)\mathbb{I}_{\{u>0\}}du \\ &= \frac{1}{\mathbb{P}(y^*>0)}\mathbb{E}(\underbrace{y^*\mathbb{I}_{\{y^*>0\}}}_{=y}),\end{aligned}$$

  and, further:

$$\begin{aligned}\mathbb{E}(y) &= \mathbb{P}(y^*>0)\mathbb{E}(y^*|y^*>0) \\ &> \mathbb{E}(y^*) = \mathbb{P}(y^*>0)\mathbb{E}(y^*|y^*>0) + \mathbb{P}(y^*<0)\underbrace{\mathbb{E}(y^*|y^*<0)}_{<0}.\end{aligned}$$

Now, let us come back to the Tobit (i.e., Gaussian case) case.

- For data that are left-truncated at 0:

$$\begin{aligned}\mathbb{E}(y) &= \beta'\mathbf{x} + \mathbb{E}(\varepsilon|\varepsilon>-\beta'\mathbf{x}) \\ &= \beta'\mathbf{x} + \sigma\underbrace{\frac{\phi(\beta'\mathbf{x}/\sigma)}{\Phi(\beta'\mathbf{x}/\sigma)}}_{=:\lambda(\beta'\mathbf{x}/\sigma)} = \sigma\left(\frac{\beta'\mathbf{x}}{\sigma} + \lambda\left(\frac{\beta'\mathbf{x}}{\sigma}\right)\right). \quad (3.20)\end{aligned}$$

  where the penultimate line is obtained by using Eq. (4.4).

- For data that are left-censored at 0:

$$\begin{aligned}\mathbb{E}(y) &= \mathbb{P}(y^* > 0)\mathbb{E}(y^*|y^* > 0) \\ &= \Phi\left(\frac{\beta'\mathbf{x}}{\sigma}\right)\sigma\left(\frac{\beta'\mathbf{x}}{\sigma} + \lambda\left(\frac{\beta'\mathbf{x}}{\sigma}\right)\right).\end{aligned}$$

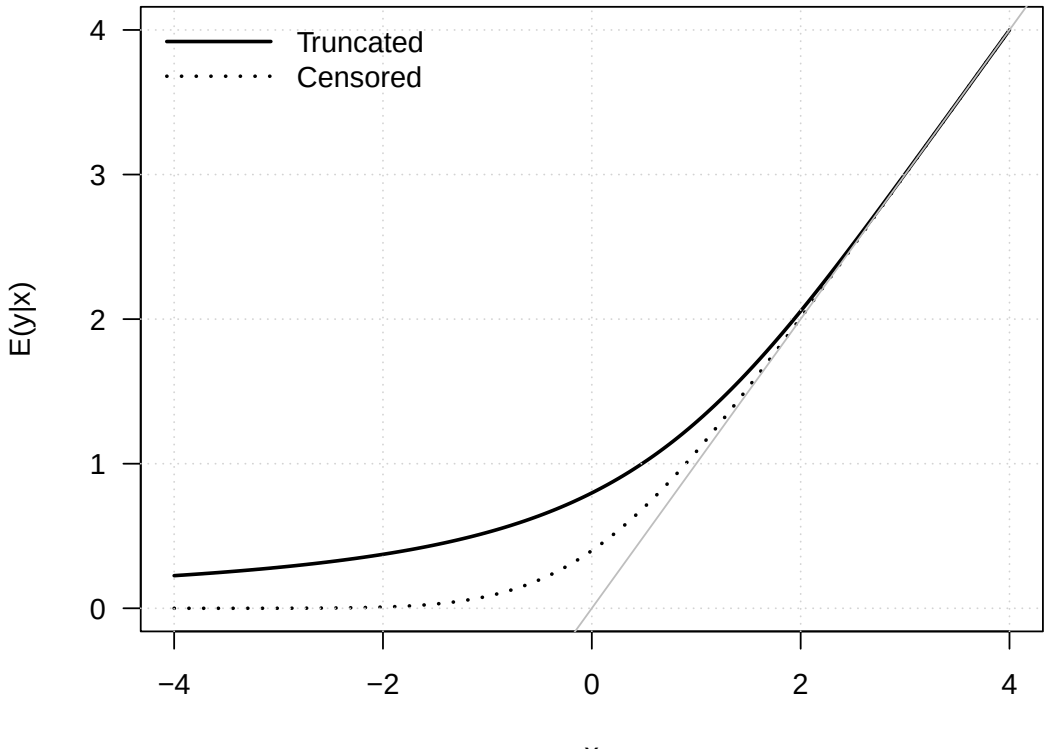


Figure 3.10: Conditional means of y in Tobit models. The model is $y_i = x_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, 1)$.

Heckit regression

The previous formula (Eq. (3.20)) can in particular be used in an alternative estimation approach, namely the Heckman two-step estimation. This approach is based on two steps:[3]

[3]See 16.10.2 of Cameron and Trivedi (2005) for the derivation of asymptotic standard errors of β.

1. Using the complete sample, fit a Probit model of $\mathbb{I}_{\{y_i>0\}}$ on $\mathbf{x}$. This provides a consistent estimate of $\frac{\beta}{\sigma}$, and therefore of $\lambda(\beta'\mathbf{x}/\sigma)$. (Indeed, if $z_i \equiv \mathbb{I}_{\{y_i>0\}}$, then $\mathbb{P}(z_i = 1|\mathbf{x}_i; \beta/\sigma) = \Phi(\beta'\mathbf{x}_i/\sigma)$.)
2. Using the truncated sample only: run an OLS regression of $\mathbf{y}$ on $\{\mathbf{x}, \lambda(\beta'\mathbf{x}/\sigma)\}$ (having Eq. (3.20) in mind). This provides a consistent estimate of (β, σ).

The underlying specification is of the form:

$$\text{Conditional mean} + \text{disturbance}.$$

where "Conditional mean" comes from Eq. (3.20) and "disturbance" is an error with zero conditional mean.

This approach is also applied to the case of **sample selection models** (Section 3.4).

Example 3.10 (Wage prediction)**.** The present example is based on the dataset used in Mroz (1987) (whicht is part of the `sampleSelection` package).

```
library(sampleSelection)
library(AER)
data("Mroz87")
Mroz87$lfp.yesno <- NaN
Mroz87$lfp.yesno[Mroz87$lfp==1] <- "yes"
Mroz87$lfp.yesno[Mroz87$lfp==0] <- "no"
Mroz87$lfp.yesno <- as.factor(Mroz87$lfp.yesno)
ols <- lm(wage ~ educ + exper + I( exper^2 ) + city,data=subset(Mroz87,lfp==1))
tobit <- tobit(wage ~ educ + exper + I( exper^2 ) + city,
               left = 0, right = Inf,
               data=Mroz87)
Heckit <- heckit(lfp ~ educ + exper + I( exper^2 ) + city, # selection equation
                 wage ~ educ + exper + I( exper^2 ) + city, # outcome equation
                 data=Mroz87 )

stargazer(ols,Heckit,tobit,no.space = TRUE,type="text",omit.stat = "f")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                        wage
##                          OLS          Heckman          Tobit
##                                      selection
##                          (1)            (2)             (3)
## ---------------------------------------------------------------
## educ                   0.481***       0.759***        0.642***
##                        (0.067)        (0.270)         (0.081)
## exper                   0.032          0.430          0.461***
##                        (0.062)        (0.369)         (0.068)
## I(exper2)              -0.0003        -0.008          -0.009***
##                        (0.002)        (0.008)         (0.002)
## city                    0.449          0.113           -0.087
##                        (0.318)        (0.522)         (0.378)
## Constant               -2.561***      -12.251         -10.395***
##                        (0.929)        (8.853)         (1.095)
## ---------------------------------------------------------------
## Observations             428            753             753
## R2                      0.125          0.128
## Adjusted R2             0.117          0.117
## Log Likelihood                                      -1,462.700
## rho                                    1.063
## Inverse Mills Ratio                5.165 (4.594)
## Residual Std. Error 3.111 (df = 423)
## Wald Test                                        153.892*** (df = 4)
## ===============================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Figure 3.11 shows that, low wages, the OLS model tends to over-predict wages. The slope between observed and Tobit-predicted wages is closer to one (the adjustment line is closer to the 45-degree line.)

Two-part model

In the standard Tobit framework, the model determining censored —or truncated— data *censoring mechanism* is the same as the one determining

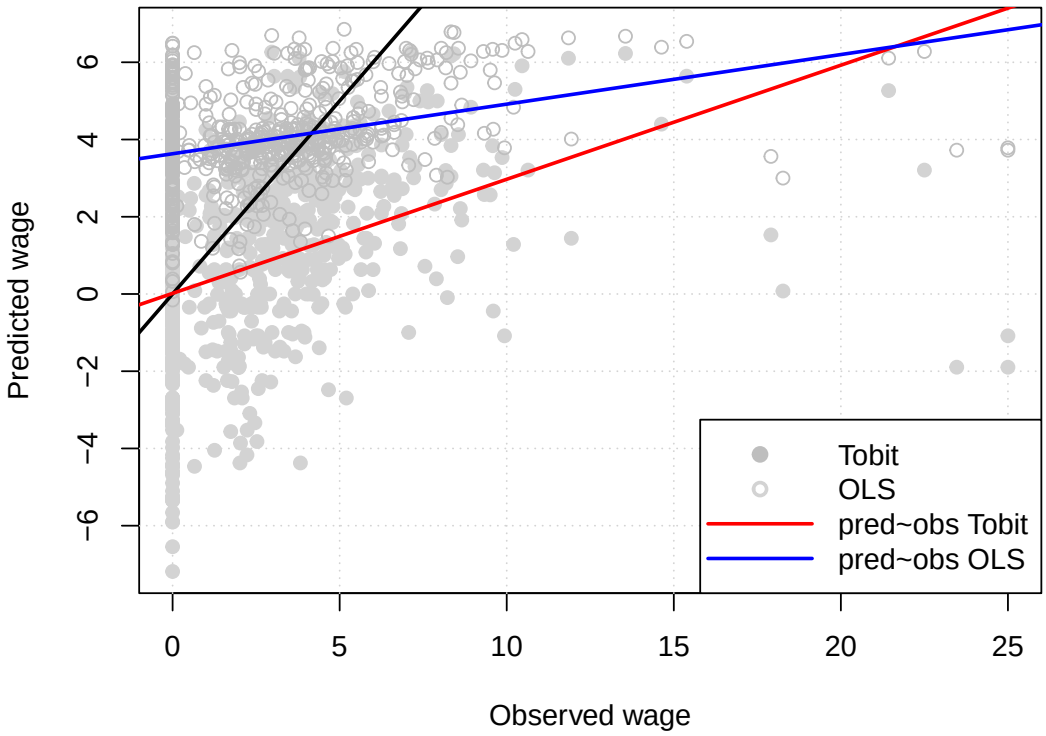


Figure 3.11: Predicted versus observed wages.

non-censored —or observed— data *outcome mechanism*. A two-part model adds flexibility by permitting the zeros and non-zeros to be generated by different densities. The second model characterizes the outcome *conditional on* the outcome being observed.

In a seminal paper, Duan et al. (1983) employ this methodology to account for individual annual hospital expenses. The two models are then as follows:

- 1^{st} model: $\mathbb{P}(hosp = 1|\mathbf{x}) = \Phi(\mathbf{x}_1'\beta_1)$,
- 2^{nd} model: $Expense = \exp(\mathbf{x}_2'\beta_2 + \eta)$, with $\eta \sim\ i.i.d.\, \mathcal{N}(0, \sigma_2^2)$.

Specifically:

$$\mathbb{E}(Expense|\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1'\beta_1) \exp\left(\mathbf{x}_2'\beta_2 + \frac{\sigma_2^2}{2}\right).$$

In sample-selection models, studied in the next section, one specifies the joint distribution for the censoring and outcome mechanisms (while the two parts are independent here).

3.4 Sample Selection Models

The situation tackled by sample-selection models is the following. The dependent variable of interest, denoted by y_2, depends on observed variables $\mathbf{x}_2$. Observing y_2, or not, depends on the value of a latent variable (y_1^*) that is correlated to observed variables $\mathbf{x}_1$. The difference w.r.t. the two-part model skethed above is that, even conditionally on $(\mathbf{x}_1, \mathbf{x}_2)$, y_1^* and y_2 may be correlated.

As in the Tobit case, even in the simplest case of population conditional mean linear in regressors (i.e. $y_2 = \mathbf{x}_2'\beta_2 + \varepsilon_2$), OLS regression leads to inconsistent parameter estimates because the sample is not representative of the population.

There are two latent variables: y_1^* and y_2^*. We observe y_1 and, if the considered entity "participates", we also observe y_2. More specifically:

$$\begin{aligned} y_1 &= \left\{ \begin{array}{ll} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0 \end{array} \right. \quad \text{(participation equation)} \\ y_2 &= \left\{ \begin{array}{ll} y_2^* & \text{if } y_1 = 1 \\ - & \text{if } y_1 = 0 \end{array} \right. \quad \text{(outcome equation).} \end{aligned}$$

Moreover:

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\beta_1 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2'\beta_2 + \varepsilon_2. \end{aligned}$$

Note that the Tobit model (Section 3.3) is the special case where $y_1^* = y_2^*$.

Usually:

$$\left[\begin{array}{c} \varepsilon_1 \\ \varepsilon_2 \end{array} \right] \sim \mathcal{N}\left(\mathbf{0}, \left[\begin{array}{cc} 1 & \rho\sigma_2 \\ \rho\sigma_2 & \sigma_2^2 \end{array} \right] \right).$$

Let us derive the likelihood associated with this model. We have:

$$\begin{aligned} f(\underbrace{0}_{=y_1}, \underbrace{\_}_{=y_2} | \mathbf{x}; \theta) &= \mathbb{P}(y_1^* \leq 0) = \Phi(-\mathbf{x}_1'\beta_1) \qquad (3.21) \\ f(1, y_2 | \mathbf{x}; \theta) &= f(y_2^* | \mathbf{x}; \theta) \mathbb{P}(y_1^* > 0 | y_2^*, \mathbf{x}; \theta) \\ &= \frac{1}{\sigma} \phi\left(\frac{y_2 - \mathbf{x}_2'\beta_2}{\sigma} \right) \mathbb{P}(y_1^* > 0 | y_2, \mathbf{x}; \theta). \qquad (3.22) \end{aligned}$$

Let us compute $\mathbb{P}(y_1^* > 0 | y_2, \mathbf{x}; \theta)$. By Prop. 4.16 (in Appendix 4.4), applied to $(\varepsilon_1, \varepsilon_2)$, we have:

$$y_1^* | y_2 \sim \mathcal{N}\left(\mathbf{x}_1'\beta_1 + \frac{\rho}{\sigma_2}(y_2 - \mathbf{x}_2'\beta_2), 1 - \rho^2 \right).$$

which leads to

$$\mathbb{P}(y_1^* > 0 | y_2, \mathbf{x}; \theta) = \Phi\left(\frac{\mathbf{x}_1'\beta_1 + \frac{\rho}{\sigma_2}(y_2 - \mathbf{x}_2'\beta_2)}{\sqrt{1 - \rho^2}} \right). \qquad (3.23)$$

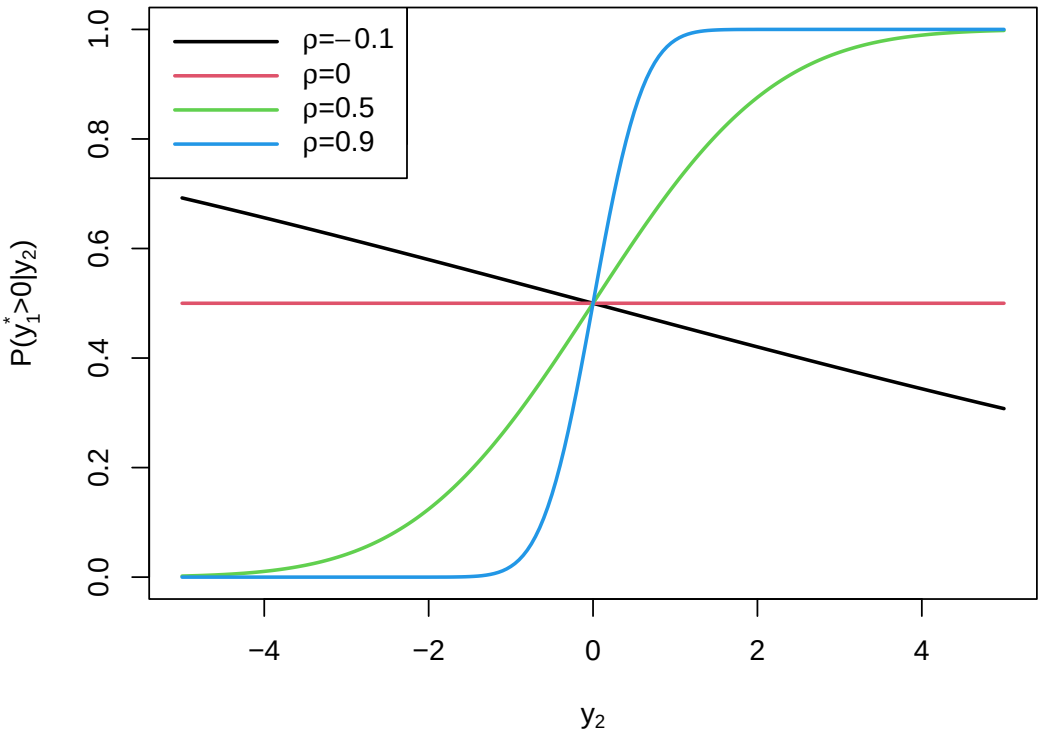


Figure 3.12: Probability of observing y_2 depending on its value, for different values of conditional correlation between y_2 and y_1^*.

Figure 3.12 displays $\mathbb{P}(y_1^* > 0|y_2, \mathbf{x}; \theta)$ for different values of y_2 and of ρ, in the case where $\beta_1 = \beta_2 = 0$.

Using Eqs. (3.21), (3.22) and (3.23), one gets the log-likelihood function:

$$\begin{aligned}\log \mathcal{L}(\theta; \mathbf{y}, \mathbf{X}) &= \sum_{i=1}^{n}(1-y_{1,i}) \log \Phi(-\mathbf{x}_{1,i}'\beta_1) + \\ &\sum_{i=1}^{n} y_{1,i} \log\left(\frac{1}{\sigma}\phi\left(\frac{y_{2,i}-\mathbf{x}_{2,i}'\beta_2}{\sigma}\right)\right) + \\ &\sum_{i=1}^{n} y_{1,i} \log\left(\Phi\left(\frac{\mathbf{x}_{1,i}'\beta_1 + \frac{\rho}{\sigma_2}(y_{2,i}-\mathbf{x}_2'\beta_2)}{\sqrt{1-\rho^2}}\right)\right).\end{aligned}$$

We can also compute conditional expectations:

$$\begin{aligned}\mathbb{E}(y_2^*|y_1=1,\mathbf{x}) &= \mathbb{E}(\mathbb{E}(y_2^*|y_1^*,\mathbf{x})|y_1=1,\mathbf{x}) \\ &= \mathbb{E}(\mathbf{x}_2'\beta_2 + \rho\sigma_2(y_1^*-\mathbf{x}_1'\beta_1)|y_1=1,\mathbf{x}) \\ &= \mathbf{x}_2'\beta_2 + \rho\sigma_2\mathbb{E}(\underbrace{y_1^*-\mathbf{x}_1'\beta_1}_{=\varepsilon_1\sim\mathcal{N}(0,1)}|\varepsilon_1 > -\mathbf{x}_1'\beta_1,\mathbf{x}) \\ &= \mathbf{x}_2'\beta_2 + \rho\sigma_2\frac{\phi(-\mathbf{x}_1'\beta_1)}{1-\Phi(-\mathbf{x}_1'\beta_1)} \\ &= \mathbf{x}_2'\beta_2 + \rho\sigma_2\frac{\phi(\mathbf{x}_1'\beta_1)}{\Phi(\mathbf{x}_1'\beta_1)} = \mathbf{x}_2'\beta_2 + \rho\sigma_2\lambda(\mathbf{x}_1'\beta_1), \quad (3.24)\end{aligned}$$

and:

$$\begin{aligned}\mathbb{E}(y_2^*|y_1=0,\mathbf{x}) &= \mathbf{x}_2'\beta_2 + \rho\sigma_2\mathbb{E}(y_1^*-\mathbf{x}_1'\beta_1|\varepsilon_1 \leq -\mathbf{x}_1'\beta_1,\mathbf{x}) \\ &= \mathbf{x}_2'\beta_2 + \rho\sigma_2\frac{\phi(-\mathbf{x}_1'\beta_1)}{1-\Phi(-\mathbf{x}_1'\beta_1)} \\ &= \mathbf{x}_2'\beta_2 - \rho\sigma_2\frac{\phi(-\mathbf{x}_1'\beta_1)}{\Phi(-\mathbf{x}_1'\beta_1)} = \mathbf{x}_2'\beta_2 - \rho\sigma_2\lambda(-\mathbf{x}_1'\beta_1).\end{aligned}$$

Heckman procedure

As for tobit models (Section 3.3), we can use the Heckman procedure to estimate this model. Eq. (3.24) shows that $\mathbb{E}(y_2^*|y_1=1,\mathbf{x}) \neq \mathbf{x}_2'\beta_2$ when

$\rho \neq 0$. Therefore, the OLS approach yields biased estimates based when it is employed only on the sub-sample where $y_1 = 1$.

The Heckman two-step procedure (or "Heckit") consists in replacing $\lambda(\mathbf{x}_1'\beta_1)$ appearing in Eq. (3.24) with a consistent estimate of it. More precisely:

1. Get an estimate $\widehat{\beta_1}$ of β_1 (probit regression of y_1 on $\mathbf{x}_1$).
2. Run the OLS regression (using only data associated with $y_1 = 1$):

$$y_2 = \mathbf{x}_2'\beta_2 + \rho\sigma_2\lambda(\mathbf{x}_1'\widehat{\beta_1}) + \varepsilon_2, \tag{3.25}$$

considering $\lambda(\mathbf{x}_1'\widehat{\beta_1})$ as a regressor.

How to estimate σ_2^2? By Eq. (4.5), we have:

$$\mathbb{V}ar(y_2|y_1^* > 0, \mathbf{x}) = \mathbb{V}ar(\varepsilon_2|\varepsilon_1 > -\mathbf{x}_1'\beta_1, \mathbf{x}).$$

Using that ε_2 can be decomposed as $\rho\sigma_2\varepsilon_1 + \xi$, where $\xi \sim \mathcal{N}(0, \sigma_2^2(1-\rho^2))$ is independent from ε_1, we get:

$$\mathbb{V}ar(y_2|y_1^* > 0, \mathbf{x}) = \sigma_2^2(1-\rho^2) + \rho^2\sigma_2^2\mathbb{V}ar(\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\beta_1, \mathbf{x}).$$

Using Eq. (4.6), we get:

$$\mathbb{V}ar(\varepsilon_1|\varepsilon_1 > -\mathbf{x}_1'\beta_1, \mathbf{x}) = 1 - \mathbf{x}_1'\beta_1\lambda(\mathbf{x}_1'\beta_1) - \lambda(\mathbf{x}_1'\beta_1)^2,$$

which gives

$$\begin{aligned}\mathbb{V}ar(y_2|y_1^* > 0, \mathbf{x}) &= \sigma_2^2(1-\rho^2) + \rho^2\sigma_2^2(1 - \mathbf{x}_1'\beta_1\lambda(\mathbf{x}_1'\beta_1) - \lambda(\mathbf{x}_1'\beta_1)^2) \\ &= \sigma_2^2 - \rho^2\sigma_2^2\left(\mathbf{x}_1'\beta_1\lambda(\mathbf{x}_1'\beta_1) + \lambda(\mathbf{x}_1'\beta_1)^2\right),\end{aligned}$$

and, finally:

$$\sigma_2^2 \approx \widehat{\mathbb{V}ar}(y_2|y_1^* > 0, \mathbf{x}) + \widehat{\rho\sigma_2}^2\left(\mathbf{x}_1'\widehat{\beta_1}\lambda(\mathbf{x}_1'\widehat{\beta_1}) + \lambda(\mathbf{x}_1'\widehat{\beta_1})^2\right).$$

The Heckman procedure is computationally simple. Although computational costs are no longer an issue, the two-step solution allows certain generalisations more easily than ML, and is more robust in certain circumstances. The computation of parameter standard errors is fairly complicated because of the two steps (see Cameron and Trivedi (2005), Subsection 16.10.2). Bootstrap can be resorted to.

Example 3.11 (Wage prediction)**.** As in Example 3.10, let us use the Mroz (1987) dataset again, with the objective of explaining wage setting.

```
library(sampleSelection)
library(AER)
data("Mroz87")
Mroz87$lfp.yesno <- NaN
Mroz87$lfp.yesno[Mroz87$lfp==1] <- "yes"
Mroz87$lfp.yesno[Mroz87$lfp==0] <- "no"
Mroz87$lfp.yesno <- as.factor(Mroz87$lfp.yesno)
#Logit & Probit (selection equation)
logitW <- glm(lfp ~ age + I( age^2 ) + kids5 + huswage + educ,
              family = binomial(link = "logit"), data = Mroz87)
probitW <- glm(lfp ~ age + I( age^2 ) + kids5 + huswage + educ,
               family = binomial(link = "probit"), data = Mroz87)
# OLS for outcome:
ols1 <- lm(log(wage) ~ educ+exper+I( exper^2 )+city,data=subset(Mroz87,lfp==1))
# Two-step Heckman estimation
heckvan <-
  heckit( lfp ~ age + I( age^2 ) + kids5 + huswage + educ, # selection equation
          log(wage) ~ educ + exper + I( exper^2 ) + city, # outcome equation
          data=Mroz87 )
# Maximun likelihood estimation of selection model:
ml <- selection(lfp~age+I(age^2)+kids5+huswage+educ,
                log(wage)~educ+exper+I(exper^2)+city, data = Mroz87)
# Print selection-equation estimates:
stargazer(logitW,probitW,heckvan,ml,type = "text",no.space = TRUE,
          selection.equation = TRUE)
```

```
##
## ===============================================================
##                                  Dependent variable:
##                     ---------------------------------------------
##                                          lfp
##                     logistic   probit     Heckman      selection
##                                          selection
##                        (1)       (2)         (3)           (4)
```

```
## ------------------------------------------------------------------
## age                    0.012     0.010       0.010         0.010
##                       (0.114)   (0.069)     (0.069)       (0.069)
## I(age2)               -0.001    -0.0005     -0.0005       -0.0005
##                       (0.001)   (0.001)     (0.001)       (0.001)
## kids5                -1.409*** -0.855***   -0.855***     -0.854***
##                       (0.198)   (0.116)     (0.115)       (0.116)
## huswage              -0.069*** -0.042***   -0.042***     -0.042***
##                       (0.020)   (0.012)     (0.012)       (0.013)
## educ                 0.244***  0.148***    0.148***      0.148***
##                       (0.040)   (0.024)     (0.023)       (0.024)
## Constant              -0.938    -0.620      -0.620        -0.615
##                       (2.508)   (1.506)     (1.516)       (1.518)
## ------------------------------------------------------------------
## Observations            753       753         753           753
## R2                                                      0.158
## Adjusted R2                                             0.148
## Log Likelihood       -459.955  -459.901                 -891.177
## Akaike Inf. Crit.     931.910   931.802
## rho                                                     0.018     0.014 (0.203)
## Inverse Mills Ratio                          0.012 (0.152)
## ==================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

```
# Print outcome-equation estimates:
stargazer(ols1,heckvan,ml,type = "text",no.space = TRUE,omit.stat = "f")
```

```
##
## ==========================================================
##                                 Dependent variable:
##                     -----------------------------------------
##                                       log(wage)
##                           OLS           Heckman      selection
##                                        selection
##                           (1)             (2)           (3)
## ----------------------------------------------------------
## educ                    0.106***       0.106***      0.106***
```

```
##                                 (0.014)          (0.017)          (0.017)
## exper                          0.041***         0.041***         0.041***
##                                 (0.013)          (0.013)          (0.013)
## I(exper2)                      -0.001**         -0.001**         -0.001**
##                                 (0.0004)         (0.0004)         (0.0004)
## city                            0.054            0.053            0.053
##                                 (0.068)          (0.069)          (0.069)
## Constant                       -0.531***         -0.547*          -0.544**
##                                 (0.199)          (0.289)          (0.272)
## ------------------------------------------------------------------------
## Observations                      428              753              753
## R2                               0.158            0.158
## Adjusted R2                      0.150            0.148
## Log Likelihood                                                   -891.177
## rho                                               0.018        0.014 (0.203)
## Inverse Mills Ratio                          0.012 (0.152)
## Residual Std. Error 0.667 (df = 423)
## ========================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

3.5 Models of Count Data

Count-data models aim at explaining dependent variables y_i that take integer values. Typically, one may want to account for the number of doctor visits, of customers, of hospital stays, of borrowers' defaults, of recreational trips, of accidents. Quite often, these data feature large proportion of zeros (see, e.g., Table 20.1 in Cameron and Trivedi (2005)), and/or are skewed to the right.

3.5.1 Poisson model

The most basic count-data model is the Poisson model. In this model, we have $y \sim \mathcal{P}(\mu)$, i.e.

$$\mathbb{P}(y = k) = \frac{\mu^k e^{-\mu}}{k!},$$

implying $\mathbb{E}(y) = \mathbb{V}ar(y) = \mu$.

the Poisson parameter, μ, is then assumed to depend on some observed variables, gathered in vector $\mathbf{x}_i$ for entity i. To ensure that $\mu_i \geq 0$, it is common to take $\mu_i = \exp(\beta' \mathbf{x}_i)$, which gives:

$$y_i \sim \mathcal{P}(\exp[\beta' \mathbf{x}_i]).$$

The Poisson regression is intrinsically heteroskedastic (since $\mathbb{V}ar(y_i) = \mu_i = \exp(\beta' \mathbf{x}_i)$).

Under the assumption of independence across entities, the log-likelihood is given by:

$$\log \mathcal{L}(\beta; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} (y_i \beta' \mathbf{x}_i - \exp[\beta' \mathbf{x}_i] - \ln[y_i!]).$$

The first-order condition to get the MLE is:

$$\sum_{i=1}^{n} (y_i - \exp[\beta' \mathbf{x}_i]) \mathbf{x}_i = \underset{K \times 1}{\mathbf{0}}. \tag{3.26}$$

Eq. (3.26) is equivalent to what would define the **Pseudo Maximum-Likelihood** estimator of β in the (misspecified) model

$$y_i \sim i.i.d. \, \mathcal{N}(\exp[\beta' \mathbf{x}_i], \sigma^2).$$

That is, Eq. (3.26) also characterizes the (true) ML estimator of β in the previous model.

Since $\mathbb{E}(y_i | \mathbf{x}_i) = \exp(\beta' \mathbf{x}_i)$, we have:

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

with $\mathbb{E}(\varepsilon_i | \mathbf{x}_i) = 0$. This notably implies that the (N)LS estimator of β is consistent.

How to interpret regression coefficients (the components of β)? We have:

$$\frac{\partial \mathbb{E}(y_i | \mathbf{x}_i)}{\partial x_{i,j}} = \beta_j \exp(\beta' \mathbf{x}_i),$$

which depends on the considered individual.

The average estimated response is:

$$\widehat{\beta}_j \frac{1}{n} \sum_{i=1}^{n} \exp(\widehat{\beta}' \mathbf{x}_i),$$

which is equal to $\widehat{\beta}_j \overline{y}$ if the model includes a constant (e.g., if $x_{1,i} = 1$ for all entities i).

The limitation of the standard Poisson model is that the distribution of y_i conditional on $\mathbf{x}_i$ depends on a single parameter (μ_i). Besides, there is often a tension between fitting the fraction of zeros, i.e. $\mathbb{P}(y_i = 0|\mathbf{x}_i) = \exp[-\exp(\beta' \mathbf{x}_i)]$, and the distribution of $y_i|\mathbf{x}_i, y_i > 0$. The following models (negative binomial, or NB model, the Hurdle model, and the Zero-Inflated model) have been designed to address these points.

3.5.2 Negative binomial model

In the negative binomial model, we have:

$$y_i|\lambda_i \sim \mathcal{P}(\lambda_i),$$

but λ_i is now random. Specifically, it takes the form:

$$\lambda_i = \nu_i \times \exp(\beta' \mathbf{x}_i),$$

where $\nu_i \sim i.i.d.\, \Gamma(\underbrace{\delta}_{\text{shape}}, \underbrace{1/\delta}_{\text{scale}})$. That is, the p.d.f. of ν_i is:

$$g(\nu) = \frac{\nu^{\delta-1} e^{-\nu\delta} \delta^{\delta}}{\Gamma(\delta)},$$

where $\Gamma : z \mapsto \int_0^{+\infty} t^{z-1} e^{-t} dt$ (and $\Gamma(k+1) = k!$).

This notably implies that:

$$\mathbb{E}(\nu_i) = 1 \quad \text{and} \quad \mathbb{V}ar(\nu) = \frac{1}{\delta}.$$

Hence, the p.d.f. of y_i conditional on μ and δ (with $\mu = \exp(\beta' \mathbf{x}_i)$) is obtained as a mixture of densities:

$$\mathbb{P}(y_i = k|\exp(\beta' \mathbf{x}_i) = \mu; \delta) = \int_0^{\infty} \frac{e^{-\mu\nu}(\mu\nu)^k}{k!} \frac{\nu^{\delta-1} e^{-\nu\delta} \delta^{\delta}}{\Gamma(\delta)} d\nu.$$

It can be shown that:

$$\mathbb{E}(y|\mathbf{x}) = \mu \quad \text{and} \quad \mathbb{V}ar(y|\mathbf{x}) = \mu\left(1 + \alpha\mu\right),$$

where $\exp(\beta'\mathbf{x}_i) = \mu$ and $\alpha = 1/\delta$.

We have one additional degree of freedom w.r.t. the Poisson model (α).

Note that $\mathbb{V}ar(y|\mathbf{x}) > \mathbb{E}(y|\mathbf{x})$ (which is often called for by the data). Moreover, the conditional variance is quadratic in the mean:

$$\mathbb{V}ar(y|\mathbf{x}) = \mu + \alpha\mu^2.$$

The previous expression is the basis of the so-called **NB2** specification. If δ is replaced with μ/γ, then we get the **NB1** model:

$$\mathbb{V}ar(y|\mathbf{x}) = \mu(1 + \gamma).$$

Example 3.12 (Number of doctor visits)**.** The following example compares different specifications, namely a linear regression model, a Poisson model, and a NB model, to account for the number of doctor visits. The dataset (‘`randdata`) is the one used in Chapter 20 of Cameron and Trivedi (2005) (available on that page).

```
library(AEC)
library(COUNT)
library(pscl) # for predprob function and hurdle model
par(plt=c(.15,.95,.1,.95))
plot(table(randdata$mdvis))
```

```
randdata$LC <- log(1 + randdata$coins)
model.OLS <- lm(mdvis ~ LC + idp + lpi + fmde + physlm + disea + hlthg +
                  hlthf + hlthp - 1,data=randdata)
model.poisson <- glm(mdvis ~ LC + idp + lpi + fmde + physlm + disea +
                       hlthg + hlthf + hlthp - 1,data=randdata,family = poisso
model.neg.bin <- glm.nb(mdvis ~ LC + idp + lpi + fmde + physlm + disea +
                          hlthg + hlthf + hlthp - 1,data=randdata)
model.neg.bin.with.intercept <-
  glm.nb(mdvis ~ LC + idp + lpi + fmde + physlm + disea + hlthg +
```

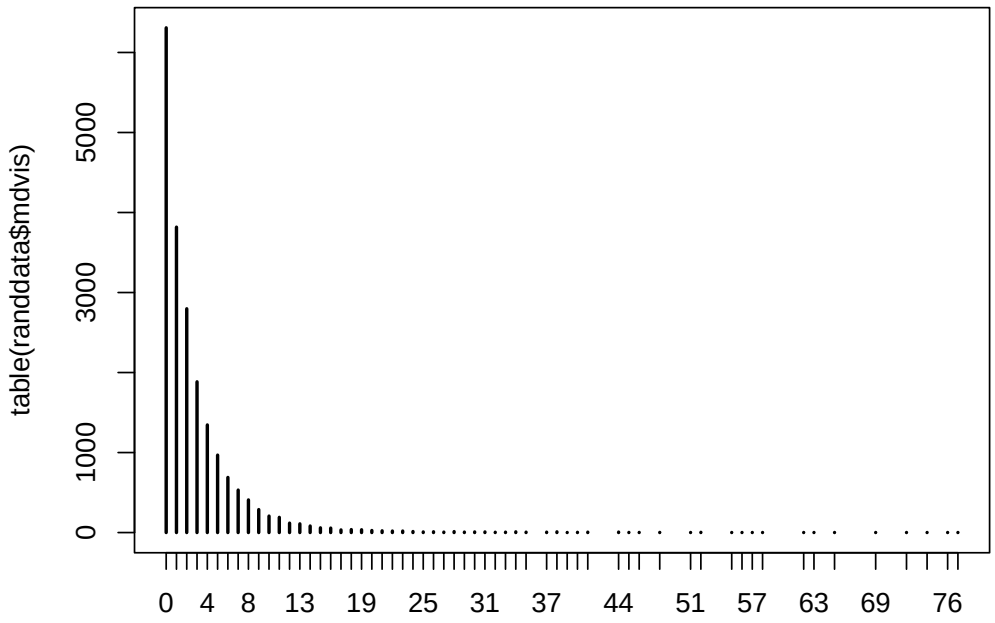


Figure 3.13: Distribution of the number of doctor visits.

```
           hlthf + hlthp,data=randdata)
stargazer::stargazer(model.OLS,model.poisson,model.neg.bin,
                     model.neg.bin.with.intercept,type="text",
                     no.space = TRUE,omit.stat = c("f","ser"))
```

```
##
## ================================================================================
##                                        Dependent variable:
##                    ------------------------------------------------------------
##                                              mdvis
##                       OLS      Poisson                negative
##                                                       binomial
##                       (1)        (2)            (3)              (4)
## --------------------------------------------------------------------------------
## LC                 -0.155***  -0.051***      -0.057***        -0.058***
##                     (0.020)    (0.003)        (0.006)          (0.006)
## idp                -0.546***  -0.183***      -0.212***        -0.268***
##                     (0.075)    (0.011)        (0.023)          (0.023)
```

```
## lpi                0.230***   0.095***       0.088***         0.041***
##                     (0.012)    (0.002)       (0.004)          (0.004)
## fmde               -0.073***  -0.029***     -0.030***        -0.038***
##                     (0.012)    (0.002)       (0.004)          (0.003)
## physlm             0.945***   0.217***       0.229***         0.269***
##                     (0.104)    (0.013)       (0.031)          (0.030)
## disea              0.177***   0.050***       0.062***         0.038***
##                     (0.004)   (0.0005)       (0.001)          (0.001)
## hlthg              0.270***   0.126***       0.068***         -0.044**
##                     (0.066)    (0.009)       (0.020)          (0.020)
## hlthf              0.455***   0.149***       0.084**           0.017
##                     (0.123)    (0.016)       (0.037)          (0.036)
## hlthp              1.537***   0.197***       0.185**          0.178**
##                     (0.263)    (0.027)       (0.076)          (0.074)
## Constant                                                     0.664***
##                                                               (0.025)
## ---------------------------------------------------------------------------
## Observations       20,190     20,190         20,190           20,190
## R2                  0.322
## Adjusted R2         0.322
## Log Likelihood              -64,221.340    -43,745.860      -43,384.660
## theta                                   0.732*** (0.010) 0.773*** (0.011)
## Akaike Inf. Crit.           128,460.700    87,509.710       86,789.320
## ===========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Models' predictions can be obtained as follows:

```
# prediction of beta'x, equivalent to "model.poisson$fitted.values":
predict_poisson.beta.x <- predict(model.poisson)
# prediction of the number of events (exp(beta'x)):
predict_poisson <- predict(model.poisson,type="response")
predict_NB <- model.neg.bin$fitted.values
```

Let us now compute the model-implied probabilities, and let's compare them with the frequencies observed in the data.

```
prop.table.data  <- prop.table(table(randdata$mdvis))
predprob.poisson <- predprob(model.poisson) # part of pscl package
predprob.nb      <- predprob(model.neg.bin)
print(rbind(prop.table.data[1:6],
            apply(predprob.poisson[,1:6],2,mean),
            apply(predprob.nb[,1:6],2,mean)))
```

```
##              0         1         2          3          4          5
## [1,] 0.3124319 0.1890540 0.1385339 0.09331352 0.06661714 0.04794453
## [2,] 0.1220592 0.2230328 0.2271621 0.17173478 0.10888940 0.06244353
## [3,] 0.3486824 0.1884640 0.1219340 0.08385899 0.05968603 0.04350209
```

It appears that the NB model is better at capturing the relatively large number of zeros than the Poisson model. This will also be the case for the Hurdle and Zero-Inflation models:

3.5.3 Hurdle model

The main objective of this model, w.r.t. the Poisson model, is to generate more zeros in the data than predicted by the previous count models. The idea is to separate the modeling of the number of zeros and that of the number of non-zero counts. Specifically, the frequency of zeros is determined by f_1, the (relative) frequencies of non-zero counts are determined by f_2:

$$f(y) = \begin{cases} f_1(0) & \text{if } y = 0, \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y) & \text{if } y > 0. \end{cases}$$

Note that we are back to the standard Poisson model if $f_1 \equiv f_2$. This model is straightforwardly estimated by ML.

3.5.4 Zero-inflated model

The objective is the same as for the Hurdle model, the modeling is slightly different. It is based on a mixture of a binary process B (p.d.f. f_1) and a

process Z (p.d.f. f_2). B and Z are independent. Formally:

$$y = BZ,$$

implying:

$$f(y) = \begin{cases} f_1(0) + (1 - f_1(0))f_2(0) & \text{if } y = 0, \\ (1 - f_1(0))f_2(y) & \text{if } y > 0. \end{cases}$$

Typically, f_1 corresponds to a logit model and f_2 is Poisson or negative binomial density. This model is easily estimated by ML techniques.

Example 3.13 (Number of doctor visits)**.** Let us come back to the data used in Example 3.12, and estimate Hurdle and a zero-inflation models:

```
model.hurdle <-
  hurdle(mdvis ~ LC + idp + lpi + fmde + physlm + disea + hlthg + hlthf +
           hlthp, data=randdata,
         dist = "poisson", zero.dist = "binomial", link = "logit")
model.zeroinfl <- zeroinfl(mdvis ~ LC + idp + lpi + fmde + physlm +
                             disea + hlthg + hlthf + hlthp, data=randdata,
                           dist = "poisson", link = "logit")
stargazer(model.hurdle,model.zeroinfl,zero.component=FALSE,
          no.space=TRUE,type="text")
```

```
##
## ============================================
##                     Dependent variable:
##                 ---------------------------
##                            mdvis
##                     hurdle     zero-inflated
##                                 count data
##                       (1)           (2)
## --------------------------------------------
## LC                -0.015***      -0.015***
##                    (0.003)        (0.003)
## idp               -0.085***      -0.086***
##                    (0.011)        (0.011)
## lpi                0.010***       0.010***
##                    (0.002)        (0.002)
```

```
## fmde              -0.021***      -0.021***
##                    (0.002)        (0.002)
## physlm             0.231***       0.231***
##                    (0.012)        (0.012)
## disea              0.022***       0.022***
##                    (0.001)        (0.001)
## hlthg              0.027***       0.026***
##                    (0.010)        (0.010)
## hlthf              0.147***       0.146***
##                    (0.016)        (0.016)
## hlthp              0.304***       0.303***
##                    (0.026)        (0.026)
## Constant           1.133***       1.133***
##                    (0.012)        (0.012)
## ------------------------------------------
## Observations       20,190         20,190
## Log Likelihood  -54,772.100    -54,772.550
## ==========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(model.hurdle,model.zeroinfl,zero.component=TRUE,
          no.space=TRUE,type="text")
```

```
##
## ==========================================
##                    Dependent variable:
##                ---------------------------
##                           mdvis
##                    hurdle     zero-inflated
##                                 count data
##                      (1)            (2)
## ------------------------------------------
## LC                -0.150***      0.154***
##                    (0.010)        (0.011)
## idp               -0.631***      0.637***
##                    (0.038)        (0.040)
## lpi                0.102***      -0.105***
```

```
##                    (0.007)         (0.007)
## fmde              -0.062***       0.060***
##                    (0.006)         (0.006)
## physlm             0.239***       -0.203***
##                    (0.056)         (0.058)
## disea              0.062***       -0.059***
##                    (0.003)         (0.003)
## hlthg             -0.142***       0.158***
##                    (0.034)         (0.036)
## hlthf             -0.352***       0.396***
##                    (0.062)         (0.064)
## hlthp               -0.181          0.233
##                    (0.149)         (0.151)
## Constant           0.411***       -0.528***
##                    (0.044)         (0.047)
## ------------------------------------------
## Observations        20,190          20,190
## Log Likelihood  -54,772.100     -54,772.550
## ==========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

Let us test the importance of `LC` in the model using a Wald test:

```
# Test whether LC is important in the model:
model.hurdle.reduced <- update(model.hurdle,.~.-LC)
lmtest::waldtest(model.hurdle, model.hurdle.reduced)
```

```
## Wald test
##
## Model 1: mdvis ~ LC + idp + lpi + fmde + physlm + disea + hlthg + hlthf +
##     hlthp
## Model 2: mdvis ~ idp + lpi + fmde + physlm + disea + hlthg + hlthf + hlthp
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  20170
## 2  20172 -2 247.64  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we compare average model-implied probabilities with the frequencies observed in the data:

```
predprob.hurdle   <- predprob(model.hurdle)
predprob.zeroinfl <- predprob(model.zeroinfl)
print(rbind(prop.table.data[1:6],
  apply(predprob.poisson[,1:6],2,mean),
  apply(predprob.nb[,1:6],2,mean),
  apply(predprob.hurdle[,1:6],2,mean),
  apply(predprob.zeroinfl[,1:6],2,mean)))
```

```
##              0          1         2          3          4          5
## [1,] 0.3124319 0.18905399 0.1385339 0.09331352 0.06661714 0.04794453
## [2,] 0.1220592 0.22303277 0.2271621 0.17173478 0.10888940 0.06244353
## [3,] 0.3486824 0.18846395 0.1219340 0.08385899 0.05968603 0.04350209
## [4,] 0.3124319 0.06056959 0.1083120 0.13262624 0.12553899 0.09847017
## [5,] 0.3124684 0.06053026 0.1082799 0.13262562 0.12556531 0.09850218
```

Chapter 4

Appendix

4.1 Principal component analysis (PCA)

Principal component analysis (PCA) is a classical and easy-to-use statistical method to reduce the dimension of large datasets containing variables that are linearly driven by a relatively small number of factors. This approach is widely used in data analysis and image compression.

Suppose that we have T observations of a n-dimensional random vector x, denoted by $x_1, x_2, \ldots, x_T$. We suppose that each component of x is of mean zero.

Denote with X the matrix given by $\begin{bmatrix} x_1 & x_2 & \ldots & x_T \end{bmatrix}'$. Denote the j^{th} column of X by X_j.

We want to find the linear combination of the x_i's $(x.u)$, with $\|u\| = 1$, with "maximum variance." That is, we want to solve:

$$\begin{array}{ll} \arg\max\limits_{u} & u'X'Xu. \\ \text{s.t.} & |u| = 1 \end{array} \tag{4.1}$$

Since $X'X$ is a positive definite matrix, it admits the following decomposi-

tion:

$$\begin{aligned} X'X &= PDP' \\ &= P \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} P', \end{aligned}$$

where P is an orthogonal matrix whose columns are the eigenvectors of $X'X$.

We can order the eigenvalues such that $\lambda_1 \geq ... \geq \lambda_n$. (Since $X'X$ is positive definite, all these eigenvalues are positive.)

Since P is orthogonal, we have $u'X'Xu = u'PDP'u = y'Dy$ where $\|y\| = 1$. Therefore, we have $y_i^2 \leq 1$ for any $i \leq n$.

As a consequence:

$$y'Dy = \sum_{i=1}^{n} y_i^2 \lambda_i \leq \lambda_1 \sum_{i=1}^{n} y_i^2 = \lambda_1.$$

It is easily seen that the maximum is reached for $y = [1, 0, \cdots, 0]'$. Therefore, the maximum of the optimization program (Eq. (4.1)) is obtained for $u = P[1, 0, \cdots, 0]'$. That is, u is the eigenvector of $X'X$ that is associated with its larger eigenvalue (first column of P).

Let us denote with F the vector that is given by the matrix product XP (note that its last column is equal to Xu). The columns of F, denoted by F_j, are called **factors**. We have:

$$F'F = P'X'XP = D.$$

Therefore, in particular, the F_j's are orthogonal.

Since $X = FP'$, the X_j's are linear combinations of the factors. Let us then denote with $\hat{X}_{i,j}$ the part of X_i that is explained by factor F_j, we have:

$$\begin{aligned} \hat{X}_{i,j} &= p_{ij} F_j \\ X_i &= \sum_j \hat{X}_{i,j} = \sum_j p_{ij} F_j. \end{aligned}$$

Consider the share of variance that is explained –through the n variables $(X_1, ..., X_n)$– by the first factor F_1:

$$\frac{\sum_i \hat{X}_{i,1}\hat{X}'_{i,1}}{\sum_i X_i X'_i} = \frac{\sum_i p_{i1}F_1F'_1p_{i1}}{tr(X'X)} = \frac{\sum_i p_{i1}^2\lambda_1}{tr(X'X)} = \frac{\lambda_1}{\sum_i \lambda_i}.$$

Intuitively, if the first eigenvalue is large, it means that the first factor embed a large share of the fluctutaions of the n X_i's.

Let us illustrate PCA on the term structure of yields. The term strucutre of yields (or yield curve) is know to be driven by only a small number of factors (e.g., Litterman and Scheinkman (1991)). One can typically employ PCA to recover such factors. The data used in the example below are taken from the Fred database (tickers: "DGS6MO","DGS1", ...). The second plot shows the factor loardings, that indicate that the first factor is a level factor (loadings = black line), the second factor is a slope factor (loadings = blue line), the third factor is a curvature factor (loadings = red line).

To run a PCA, one simply has to apply function `prcomp` to a matrix of data:

```
library(AEC)
USyields <- USyields[complete.cases(USyields),]
yds <- USyields[c("Y1","Y2","Y3","Y5","Y7","Y10","Y20","Y30")]
PCA.yds <- prcomp(yds,center=TRUE,scale. = TRUE)
```

Let us know visualize some results. The first plot of Figure 4.1 shows the share of total variance explained by the different principal components (PCs). The second plot shows the facotr loadings. The two bottom plots show how yields (in black) are fitted by linear combinations of the first two PCs only.

```
par(mfrow=c(2,2))
par(plt=c(.1,.95,.2,.8))
barplot(PCA.yds$sdev^2/sum(PCA.yds$sdev^2),
        main="Share of variance expl. by PC's")
axis(1, at=1:dim(yds)[2], labels=colnames(PCA.yds$x))
nb.PC <- 2
plot(-PCA.yds$rotation[,1],type="l",lwd=2,ylim=c(-1,1),
     main="Factor loadings (1st 3 PCs)",xaxt="n",xlab="")
```

```
axis(1, at=1:dim(yds)[2], labels=colnames(yds))
lines(PCA.yds$rotation[,2],type="l",lwd=2,col="blue")
lines(PCA.yds$rotation[,3],type="l",lwd=2,col="red")
Y1.hat <- PCA.yds$x[,1:nb.PC] %*% PCA.yds$rotation["Y1",1:2]
Y1.hat <- mean(USyields$Y1) + sd(USyields$Y1) * Y1.hat
plot(USyields$date,USyields$Y1,type="l",lwd=2,
     main="Fit of 1-year yields (2 PCs)",
     ylab="Obs (black) / Fitted by 2PCs (dashed blue)")
lines(USyields$date,Y1.hat,col="blue",lty=2,lwd=2)
Y10.hat <- PCA.yds$x[,1:nb.PC] %*% PCA.yds$rotation["Y10",1:2]
Y10.hat <- mean(USyields$Y10) + sd(USyields$Y10) * Y10.hat
plot(USyields$date,USyields$Y10,type="l",lwd=2,
     main="Fit of 10-year yields (2 PCs)",
     ylab="Obs (black) / Fitted by 2PCs (dashed blue)")
lines(USyields$date,Y10.hat,col="blue",lty=2,lwd=2)
```

4.2 Linear algebra: definitions and results

Definition 4.1 (Eigenvalues)**.** The eigenvalues of of a matrix M are the numbers λ for which:

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

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

Proposition 4.1 (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 4.2 (Moore-Penrose inverse)**.** If $M \in \mathbb{R}^{m \times n}$, then its Moore-Penrose pseudo inverse (exists and) is the unique matrix $M^* \in \mathbb{R}^{n \times m}$ that satisfies:

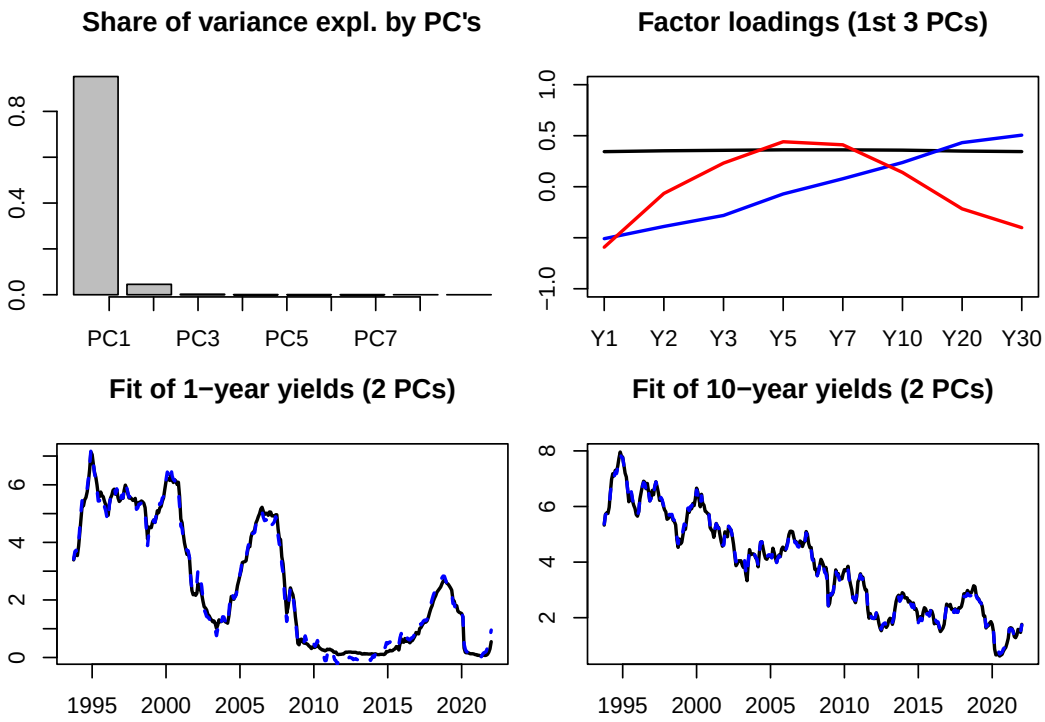


Figure 4.1: Some PCA results. The dataset contains 8 time series of U.S. interest rates of different maturities.

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

Proposition 4.2 (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 4.3 (Idempotent matrix)**.** Matrix M is idempotent if $M^2 = M$.

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

Proposition 4.3 (Roots of an idempotent matrix)**.** *The eigenvalues of an idempotent matrix are either 1 or 0.*

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

Proposition 4.4 (Idempotent matrix and chi-square distribution)**.** *The rank of a symmetric idempotent matrix is equal to its trace.*

Proof. The result follows from Prop. 4.3, combined with the fact that the rank of a symmetric matrix is equal to the number of its nonzero eigenvalues. □

Proposition 4.5 (Constrained least squares)**.** *The solution of the following optimisation problem:*

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

is given by:

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

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

Proof. See for instance Jackman, 2007. □

Proposition 4.6 (Inverse of a partitioned matrix)**.** *We have:*

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix}.$$

Definition 4.4 (Matrix derivatives)**.** Consider a fonction $f : \mathbb{R}^K \to \mathbb{R}$. Its first-order derivative is:

$$\frac{\partial f}{\partial \mathbf{b}}(\mathbf{b}) = \begin{bmatrix} \frac{\partial f}{\partial b_1}(\mathbf{b}) \\ \vdots \\ \frac{\partial f}{\partial b_K}(\mathbf{b}) \end{bmatrix}.$$

We use the notation:

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

Proposition 4.7. *We have:*

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

Proposition 4.8 (Square and absolute summability)**.** *We have:*

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

Proof. See Appendix 3.A in Hamilton. Idea: Absolute summability implies that there exist N such that, for $j > N$, $|\theta_j| < 1$ (deduced from Cauchy criterion, Theorem 4.2 and therefore $\theta_j^2 < |\theta_j|$. □

4.3 Statistical analysis: definitions and results

4.3.1 Moments and statistics

Definition 4.5 (Partial correlation)**.** The **partial correlation** between y and z, controlling for some variables $\mathbf{X}$ is the sample correlation between y^* and z^*, where the latter two variables are the residuals in regressions of y on $\mathbf{X}$ and of z on $\mathbf{X}$, respectively.

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

$$r_{yz}^{\mathbf{X}} = \frac{\mathbf{z}^{*\prime}\mathbf{y}^*}{\sqrt{(\mathbf{z}^{*\prime}\mathbf{z}^*)(\mathbf{y}^{*\prime}\mathbf{y}^*)}}. \tag{4.2}$$

Definition 4.6 (Skewness and kurtosis)**.** Let Y be a random variable whose fourth moment exists. The expectation of Y is denoted by μ.

- The skewness of Y is given by:

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

- The kurtosis of Y is given by:

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

Theorem 4.1 (Cauchy-Schwarz inequality)**.** *We have:*

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

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

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

variance of $Y = Z + \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$ is equal to the sum of the variance of Z and of the variance of $\frac{\mathbb{C}ov(X,Y)}{\mathbb{V}ar(X)}X$, that is:

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

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

Definition 4.7 (Asymptotic level)**.** An asymptotic test with critical region Ω_n has an asymptotic level equal to α if:

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

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

Definition 4.8 (Asymptotically consistent test)**.** An asymptotic test with critical region Ω_n is consistent if:

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

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

Definition 4.9 (Kullback discrepancy)**.** Given two p.d.f. f and f^*, the Kullback discrepancy is defined by:

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

Proposition 4.9 (Properties of the Kullback discrepancy)**.** *We have:*

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

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

Definition 4.10 (Characteristic function)**.** For any real-valued random variable X, the characteristic function is defined by:

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

4.3.2 Standard distributions

Definition 4.11 (F distribution)**.** Consider $n = n_1 + n_2$ i.i.d. $\mathcal{N}(0,1)$ r.v. X_i. If the r.v. F is defined by:

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

then $F \sim \mathcal{F}(n_1, n_2)$. (See Table 4.4 for quantiles.)

Definition 4.12 (Student-t distribution)**.** Z follows a Student-t (or t) distribution with ν degrees of freedom (d.f.) if:

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

We have $\mathbb{E}(Z) = 0$, and $\mathbb{V}ar(Z) = \frac{\nu}{\nu-2}$ if $\nu > 2$. (See Table 4.2 for quantiles.)

Definition 4.13 (Chi-square distribution)**.** Z follows a χ^2 distribution with ν d.f. if $Z = \sum_{i=1}^{\nu} X_i^2$ where $X_i \sim i.i.d. \mathcal{N}(0,1)$. We have $\mathbb{E}(Z) = \nu$. (See Table 4.3 for quantiles.)

Definition 4.14 (Cauchy distribution)**.** The probability distribution function of the Cauchy distribution defined by a location parameter μ and a scale parameter γ is:

$$f(x) = \frac{1}{\pi\gamma \left(1 + \left[\frac{x-\mu}{\gamma}\right]^2\right)}.$$

The mean and variance of this distribution are undefined.

Proposition 4.10 (Inner product of a multivariate Gaussian variable)**.** *Let X be a n-dimensional multivariate Gaussian variable: $X \sim \mathcal{N}(0, \Sigma)$. We have:*

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

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

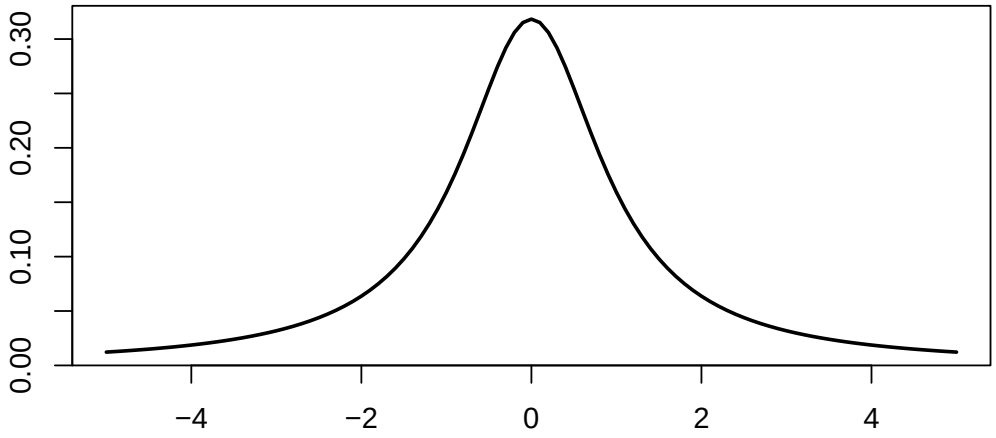


Figure 4.2: Pdf of the Cauchy distribution ($\mu = 0$, $\gamma = 1$).

easily checked that the covariance matrix of $Y := \sqrt{D^{-1}}P'X$ is Id. Therefore Y is a vector of uncorrelated Gaussian variables. The properties of Gaussian variables imply that the components of Y are then also independent. Hence $Y'Y = \sum_i Y_i^2 \sim \chi^2(n)$.

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

Definition 4.15 (Generalized Extreme Value (GEV) distribution)**.** The vector of disturbances $\varepsilon = [\varepsilon_{1,1}, \dots, \varepsilon_{1,K_1}, \dots, \varepsilon_{J,1}, \dots, \varepsilon_{J,K_J}]'$ follows the Generalized Extreme Value (GEV) distribution if its c.d.f. is:

$$F(\varepsilon, \rho) = \exp(-G(e^{-\varepsilon_{1,1}}, \dots, e^{-\varepsilon_{J,K_J}}; \rho))$$

with

$$\begin{aligned} G(\mathbf{Y}; \rho) &\equiv G(Y_{1,1}, \dots, Y_{1,K_1}, \dots, Y_{J,1}, \dots, Y_{J,K_J}; \rho) \\ &= \sum_{j=1}^{J} \left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j} \right)^{\rho_j} \end{aligned}$$

4.3.3 Stochastic convergences

Proposition 4.11 (Chebychev's inequality)**.** *If* $\mathbb{E}(|X|^r)$ *is finite for some* $r > 0$ *then:*

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

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

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

Proof. Remark that $\varepsilon^r \mathbb{1}_{\{|X| \geq \varepsilon\}} \leq |X|^r$ and take the expectation of both sides. □

Definition 4.16 (Convergence in probability)**.** The random variable sequence x_n converges in probability to a constant c if $\forall \varepsilon$, $\lim_{n \to \infty} \mathbb{P}(|x_n - c| > \varepsilon) = 0$.

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

Definition 4.17 (Convergence in the Lr norm)**.** x_n converges in the r-th mean (or in the L^r-norm) towards x, if $\mathbb{E}(|x_n|^r)$ and $\mathbb{E}(|x|^r)$ exist and if

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

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

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

Definition 4.18 (Almost sure convergence)**.** The random variable sequence x_n converges almost surely to c if $\mathbb{P}(\lim_{n \to \infty} x_n = c) = 1$.

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

Definition 4.19 (Convergence in distribution)**.** x_n is said to converge in distribution (or in law) to x if

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

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

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

Proposition 4.12 (Rules for limiting distributions (Slutsky))**.** *We have:*

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

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

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

Proposition 4.13 (Implications of stochastic convergences)**.** *We have:*

$$\boxed{\overset{L^s}{\to}} \underset{1\leq r\leq s}{\Rightarrow} \boxed{\overset{L^r}{\to}}$$

$$\Downarrow$$

$$\boxed{\overset{a.s.}{\to}} \Rightarrow \boxed{\overset{p}{\to}} \Rightarrow \boxed{\overset{d}{\to}}.$$

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

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

Besides,

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

which implies:

$$F(x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \leq F_n(x). \tag{4.3}$$

Eqs. (4.3) and (4.3) 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 4.14 (Convergence in distribution to a constant)**.** *If X_n converges in distribution to a constant c, then X_n converges in probability to c.*

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

Example 4.1 (Convergence in probability but not L^r)**.** Let $\{x_n\}_{n\in\mathbb{N}}$ be a series of random variables defined by:

$$x_n = nu_n,$$

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

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

Theorem 4.2 (Cauchy criterion (non-stochastic case))**.** *We have that $\sum_{i=0}^{T} a_i$ converges ($T \to \infty$) iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

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

Theorem 4.3 (Cauchy criterion (stochastic case))**.** *We have that $\sum_{i=0}^{T} \theta_i \varepsilon_{t-i}$ converges in mean square ($T \to \infty$) to a random variable iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

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

4.3.4 Central limit theorem

Theorem 4.4 (Law of large numbers)**.** *The sample mean is a consistent estimator of the population mean.*

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

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

The properties of the characteristic function (see Def. 4.10) imply that:

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

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

Theorem 4.5 (Lindberg-Levy Central limit theorem, CLT)**.** *If x_n is an i.i.d. sequence of random variables with mean μ and variance σ^2 ($\in]0, +\infty[$), then:*

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

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

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

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

4.4 Some properties of Gaussian variables

Proposition 4.15. *If* $\mathbf{A}$ *is idempotent and if* $\mathbf{x}$ *is Gaussian,* $\mathbf{Lx}$ *and* $\mathbf{x}'\mathbf{Ax}$ *are independent if* $\mathbf{LA} = \mathbf{0}$.

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

Proposition 4.16 (Bayesian update in a vector of Gaussian variables)**.** *If*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}\right),$$

then

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

Proposition 4.17 (Truncated distributions)**.** *If* X *is a random variable distributed according to some p.d.f.* f*, with c.d.f.* F*, with infinite support. Then the p.d.f. of* $X|a \leq X < b$ *is*

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

for any $a < b$.

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

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

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

Moreover:

$$\mathbb{E}(X|a \leq X < b) \quad = \quad \mu - \frac{\phi(\beta) - \phi(\alpha)}{Z}\sigma. \tag{4.4}$$

We also have:

$$\begin{aligned} & \mathbb{V}ar(X|a \leq X < b) \\ = \; & \sigma^2 \left[1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{Z} - \left(\frac{\phi(\beta) - \phi(\alpha)}{Z} \right)^2 \right] \end{aligned} \tag{4.5}$$

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

$$\mathbb{V}ar(X|a < X) = \sigma^2 \left[1 + \alpha\lambda(-\alpha) - \lambda(-\alpha)^2 \right], \tag{4.6}$$

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

Consider the case where $a \to -\infty$ (i.e. the conditioning set is $X < b$) and $\mu = 0$, $\sigma = 1$. Then Eq. (4.4) gives $\mathbb{E}(X|X < b) = -\lambda(b) = -\dfrac{\phi(b)}{\Phi(b)}$, where λ is the function computing the inverse Mills ratio.

Proposition 4.18 (p.d.f. of a multivariate Gaussian variable)**.** *If $Y \sim \mathcal{N}(\mu, \Omega)$ and if Y is a n-dimensional vector, then the density function of Y is:*

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

4.5 Proofs

Proof of Proposition 2.4

Proof. Assumptions (i) and (ii) (in the set of Assumptions 2.1) imply that θ_{MLE} exists ($= \text{argmax}_\theta (1/n) \log \mathcal{L}(\theta; \mathbf{y})$).

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

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

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

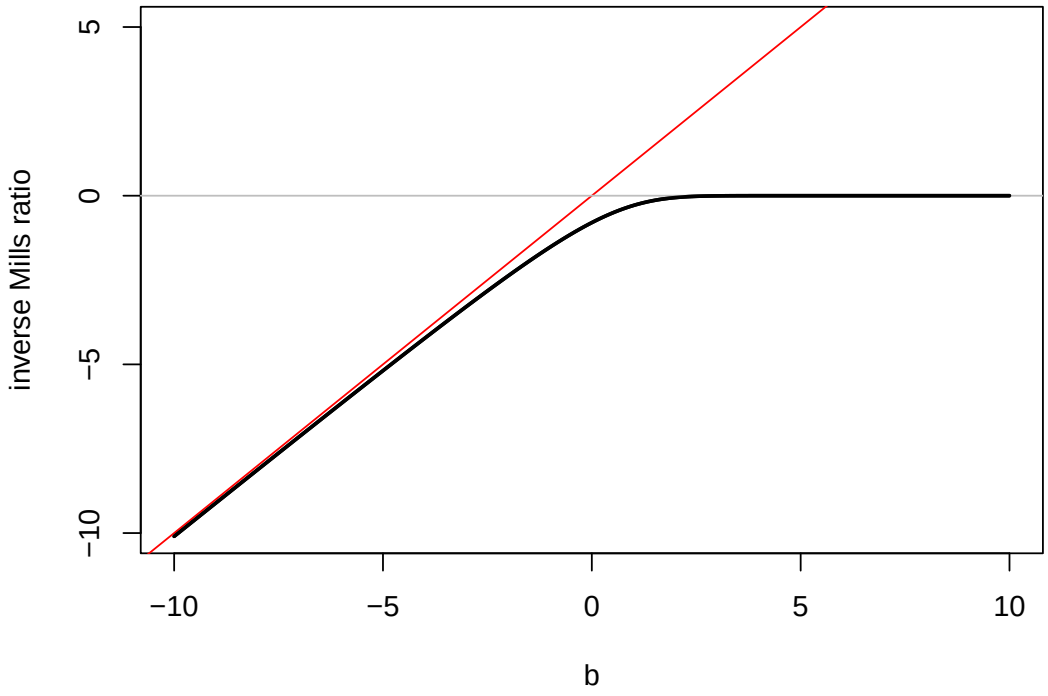


Figure 4.3: $\mathbb{E}(X|X < b)$ as a function of b when $X \sim \mathcal{N}(0, 1)$ (in black).

Properties of the Kullback information measure (see Prop. 4.9), together with the identifiability assumption (ii) implies that the solution to the limit problem is unique and equal to θ_0.

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

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

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

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

or equivalently:

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

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

Besides, we have:

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

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

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

Proof of Proposition 2.5

Proof. We have $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$ (Eq. (2.10)). A Taylor expansion around θ_0 yields to:

$$\sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{4.7}$$

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

$$\sqrt{n}h(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{4.8}$$

Hence

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

Taking the quadratic form, we obtain:

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

The fact that the test has asymptotic level α directly stems from what precedes. **Consistency of the test**: Consider $\theta_0 \in \Theta$. Because the MLE is consistent, $h(\hat{\theta}_n)$ converges to $h(\theta_0) \neq 0$. Eq. (4.7) is still valid. It implies that ξ_n^W converges to $+\infty$ and therefore that $\mathbb{P}_\theta(\xi_n^W \geq \chi^2_{1-\alpha}(r)) \to 1$. □

Proof of Proposition 2.6

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

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

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{4.9}$$

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

$$\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \tag{4.10}$$

and:

$$0 = \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n;\mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0). \tag{4.11}$$

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

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \approx \mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}\mathcal{I}(\theta_0). \tag{4.12}$$

Eqs. (4.9) and (4.12) yield to:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}. \tag{4.13}$$

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

$$\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta} + \frac{\partial h'(\hat{\theta}_n^0;\mathbf{y})}{\partial \theta}\hat{\lambda}_n = 0. \tag{4.14}$$

Substituting the latter equation in Eq. (4.13) gives:

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

which yields:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \approx -\left(\frac{\partial h(\theta_0)}{\partial \theta'}\mathcal{I}(\theta_0)^{-1}\frac{\partial h'(\theta_0;\mathbf{y})}{\partial \theta}\right)^{-1}\sqrt{n}h(\hat{\theta}_n). \tag{4.15}$$

It follows, from Eq. (4.8), that:

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

Taking the quadratic form of the last equation gives:

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

Using Eq. (4.14), it appears that the left-hand side term of the last equation is ξ^{LM} as defined in Eq. (2.18). Consistency: see Remark 17.3 in Gouriéroux and Monfort (1995). □

Proof of Proposition 2.7

Proof. Let us first demonstrate the asymptotic equivalence of ξ^{LM} and ξ^{LR}.

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

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

Taking the difference, we obtain:

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

Using $\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$ (Eq. (4.11)), we have:

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

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

$$\xi_n^{LR} \approx n(\hat{\theta}_n^0 - \hat{\theta}_n)' \mathcal{I}(\theta_0)^{-1} (\hat{\theta}_n^0 - \hat{\theta}_n). \tag{4.16}$$

The difference between Eqs. (4.10) and (4.11) implies:

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

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

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

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

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

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

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

Let us now demonstrate the equivalence of $\xi^{LM} and \xi^W$.

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

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

Since, under H_0, $\hat{\theta}_n^0 \approx \hat{\theta}_n \approx \theta_0$, Eq. (4.15) 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. □

Proof of Eq. (??)

Proof. We have:

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

Therefore:

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

And then:

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

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

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

Consider $\varepsilon > 0$. The fact that the autocovariances are absolutely summable implies that there exists q_0 such that (Cauchy criterion, Theorem 4.2):

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

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

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

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

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

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

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

□

Proof of Proposition ??

Proof. We have:

$$\begin{aligned}
\mathbb{E}([y_{t+1} - y^*_{t+1}]^2) &= \mathbb{E}\left([\{y_{t+1} - \mathbb{E}(y_{t+1}|x_t)\} + \{\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}\}]^2\right) \\
&= \mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)]^2\right) + \mathbb{E}\left([\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]^2\right) \\
&\quad +2\mathbb{E}\left([y_{t+1} - \mathbb{E}(y_{t+1}|x_t)][\mathbb{E}(y_{t+1}|x_t) - y^*_{t+1}]\right). \quad (4.17)
\end{aligned}$$

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. (4.17) becomes:

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

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

Proof of Proposition ??

Proof. Using Proposition 4.18, we obtain that, conditionally on x_1, the log-likelihood is given by

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

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):

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

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

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

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 $\Omega \xrightarrow{\ell} L(Y_T;\hat{\Pi},\Omega)$. We have:

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

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:

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

which leads to the result. □

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. □

4.6 Additional codes

4.6.1 Simulating GEV distributions

The following lines of code have been used to generate Figure 3.7.

```
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)
}
```

4.6.2 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$ED3_TC,na.rm=TRUE),T=60,
                  y.0=rep(0,length(x$phi)),nb.sim=1,make.IRF=1,
                  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)
```

4.7 Statistical Tables

Table 4.1: 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 4.2: 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 4.3: 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 4.4: 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
alpha = 0.9									
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695
500	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644
alpha = 0.95									
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899
alpha = 0.99									
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443

Bibliography

Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10(1):465–503.

Anderson, T. W. and Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1):47–82.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58(2):277–297.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, 1(2):115–126.

Fritsch, M., Pua, A. A. Y., and Schnurbus, J. (2019). Pdynmc - An R-package for estimating linear dynamic panel data models based on linear and nonlinear moment conditions. Passauer Diskussionspapiere, Betriebswirtschaftliche Reihe B-39-19, University of Passau, Faculty of Business and Economics.

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

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

Heiss, F. (2002). Structural choice analysis with nested logit models. *Stata Journal*, 2(3):227–252(26).

Hensher, D. and Greene, W. (2002). Specification and estimation of the nested logit model: alternative normalisations. *Transportation Research Part B: Methodological*, 36(1):1–17.

Jordà, O., Schularick, M., and Taylor, A. M. (2017). Macrofinancial History and the New Business Cycle Facts. *NBER Macroeconomics Annual*, 31(1):213–263.

Litterman, R. and Scheinkman, J. (1991). Common Factors Affecting Bond Returns. *Journal of Fixed Income*, (1):54–61.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. (1995). Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *American Economic Review*, 85(3):322–340.

Mroz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4):765–799.

Nakosteen, R. A. and Zimmer, M. (1980). Migration and income: The question of self-selection. *Southern Economic Journal*, 46(3):840–851.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

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

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

Tobin, J. (1956). Estimation of relationships for limited dependent variables. Cowles Foundation Discussion Papers 3R, Cowles Foundation for Research in Economics, Yale University.