Advanced Econometrics

Jean-Paul Renne

2022-08-29

Contents

Chapter 1

Prerequisites

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

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

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

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

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

Chapter 2

Introduction

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

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

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

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

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

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

Below is an example borrowed from Petersen.

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

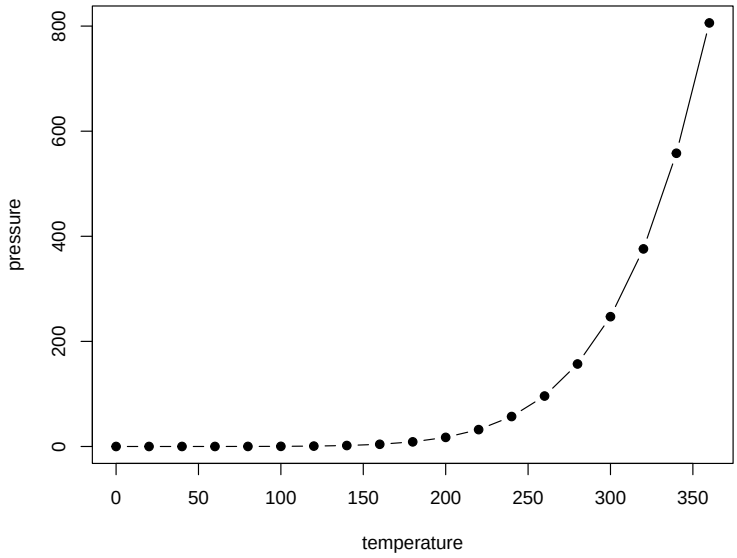


Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

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

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

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

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

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

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

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

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

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

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

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

XXXX

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

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

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

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

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

Diff-in-Diff: Meyer et al. (1995) with data in the `wooldridge` package. See this page

Chapter 3

Microeconometrics

3.1 Binary-choice models

In many instances, the variables to be explained (y_is) have only two possible values (0 and 1, say).

Assume we suspect some variable $\mathbf{x}_i$ ($K \times 1$) to be able to account for the probability that $y_i = 1$.

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 $\mathbf{X}$:

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

where θ is a vector of parameters to be estimated.

The objective is to estimate the vector of population parameters θ.

Binary-choice models can be used to account for…

- any 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).

A possibility is to run a linear regression (a situation called **Linear Probability Model, LPM**):

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

Such a regression could be consistent with the *conditional-mean-zero assumption* (Hypothesis **??**) and with the *assumption of non-correlated residuals* (Hypothesis **??**), but more difficultly with the **homoskedasticity assumption** (Hypothesis **??**). Moreover, the ε_is cannot be Gaussian (because $y_i \in \{0, 1\}$). Therefore, using a linear regression to study the relationship between $\mathbf{x}_i$ and y_i can be consistent but it is inefficient.

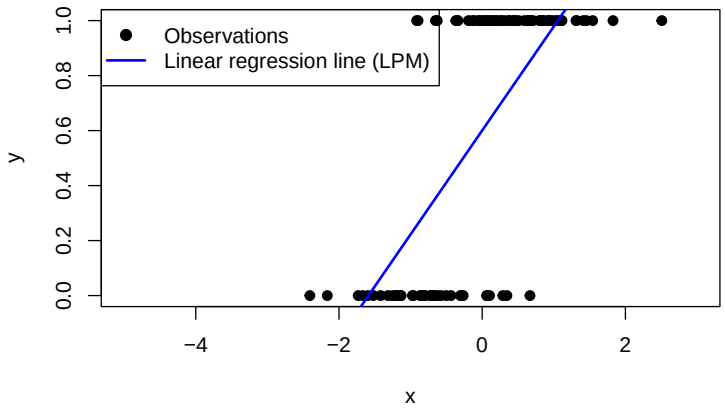


Figure 3.1: Fitting a binary variable with a linear model (Linear Probability Model, LPM). The model is $\mathbb{P}(y_i = 1|x_i) = \Phi(0.5 + 2x_i)$, where Φ is the c.d.f. of the normal distribution and where $x_i \sim i.i.d.\, \mathcal{N}(0, 1)$.

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 "linear" case is given for comparison, but note that it does not satisfy $g(\theta'\mathbf{x}_i)$ 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	x	1

Two prominent models to tackle this situation. In both models, we have:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = g(\theta'\mathbf{x}_i).$$

In the **Probit model**, we have

$$g(z) = \Phi(z), \tag{3.2}$$

where Φ is the c.d.f. of the normal distribution.

And for the **logit model**:

$$g(z) = \frac{1}{1 + \exp(-z)}. \tag{3.3}$$

Figure 3.2 displays the functions g appearing in Table 3.1.

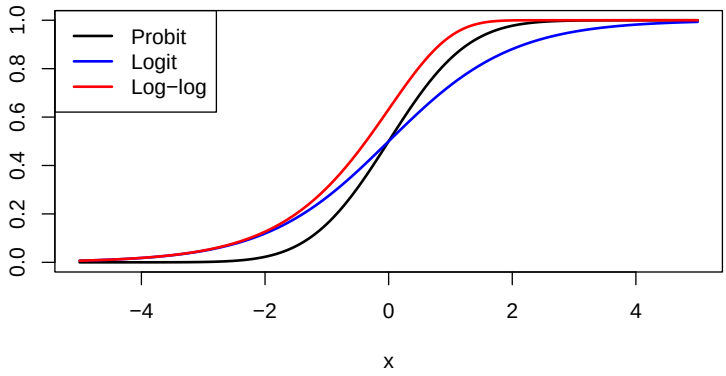


Figure 3.2: Probit, Logit, and Log-log functions.

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

The probit model has an interpretation in terms of **latent variables**. In this model, we indeed 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 $y_i = \mathbb{I}_{\{y_i^* > 0\}}$.

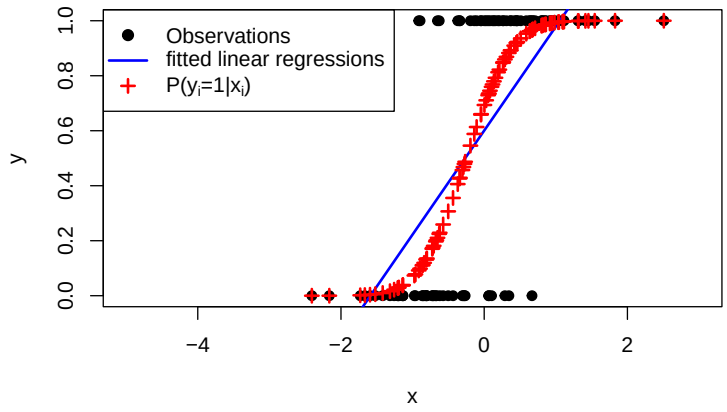


Figure 3.3: The model is $\mathbb{P}(y_i = 1|x_i) = \Phi(0.5 + 2x_i)$, where Φ is the c.d.f. of the normal distribution and where $x_i \sim i.i.d.\,\mathcal{N}(0, 1)$. Crosses give the model-implied probabilities of having $y_i = 1$.

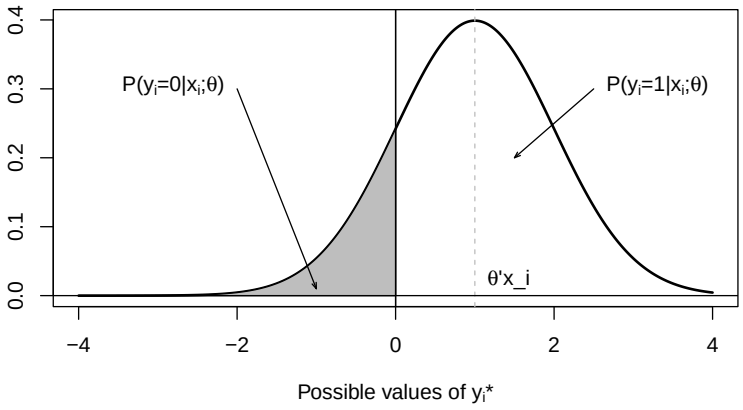


Figure 3.4: Distribution of y_i^* conditional on $\mathbf{x}_i$.

Random Utility Models (RUM) are based on such a view of probit models. 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 $y_i = 0$ $(U_{i,0})$. 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 $\varepsilon_{i,j}$ is a random (agent-specific) component.

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

$$\begin{aligned} \mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) &= \mathbb{P}(\theta_1' \mathbf{x}_i + \varepsilon_{i,1} > \theta_0' \mathbf{x}_i + \varepsilon_{i,0}) \\ &= F(\theta_1' \mathbf{x}_i - \theta_0' \mathbf{x}_i) = F([\theta_1 - \theta_0]' \mathbf{x}_i), \end{aligned} \tag{3.4}$$

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 $\Leftrightarrow$ scaling issue, solved by fixing the variance of $\varepsilon_{i,0} - \varepsilon_{i,1}$.

Other types of structural models –based on the comparison of marginal costs and benefits– give rise to the existence of latent variable and to probit models. An example is Nakosteen and Zimmer (1980). 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, the 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. It cannot be included among the regressors given it is not observed if the considered agent has decided not to work.

- In the context of the choice of transportation mode, the *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. (3.4) becomes:

$$\mathbb{P}(y_i = 1|\mathbf{x}_i; \theta) = F\left(\theta_1^{(u)'} \mathbf{u}_{i,1} - \theta_0^{(u)'} \mathbf{u}_{i,0} + [\theta_1^{(v)} - \theta_0^{(v)}]' \mathbf{v}_i\right), \tag{3.5}$$

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

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

The fishing-mode dataset used in Cameron and Trivedi (2005) (Chapter 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(Ecdat)
data(Fishing)
stargazer::stargazer(Fishing,type="text")
```

```
##
## ======================================================
## Statistic   N     Mean    St. Dev.    Min      Max
## ------------------------------------------------------
## price     1,182  52.082    53.830    1.290   666.110
## catch     1,182   0.389     0.561   0.0002    2.310
## pbeach    1,182  103.422   103.641   1.290   843.186
## ppier     1,182  103.422   103.641   1.290   843.186
## pboat     1,182  55.257    62.713    2.290   666.110
## pcharter  1,182  84.379    63.545   27.290   691.110
## cbeach    1,182   0.241     0.191    0.068    0.533
## cpier     1,182   0.162     0.160    0.001    0.452
## cboat     1,182   0.171     0.210   0.0002    0.737
## ccharter  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 **??**).

To simplify the exposition, we consider the $\mathbf{x}_i$ to be deterministic. Also, we assume that the r.v. are independent across entities i. How to write the likelihood here? It can be seen that:

$$\begin{aligned} f(y_i|\mathbf{x}_i;\theta) &= y_i g(\theta'\mathbf{x}_i) + (1-y_i)(1-g(\theta'\mathbf{x}_i)) \\ &= g(\theta'\mathbf{x}_i)^{y_i}(1-g(\theta'\mathbf{x}_i))^{1-y_i}. \end{aligned} \tag{3.7}$$

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

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

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

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

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 equation that generally has to be numerically solved. Under regularity conditions (Hypotheses **??**), we have (Prop. **??**):

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

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. (**??**)):

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

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

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

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

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

3.1.4 Marginal effectss

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.

It can be estimated by $g'(\theta'_{MLE} \mathbf{x}_i) \theta_{MLE,k}$. It is important to see that the marginal effect depends on $\mathbf{x}_i$: increases by 1 unit of $x_{i,k}$ (entity i) and of $x_{j,k}$ (entity j) do not necessarily have the same effects on $\mathbb{P}(y_i = 1 | \mathbf{x_i}; \theta)$ and on $\mathbb{P}(y_j = 1 | \mathbf{x_j}; \theta)$, respectively.

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:

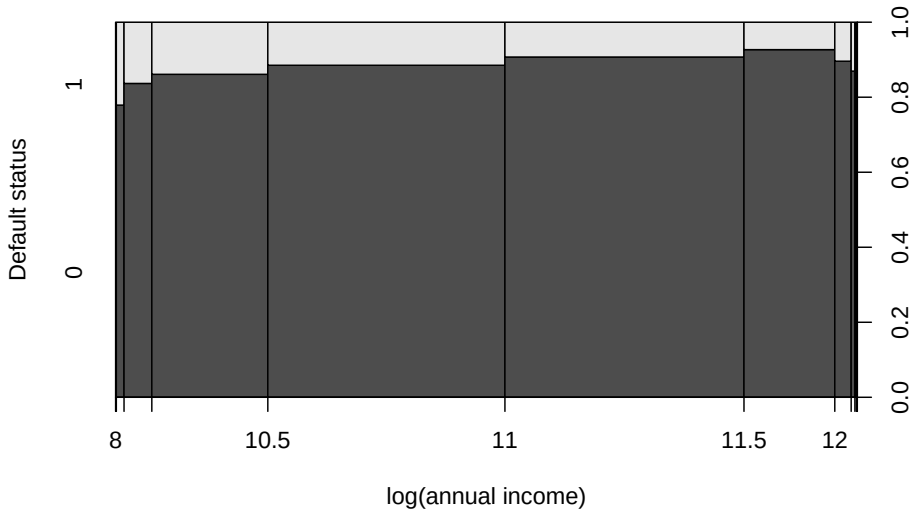
$$R^2_{MF} = 1 - \frac{\log \mathcal{L}(\theta; \mathbf{y})}{\log \mathcal{L}_0(\mathbf{y})}.$$

Intuitively, $R^2_{MF} = 0$ if the explanatory variables do not allow to predict the decision (y).

3.1.6 Example: Credit data

This example makes use of the `credit` data of package `AEC`. The objective is to model the default probabilities of lenders.

```
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)")
```



We consider three specifications. The first one, with no explanatory variables, is trivial. It will just be used to compute the pseudo-R^2.

```
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_in
                data=credit,family=binomial(link="probit"))
eq2 <- glm(Default ~ grade + log(loan_amnt) + amt2income + delinq_2yrs + log(annual_inc)+ I(log(a
                data=credit,family=binomial(link="probit"))
logL0 <- logLik(eq0)
logL1 <- logLik(eq1)
logL2 <- logLik(eq2)
1 - logL1/logL0 # pseudo R2
```

```
## 'log Lik.' 0.01173993 (df=6)
```

```
1 - logL2/logL0 # pseudo R2
```

```
## 'log Lik.' 0.0558487 (df=12)
```

```
stargazer::stargazer(eq0,eq1,eq2,type="text")
```

```
##
## ================================================
##                          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 now compute marginal effects.

```
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, however, an issue for the `annual_inc` variable. Indeed, the previous computation does not realize that this variable appears twice among the explanatory variables. To address this, one can proceed as follows.

```
new_credit <- credit
new_credit$annual_inc <- 1.01 * new_credit$annual_inc # increase of income by 1%
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
```

This average effect is pretty low. To compare, 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. **??**) 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.

3.1.7 Replicating Table 14.2 of Cameron and Trivedi (2005)

The following lines of codes replicate Table 14.2 of Cameron and Trivedi (2005).

```
data.reduced <- subset(Fishing,mode %in% c("charter","pier"))
data.reduced$lnrelp <- log(data.reduced$pcharter/data.reduced$ppier)
data.reduced$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,
                     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
```

3.1.8 Predictions

How to define 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 $c = 0.5$.However, 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. It works as follows:

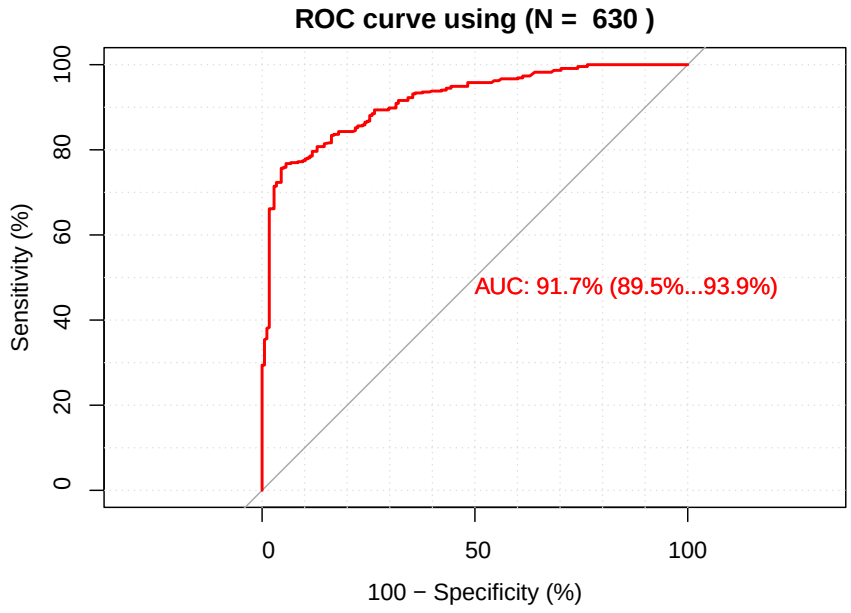
For each potential cutoff $c \in [0, 1]$, compute (and plot):

- 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) [situation when $c = 1$] 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).

```
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),")") )
```




```
##
## Call:
## roc.default(response = data.reduced$y, predictor = predict_model,     percent = T, c
##
## Data: predict_model in 178 controls (data.reduced$y 0) < 452 cases (data.reduced$y 1
## Area under the curve: 91.69%
## 95% CI: 89.5%-93.87% (DeLong)
```

Chapter 4

Appendix

4.1 Statistical Tables

4.2 Statistics: defintitions and results

Definition 4.1 (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:

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

Definition 4.2 (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:*

- $|MN| = |M| \times |N|$.
- $|M^{-1}| = |M|^{-1}$.

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	10
alpha = 0.9										
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663
500	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644	1.612
alpha = 0.95										
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850
alpha = 0.99										
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443	2.356

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

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

Proposition 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 4.4 (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 4.5 (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 4.6 (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 4.7 (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 (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 4.8 (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 4.9 (Convergence in the Lr norm)**.** x_n converges in the r-th mean (or in the L^r-norm) towards x, if $\mathbb{E}(|x_n|^r)$ and $\mathbb{E}(|x|^r)$ exist and if

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

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

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

Definition 4.10 (Almost sure convergence)**.** The random variable sequence x_n converges almost surely to c if $\mathbb{P}(\lim_{n\to\infty} x_n = c) = 1$.

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

Definition 4.11 (Convergence in distribution)**.** x_n is said to converge in distribution (or in law) to x if

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

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

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

Proposition 4.7 (Rules for limiting distributions (Slutsky))**.** *We have:*

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

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

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

Proposition 4.8 (Implications of stochastic convergences)**.** *We have:*

$$\boxed{\xrightarrow{L^s}} \underset{1\le r\le s}{\Rightarrow} \boxed{\xrightarrow{L^r}} \quad \Downarrow \quad \boxed{\xrightarrow{a.s.}} \Rightarrow \boxed{\xrightarrow{p}} \Rightarrow \boxed{\xrightarrow{d}}.$$

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

$$F_n(x) = \mathbb{P}(X_n \le x, X \le x+\varepsilon) + \mathbb{P}(X_n \le x, X > x+\varepsilon) \le F(x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon). \tag{4.1}$$

Besides,

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

which implies:

$$F(x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \le F_n(x). \tag{4.2}$$

Eqs. (4.1) and (4.2) imply:

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

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

Proof. If $\varepsilon > 0$, we have $\mathbb{P}(X_n < c-\varepsilon) \underset{n\to\infty}{\to} 0$ 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 of *plim* but not L^r convergence: Let $\{x_n\}_{n\in\mathbb{N}}$ be a series of random variables defined by:

$$x_n = nu_n,$$

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

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

Theorem 4.1 (Cauchy criterion (non-stochastic case))**.** *We have that $\sum_{i=0}^{T} a_i$ converges ($T \to \infty$) iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

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

Theorem 4.2 (Cauchy criterion (stochastic case))**.** *We have that $\sum_{i=0}^{T} \theta_i \varepsilon_{t-i}$ converges in mean square ($T \to \infty$) to a random variable iff, for any $\eta > 0$, there exists an integer N such that, for all $M \geq N$,*

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

Definition 4.12 (Characteristic function)**.** For any real-valued random variable X, the characteristic function is defined by:

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

Theorem 4.3 (Law of large numbers)**.** *The sample mean is a consistent estimator of the population mean.*

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

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

The properties of the characteristic function (see Def. 4.12) 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 4.4 (Lindberg-Levy Central limit theorem, CLT)**.** *If x_n is an i.i.d. sequence of random variables with mean μ and variance σ^2 ($\in]0, +\infty[$), then:*

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

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 $\mathcal{N}(0, \sigma^2)$. □

Proposition 4.10 (Inverse of a partitioned matrix)**.** *We have:*

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \\ \begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix}.$$

Proposition 4.11. *If $\mathbf{A}$ is idempotent and if $\mathbf{x}$ is Gaussian, $\mathbf{Lx}$ and $\mathbf{x}'\mathbf{Ax}$ are independent if $\mathbf{LA} = \mathbf{0}$.*

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

Proposition 4.12 (Inner product of a multivariate Gaussian variable)**.** *Let X be a n-dimensional multivariate Gaussian variable: $X \sim \mathcal{N}(0, \Sigma)$. We have:*

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

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

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

Theorem 4.5 (Cauchy-Schwarz inequality)**.** *We have:*

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

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

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

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

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

Definition 4.13 (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 4.13. *We have:*

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

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

Definition 4.14 (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 4.15 (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 4.16 (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 4.14 (Properties of the Kullback discrepancy)**.** *We have:*

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

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

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

4.3 Proofs

4.3.1 Proof of Proposition ??

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

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

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

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

Properties of the Kullback information measure (see Prop. 4.14), together with the identifiability assumption (ii) implies that the solution to the limit problem is unique and equal to θ_0.

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

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

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

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

or equivalently:

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

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

Besides, we have:

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

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

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

4.3.2 Proof of Proposition ??

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

$$\sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{4.3}$$

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

$$\sqrt{n}h(\hat{\theta}_n) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h(\theta_0)'}{\partial \theta}\right). \tag{4.4}$$

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. (4.3) is still valid. It implies that ξ_n^W converges to $+\infty$ and therefore that $\mathbb{P}_\theta(\xi_n^W \geq \chi^2_{1-\alpha}(r)) \to 1$. □

4.3.3 Proof of Proposition ??

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:

$$\sqrt{n}h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'} \sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0). \tag{4.5}$$

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

$$\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n^0 - \theta_0) \tag{4.6}$$

and:

$$0 = \frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\hat{\theta}_n; \mathbf{y})}{\partial \theta} \approx \frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0; \mathbf{y})}{\partial \theta} - \mathcal{I}(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0). \tag{4.7}$$

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

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^0) \approx \mathcal{I}(\theta_0)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \mathcal{I}(\theta_0). \tag{4.8}$$

Eqs. (4.5) and (4.8) yield to:

$$\sqrt{n} h(\hat{\theta}_n) \approx \frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta}. \tag{4.9}$$

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:

$$\frac{\partial \log \mathcal{L}(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} + \frac{\partial h'(\hat{\theta}_n^0; \mathbf{y})}{\partial \theta} \hat{\lambda}_n = 0. \tag{4.10}$$

Substituting the latter equation in Eq. (4.9) gives:

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

which yields:

$$\frac{\hat{\lambda}_n}{\sqrt{n}} \approx -\left(\frac{\partial h(\theta_0)}{\partial \theta'} \mathcal{I}(\theta_0)^{-1} \frac{\partial h'(\theta_0; \mathbf{y})}{\partial \theta} \right)^{-1} \sqrt{n} h(\hat{\theta}_n). \tag{4.11}$$

It follows, from Eq. (4.4), 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. (4.10), it appears that the left-hand side term of the last equation is ξ^{LM} as defined in Eq. (**??**). Consistency: see Remark 17.3 in Gouriéroux and Monfort (1995). □

4.3.4 Proof of Proposition ??

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

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

Since, under H_0, $\hat{\theta}_n^0 \approx \hat{\theta}_n \approx \theta_0$, Eq. (4.11) 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. □

4.3.5 Proof of Proposition ??

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

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

Taking the difference, we obtain:

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

Using $\frac{1}{\sqrt{n}} \frac{\partial \log \mathcal{L}(\theta_0;\mathbf{y})}{\partial \theta} \approx \mathcal{I}(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$ (Eq. (4.7)), we have:

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

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

$$\xi_n^{LR} \approx n(\hat{\theta}_n^0 - \hat{\theta}_n)' \mathcal{I}(\theta_0)^{-1} (\hat{\theta}_n^0 - \hat{\theta}_n). \tag{4.12}$$

The difference between Eqs. (4.6) and (4.7) 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 test is consistent. For this, note that:

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

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

Bibliography

Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10(1):465–503.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

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

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

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.

Meyer, B. D., Viscusi, W. K., and Durbin, D. L. (1995). Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *American Economic Review*, 85(3):322–340.

Nakosteen, R. A. and Zimmer, M. (1980). Migration and income: The question of self-selection. *Southern Economic Journal*, 46(3):840–851.

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

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

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

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