UNIVERSITAS
FRIBURGENSIS

# Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logistic Regression Models

**Diploma Thesis**

submitted to the

Department of Mathematics
Faculty of Sciences
University of Fribourg Switzerland

in partial fulfilment of the requirements
for the attainment of the

Diploma in Mathematics

by

**Michael Beer**

⋆ ⋆ ⋆

Academic supervisor: Prof Dr André Antille
Submission date: December 2001

Address of the author: Henzenmoos 41, 3182 Überstorf, Switzerland
E-mail: `Michael.Beer@unifr.ch`

## Abstract

This diploma thesis investigates the consistency and asymptotic normality of the maximum likelihood estimator in dichotomous logistic regression models. The consistency is proved under hypotheses which allow the number of exogenous variables $p$ to grow along with the sample size $n$. Conditions on the interdependence of $p$ and $n$ necessary for the consistency of the estimator are established. The asymptotic normality is proved in the case where $p$ is constant. Some aspects on the motivation and the origins of the model are presented at the beginning. Finally, a short example based on a medical study outlines the application of the logistic regression model.

# Contents

# 1 Introduction: A Regression Model for Dichotomous Outcome Variables

Among the different regression models, logistic regression plays a particular role. The base concept, however, is universal. In a statistical investigation, several variables are defined and their values are determined for a set of objects. Some of these variables are a priori considered to be dependent on others within the same framework. Others are thought to be exclusively dependent on exterior and not quantified (or not quantifiable) aspects, but not on further variables inside the same model. The main aim of a regression analysis is to reveal and specify the impact of the so-called *independent* or *exogenous* explanatory variables on the *dependent* or *endogenous* response.

The prevalent linear regression model is – under certain hypotheses – in many circumstances a valuable tool for quantifying the effects of several exogenous on one dependent continuous variable. For situations where the dependent variable is qualitative, however, other methods had to be developed. One of these is the logistic regression model which specifically covers the case of a binary (dichotomous) response.

In a first section of this thesis, the logistic regression model as well as the maximum likelihood procedure for the estimation of its parameters are introduced in detail. A notation is fixed and will be used throughout the remaining parts of this document. Furthermore, several ideas on the motivation of the model and the interpretation of its parameters are outlined. At the end, a short paragraph illuminates a few historical aspects in relation to the development of this model.

The second part examines the consistency of the maximum likelihood estimator for the model parameters. It will be shown that, if certain hypotheses are satisfied, this estimator converges in probability to the true parameter values as the sample size increases. As a peculiarity of the procedure presented in this thesis, the number of explanatory variables is allowed to grow simultaneously with the number of observations. Conditions on the admissible interdependence of both of these characteristics are given in order to ensure the consistency of the estimator. Finally, these conditions are compared to the assumptions of an existing consistency theorem.

As another important property, the asymptotic normality of the estimator being discussed is proved in section 4. Picking up the thread of the previous part, this aspect is established in a similar context. In contrast to the proof of consistency, however, this section will be based on the “classical” approach where the number of independent variables is held constant.

To conclude, a short example attempts to illustrate the application of the logistic regression model. Proceeding from a survey of health enhancing physical activity carried out by the Swiss Federal Offices of Sports and Statistics, an estimation will be made what effects the linguistic background of a Swiss inhabitant has on his or her daily or weekly physical activity. The results are presented along with the *Mathematica*[1] code used for the calculations.

[1] *Mathematica* is a registered trademark of Wolfram Research, Inc.

# 2 Logistic Regression Model

## 2.1 Model Specification

We consider a binary random variable $y$ having a Bernoulli distribution,

$$y \overset{\mathrm{L}}{\sim} B\big(1, \pi(\boldsymbol{x})\big) \,,$$

i.e. the variable $y$ takes either the value 1 or the value 0 with probabilities $\pi(\boldsymbol{x})$ or $1 - \pi(\boldsymbol{x})$ respectively. $\boldsymbol{x} \in \mathbb{R}^p$ is a vector of $p$ exogenous variables and $\pi : \mathbb{R}^p \longrightarrow [0, 1]$ a real-valued function. In fact, $\pi(\boldsymbol{x})$ represents the conditional probability $P(y = 1 \,|\, \boldsymbol{x})$ of $y = 1$, given $\boldsymbol{x}$.

Let $r := y - \pi(\boldsymbol{x})$, which allows us to rewrite our model as

$$y = \pi(\boldsymbol{x}) + r \,,$$

where $r$ has an expectation of

$$\mathrm{E}(r) = \mathrm{E}\big(y - \pi(\boldsymbol{x})\big) = \mathrm{E}(y) - \pi(\boldsymbol{x}) = \pi(\boldsymbol{x}) - \pi(\boldsymbol{x}) = 0 \tag{2.1a}$$

and a variance of

$$\mathrm{Var}(r) = \mathrm{Var}(y) = \pi(\boldsymbol{x})\big(1 - \pi(\boldsymbol{x})\big) \,. \tag{2.1b}$$

For the forthcoming analysis we are going to define the so-called *logistic transformation* $\sigma_{\mathrm{LR}} : \mathbb{R} \longrightarrow [0, 1]$ by

$$\sigma_{\mathrm{LR}}(z) := \frac{\exp z}{1 + \exp z} = \frac{1}{1 + \exp -z}$$

which allows us to specify the probability function $\pi$ as

$$\pi(\boldsymbol{x}) := \sigma_{\mathrm{LR}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$$

with a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ of unknown parameters. This specification yields the *logistic regression model* with parameter $\boldsymbol{\beta}$.

If we denote the inverse function of $\sigma_{\mathrm{LR}}$, which is called *logit transformation*, by

$$\operatorname{logit} \pi = \ln \frac{\pi}{1 - \pi} \,,$$

we can write our regression model as

$$\operatorname{logit} \pi(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} \,. \tag{2.2}$$

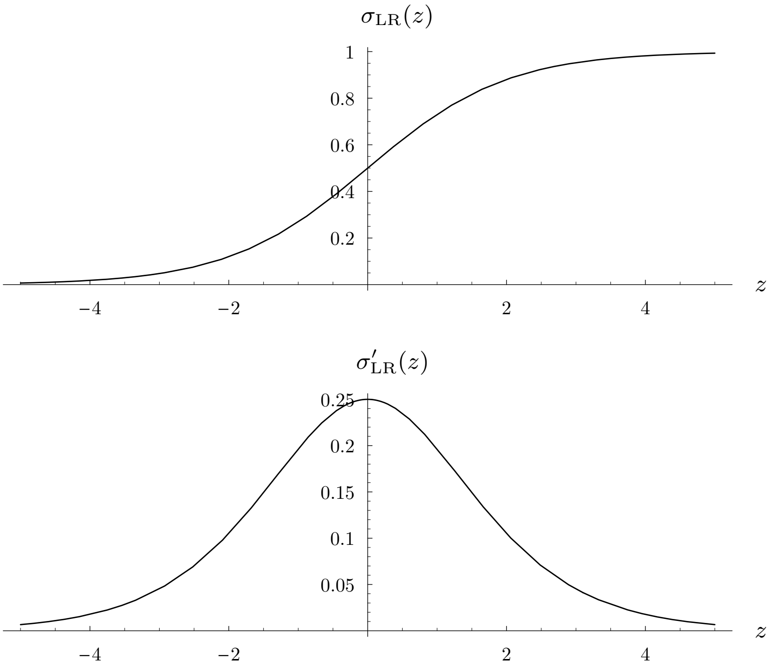

Figure 1: The logistic transformation $\sigma_{\mathrm{LR}}$ and its first derivative.

Furthermore, it will be important to note that

$$\begin{aligned}\sigma'_{\mathrm{LR}}(z) &= \frac{\partial}{\partial z}\left(\frac{\exp z}{1+\exp z}\right) = \frac{(\exp z)(1+\exp z)-(\exp z)^2}{(1+\exp z)^2} \\ &= \frac{\exp z}{(1+\exp z)^2} > 0 \quad \forall\, z \in \mathbb{R}\,.\end{aligned} \tag{2.3}$$

The shape of $\sigma_{\mathrm{LR}}$ and its first derivative $\sigma'_{\mathrm{LR}}$ are displayed in figure 1. Possible motivations for this specific model shall be discussed in section 2.3.

## 2.2 Maximum Likelihood Estimation of the Parameter $\boldsymbol{\beta}$

In order to estimate the true value $\boldsymbol{\beta}^0$ of the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$, we consider a set of $n$ observations $\{(y_i, \boldsymbol{x}_i)\}_{i\in\{1,\ldots,n\}} \in \big(\{0,1\} \times \mathbb{R}^p\big)^n$ of mutually independent responses $y_i$ with respective explanatory variables $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$. Assuming

$$y_i \overset{\mathrm{L}}{\sim} B\big(1, \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)\big) \quad (i = 1, \ldots, n) \tag{2.4}$$

we consider the model

$$\boldsymbol{y} = \sigma_{\mathrm{LR}}(\boldsymbol{X}\boldsymbol{\beta}^0) + \boldsymbol{r}\,,$$

where

$$\begin{aligned}
\boldsymbol{y} &:= (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \in \{0,1\}^n \,, \\
\boldsymbol{X} &:= \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{n \times p} \,, \text{ and} \\
\boldsymbol{r} &:= \boldsymbol{y} - \sigma_{\mathrm{LR}}(\boldsymbol{X}\boldsymbol{\beta}^0) \,.
\end{aligned}$$

The function $\sigma_{\mathrm{LR}}$ applied to a vector $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}} \in \mathbb{R}^n$ has to be interpreted as $\sigma_{\mathrm{LR}}(\boldsymbol{z}) := (\sigma_{\mathrm{LR}}(z_1), \ldots, \sigma_{\mathrm{LR}}(z_n))^{\mathrm{T}}$. By (2.1a) and (2.1b), we know that

$$\mathrm{E}(r_i) = 0 \tag{2.5a}$$

and

$$\begin{aligned}
s_i^2 := \mathrm{Var}(r_i) &= \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)\big(1 - \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)\big) \\
&= \frac{\exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0}{1 + \exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0} \cdot \frac{1 + \exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0 - \exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0}{1 + \exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0} \\
&= \frac{\exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0}{(1 + \exp \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)^2} = \sigma_{\mathrm{LR}}'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0) \,.
\end{aligned} \tag{2.5b}$$

In this context, we shall analyse the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of the true parameter vector $\boldsymbol{\beta}^0$. The likelihood function is defined as

$$\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{i=1}^n \mathrm{P}(y = y_i) \\
&= \prod_{i=1}^n \binom{1}{y_i} \big(\sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)^{y_i} \big(1 - \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)^{1-y_i} \\
&= \prod_{i=1}^n \left(\frac{\sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 - \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right)^{y_i} \big(1 - \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big) \\
&= \prod_{i=1}^n \left(\frac{\frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}}{\frac{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})-\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}}\right)^{y_i} \left(\frac{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}\right) \\
&= \prod_{i=1}^n \frac{\big(\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big)^{y_i}}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} = \prod_{i=1}^n \frac{\exp(y_i\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}
\end{aligned}$$

and yields the log-likelihood function

$$\begin{aligned}\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \ln\bigl(\exp(y_i\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\bigr) - \sum_{i=1}^{n} \ln\bigl(1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\bigr) \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} - \sum_{i=1}^{n} \ln\bigl(1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\bigr)\,.\end{aligned} \tag{2.6}$$

To determine the maximum of this function, we consider the gradient

$$\begin{aligned}\nabla \ln L(\boldsymbol{\beta}) &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \sum_{i=1}^{n} \underbrace{\frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}}_{=\sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \boldsymbol{x}_i = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{X}^{\mathrm{T}}\sigma_{\mathrm{LR}}(\boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \sigma_{\mathrm{LR}}(\boldsymbol{X}\boldsymbol{\beta})\right).\end{aligned}$$

A necessary condition for $\hat{\boldsymbol{\beta}}$ being the maximum likelihood estimator of $\boldsymbol{\beta}^0$ is therefore

$$\boldsymbol{X}^{\mathrm{T}}\Bigl(\boldsymbol{y} - \sigma_{\mathrm{LR}}\bigl(\boldsymbol{X}\hat{\boldsymbol{\beta}}\bigr)\Bigr) = \boldsymbol{0}\,. \tag{2.7}$$

From the second partial derivatives

$$\begin{aligned}\frac{\partial^2}{\partial\beta_s\,\partial\beta_r} \ln L(\boldsymbol{\beta}) &= \frac{\partial}{\partial\beta_s}\left(\sum_{i=1}^{n} x_{ir}\, y_i - \sum_{i=1}^{n} x_{ir}\, \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\right) \\ &= -\sum_{i=1}^{n} x_{ir} \frac{\partial}{\partial\beta_s}\, \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) = -\sum_{i=1}^{n} x_{ir}\, x_{is}\, \sigma_{\mathrm{LR}}'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\end{aligned}$$

we can derive the Hessian matrix $\mathbf{H}_{\ln L}(\boldsymbol{\beta}) \in \mathbb{R}^{p\times p}$ of the log-likelihood function $\ln L(\boldsymbol{\beta})$ given by

$$\mathbf{H}_{\ln L}(\boldsymbol{\beta}) = -\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{X}\,,$$

where $\boldsymbol{D}(\boldsymbol{\beta}) = (d_{ij}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix defined by

$$d_{ij} = \begin{cases} \sigma_{\mathrm{LR}}'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.8}$$

**Affirmation.** $\mathbf{H}_{\ln L}(\boldsymbol{\beta})$ *is negative semi-definite for any* $\boldsymbol{\beta} \in \mathbb{R}^p$*.* ◇

**Verification.** We have

$$\boldsymbol{u}^{\mathrm{T}}\, \mathbf{H}_{\ln L}(\boldsymbol{\beta})\, \boldsymbol{u} = -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{X}\boldsymbol{u} = -\sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{u})^2\, \sigma_{\mathrm{LR}}'(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\,. \tag{2.9}$$

As the first derivative of $\sigma_{\mathrm{LR}}$ is always positive (see (2.3)), we can see from equation (2.9) that $\boldsymbol{u}^{\mathrm{T}}\, \mathbf{H}_{\ln L}(\boldsymbol{\beta})\, \boldsymbol{u} \leq 0$ for all $\boldsymbol{u} \in \mathbb{R}^p$ and all $\boldsymbol{\beta} \in \mathbb{R}^p$. ■

We therefore know that, in any case, the log-likelihood function $\ln L(\boldsymbol{\beta})$ is concave. Consequently, any root of $\nabla \ln L(\boldsymbol{\beta})$ is a global maximum of $\ln L(\boldsymbol{\beta})$ and a maximum likelihood estimator of $\boldsymbol{\beta}^0$. The condition (2.7) therefore is not only necessary but also sufficient for $\hat{\boldsymbol{\beta}}$ being a maximum likelihood estimator of $\boldsymbol{\beta}^0$. However, neither its existence nor its uniqueness are a priori guaranteed.

**Remark.** If we assumed $p \leq n$ and $\operatorname{rg} \boldsymbol{X} = p$, we would obtain that $\boldsymbol{X}\boldsymbol{u} = \mathbf{0}$ is equivalent to $\boldsymbol{u} = \mathbf{0}$, such that $\mathbf{H}_{\ln L}(\boldsymbol{\beta})$ were negative definite for all $\boldsymbol{\beta} \in \mathbb{R}^p$ and thus $\ln L(\boldsymbol{\beta})$ strictly concave. Any $\hat{\boldsymbol{\beta}}$ satisfying condition (2.7) therefore would be the *unique* maximum likelihood estimator of $\boldsymbol{\beta}^0$. $\diamond$

Even though the existence of $\hat{\boldsymbol{\beta}}$ cannot be guaranteed for finite sample spaces, we shall prove in section 3 that the probability of non-existence approaches zero as $n$ tends to infinity.

Iterative methods for obtaining the maximum likelihood estimator are presented for example by Amemiya (1985, pp. 274-275).

## 2.3 Motivation and Interpretation

The logistic regression model is a common tool for handling dichotomous response variables. Several reasons account for this circumstance. We now want to focus on two of them: the interpretation of the regression parameters and the relationship with the logistic distribution. Other motivations, such as the "formal connection of the logistic model with loglinear model theory" and "theoretical statistical considerations including the availability of 'exact' analyses of individual parameters" mentioned by Santner and Duffy (1989, p. 206), shall not be investigated within the scope of this thesis.

### 2.3.1 Odds, Log Odds and the Odds Ratio

In order to discuss binary data and to interpret the regression coefficients, it is essential to mention the terms of *odds* and *odds ratios* as counterparts to probabilities and probability ratios.

Suppose that an event $A$ has a probability $\pi$. We then define the *odds* of $A$ as the probability that $A$ occurs divided by the probability it does not occur, i.e.

$$Odds(A) := \frac{P(A)}{P(A^c)} = \frac{\pi}{1-\pi}\,, \tag{2.10}$$

where $A^c$ denotes the complementary event of $A$ in the respective sample space.

Such a way of thinking is quite common in practice. If in a sporting competition, for example, the probability of one team winning is about 80 percent, many people say that the odds on this team winning are four to one. However, as Christensen (1997, p. 2) mentions, odds are not infrequently confused with probabilities. In a document entitled "What Are the Odds of Dying?" the US National Safety Council (2001) states for instance: "The odds of dying from an injury in 1998 were 1 in 1,796." This number was approximated by dividing the 1998 US population (270,248,000) by the respective number of deaths (150,445). From a statistical point of view, the number 1,796 is an estimation of the reciprocal probability, whereas an estimation of the real odds on dying from an injury would rather be 150,445/(270,248,000 − 150,445), i.e. 1 to 1,795. The expression "1 in" instead of "1 to" and the large numbers involved in this study justify such an approximation procedure to a certain extent.

This latter example points out the need of careful distinction between odds and probability. Note however, that there is a one-to-one relationship between both of them.[2] Given the probability $\pi$, the respective odds $o$ are obtained by formula (2.10), while $\pi$ is calculated from any given $o$ by $\pi = o/(1+o)$. As Christensen (1997, p. 3) emphasises, examining odds therefore amounts to a rescaling of the measure of uncertainty: Probabilities between 0 and 1/2 correspond to odds between 0 and 1 whereas odds between 1 and $\infty$ correspond to probabilities between 1/2 and 1.

The loss of symmetry inherent in the transformation of probabilities to odds can be offset by a subsequent application of the natural logarithm function. Looking at the so-called *log odds*, we observe that they are "symmetric about zero just as probabilities are symmetric about one half" (Christensen 1997, p. 3). This is why mathematical analyses often deal with log odds instead of "simple" odds. It is worth noting that the log odds are usually allowed to take the values $-\infty$ and $+\infty$ in order to establish a one-to-one relationship with the respective probabilities 0 and 1, and, moreover, that the transformation of probabilities to log odds is exactly the logit transformation introduced in section 2.1.

In order to compare the odds of two different events, it may be useful to examine not only odds but also *odds ratios*. If the odds of an event $A$ are $Odds(A)$ and the odds of an event $B$ are $Odds(B)$, then the odds ratio of $B$ to $A$ is defined as $Odds(A)\big/Odds(B)$.[3] In the above study on the "Odds of Dying", for instance, the odds of dying in a railway accident were 1 to 524,752 whereas the odds of an accidental death in a motor-vehicle

[2] This holds under the condition that the odds are allowed to take the value $+\infty$ when the probability $\pi$ equals 1.

[3] Note that this syntax is not used identically throughout the literature. While many authors just speak of the ratio of the odds of $A$ to the odds of $B$, Garson (2001, "Log-linear Models, Logit, and Probit") explicitly refers to the expression "odds ratio of $B$ to $A$".

were 1 to 6,211. The odds ratio of dying in a railway accident to dying in a motor-vehicle in the USA thus was about 84, i.e. the odds of dying in a motor-vehicle were about 84 times as important as those of dying in a railway crash.

### 2.3.2 Interpretation of the Parameter $\boldsymbol{\beta}$

For an interpretation of the parameter $\boldsymbol{\beta}$, let $\boldsymbol{x} \in \mathbb{R}^p$ and $\check{\boldsymbol{x}} := \boldsymbol{x} + \boldsymbol{e}_i$ where $\boldsymbol{e}_i$ denotes the $i$th canonical base vector of $\mathbb{R}^p$ for an arbitrary $i \in \{1, \ldots, p\}$. A comparison of the two model equations (2.2) for $\boldsymbol{x}$ and $\check{\boldsymbol{x}}$ gives

$$\begin{aligned} \operatorname{logit} \pi(\boldsymbol{x}) &= \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta} \\ \operatorname{logit} \pi(\check{\boldsymbol{x}}) &= \check{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{\beta} = (\boldsymbol{x} + \boldsymbol{e}_i)^{\mathrm{T}} \boldsymbol{\beta} = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{\beta} \end{aligned}$$

from where we get the difference

$$\operatorname{logit} \pi(\check{\boldsymbol{x}}) - \operatorname{logit} \pi(\boldsymbol{x}) = \boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{\beta} = \beta_i \, .$$

On the other hand,

$$\operatorname{logit} \pi(\check{\boldsymbol{x}}) - \operatorname{logit} \pi(\boldsymbol{x}) = \ln \left( \frac{\pi(\check{\boldsymbol{x}})}{1 - \pi(\check{\boldsymbol{x}})} \right) - \ln \left( \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} \right) = \ln \left( \frac{\frac{\pi(\check{\boldsymbol{x}})}{1 - \pi(\check{\boldsymbol{x}})}}{\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}} \right) .$$

We thus see that $\exp \beta_i$ is the odds ratio of $\{y = 1\}$ given $\boldsymbol{x}$ to $\{y = 1\}$ given $\check{\boldsymbol{x}}$. In other words, the odds on the event that $y$ equals 1 increase (or decrease) by the factor $\exp \beta_i$ if $x_i$ grows by one unit. If in a certain study, for example, $\beta_2$ has an estimated value of about 0.693, a unit increase in $x_2$ is likely to double the odds on getting a positive response ($y = 1$), as $\exp 0.693 \approx 2$.

According to Hosmer and Lemeshow (1989, p. 41), “this fact concerning the interpretability of the coefficients is the fundamental reason why logistic regression has proven such a powerful analytic tool for epidemiologic research.” At least, this argumentation holds whenever the explanatory variables $\boldsymbol{x}$ are quantitative.

Collett (1991, pp. 242-246) gives an outline of a suitable procedure for the case of qualitative exogenous variables. Assume, for instance, that the probability $\pi$ is to be related to a single exposure factor $g$ that has $m$ levels, expressed through integer values between 0 and $m - 1$. Instead of introducing $g$ into a model like

$$\operatorname{logit} \pi = \beta_0 + \beta_1 \, g \, ,$$

it is preferable to define a set of dummy variables $x_1, \ldots, x_{m-1}$ as

$$x_i := \begin{cases} 1 & \text{if } g = i, \\ 0 & \text{if } g \neq i, \end{cases} \qquad (i = 1, \ldots, m-1)$$

and to analyse the model

$$\operatorname{logit} \pi = \beta_0 + \beta_1 \, x_1 + \cdots + \beta_{m-1} \, x_{m-1}$$

for instance. This procedure allows to distinguish between the effects on $y$ of every single factor level of $g$. An application of this idea will be shown in section 5.

The possibility of interpreting the coefficients of $\boldsymbol{\beta}$ as logarithms of odds ratios provides the foundation of a second important motivation of the logistic model. Both Santner and Duffy (1989, pp. 206-207) and Christensen (1997, pp. 118-120, p. 387) emphasise on the difference between *prospective* and *retrospective* studies. Consider for instance an experiment in which 250 people of an arbitrary population are sampled. A binary response "diseased" ($D$) or "non-diseased" ($D^c$) is observed for each person. Moreover, there is a single explanatory variable "exposed" ($E$) or "non-exposed" ($E^c$) involved. This kind of study is called *prospective*. Let $\psi_P$ denote the (prospective) ratio of odds of disease for the exposed group to odds of disease for the non-exposed group as

$$\psi_P = \frac{P(D|E)}{1 - P(D|E)} \bigg/ \frac{P(D|E^c)}{1 - P(D|E^c)} \ .$$

According to the nature of the study, diseased individuals may be very rare in a random sample of 250 people. So most of the collected data is about non-diseased persons. It is therefore sometimes useful to fix the sample size in the rare event category by design. In our example, one could possibly study separated samples of 100 diseased and 150 non-diseased individuals while determining for every person whether he or she had been exposed or not. This procedure is called *retrospective* and leads directly to information about the probability of exposure among the diseased and among the healthy groups. We thus get the (retrospective) odds ratio

$$\psi_R = \frac{P(E|D)}{1 - P(E|D)} \bigg/ \frac{P(E|D^c)}{1 - P(E|D^c)} \ .$$

However, we obtain by Bayes's rule that

$$\psi_P = \frac{\frac{P(D|E)}{P(D^c|E)}}{\frac{P(D|E^c)}{P(D^c|E^c)}} = \frac{\frac{P(E|D)P(D)}{P(E|D^c)P(D^c)}}{\frac{P(E^c|D)P(D)}{P(E^c|D^c)P(D^c)}} = \frac{\frac{P(E|D)}{P(E|D^c)}}{\frac{1-P(E|D)}{1-P(E|D^c)}} = \psi_R$$

so that we are able to make inferences about $\psi_P$ even from a retrospective study. The generalisation of this result motivates the inspection of odds ratios.

### 2.3.3 Relationship with the Logistic Distribution

Another important issue in connection with the logistic model is outlined by Amemiya (1985, pp. 269-270) and Cramer (1991, pp. 11-13). Both of them cite two examples, one a biometric threshold model and the other an econometric utility model.

In the biometric model, they suppose a dosage $x_i$ of an insecticide is given to the $i$th insect of a population. Furthermore, it is assumed that every insect has its own tolerance or threshold level $y_i^*$ against this insecticide. If the dosage $x_i$ is higher than $y_i^*$, the insect dies, if it is lower, the insect survives. The binary variable $y_i$ expresses whether the $i$th insect dies ($y_i = 1$) or not ($y_i = 0$). We thus get

$$P(y_i = 1) = P(y_i^* < x_i) = F(x_i)\,,$$

where $F$ is the cumulative distribution function of the random variable $y_i^*$.

The econometric model, on the other hand, attributes separate random utilities $u_1$ and $u_0$ to the two possible states $y = 1$ and $y = 0$ of a certain variable $y$. For example, $y = 1$ could express that a person drives a car whereas $y = 0$ would mean that this person travels by transit to work. Both utilities $u_1$ and $u_0$ are assumed to depend on certain characteristics, represented by a vector $\boldsymbol{x}_i$, as

$$u_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i \qquad (i = 0, 1)\,,$$

where $\varepsilon_i$ is a random error term. If we suppose that our individual maximises its utility, the probability of his decision in favour of the state $y = 1$ is therefore given by

$$\begin{aligned} P(y = 1) = P(u_0 < u_1) &= P(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_0 < \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_1) \\ &= P(\varepsilon_0 - \varepsilon_1 < \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta}) = F\big((\boldsymbol{x}_1 - \boldsymbol{x}_0)^{\mathrm{T}}\boldsymbol{\beta}\big)\,, \end{aligned}$$

where $F$ is the cumulative distribution function of the random variable $\epsilon := \varepsilon_0 - \varepsilon_1$.

In both of these examples, an assumption on the shape of the cumulative distribution function $F$ or rather on the distribution of the underlying random variable has to be made in order to estimate the probability of the event $\{y = 1\}$. With respect to the central limit theorem, the assumption $F = \Phi(0, 1)$, where the so-called *probit function* $\Phi(0, 1)$ is the cumulative distribution function of the standardised normal distribution, could most probably be justified, at least after an appropriate standardisation of the relevant data. This specific choice would lead to what is called the *probit model.*

However, the logistic transformation $\sigma_{\mathrm{LR}}$ itself provides a very similar and analytically sometimes more convenient alternative. Being viewed as a distribution function, $\sigma_{\mathrm{LR}}$ gives rise to the *logistic distribution*. Its density $\sigma'_{\mathrm{LR}}$ has mean zero and variance $\pi^2/3$, so that it is appropriate to define the cumulative distribution function of the *standardised* logistic distribution with zero mean and unit variance as

$$\Lambda(x) := \sigma_{\mathrm{LR}}(\lambda x) = \frac{\exp \lambda x}{1 + \exp \lambda x}$$

where $\lambda = \pi\big/\sqrt{3}$. In the literature, for example by Cramer (1991), the function $\Lambda$ is sometimes called *logit function* and is therefore not to be confused with the logit transformation introduced in section 2.1, which is nearly its inverse.

Cramer (1991, section 2.3) shows in an example that "by judicious adjustment of the linear transformations of the argument $x$, the logit and probit probability functions can be made to coincide over a fairly wide range." Moreover, he states that "logit and probit functions which have been fitted to the same data are therefore virtually indistinguishable, and it is impossible to choose between the two on empirical grounds."

As a result, it is justifiable in most cases to assume a logistic distribution instead of a normal distribution for $y_i^*$ or $\epsilon$ in examples as those mentioned above. This, however, guides us again to the logistic regression approach.

## 2.4 History of the Logistic Regression Model

An overview of the development of the logistic regression model is given by Cramer (1991, section 2.9). He identifies three sources having lead to this model as it is known today: applied mathematics, experimental statistics, and economic theory.

While studying the development of the human population, Thomas Robert Malthus (1766-1834) described an increase "in geometric progression". It can be argued that at the basis of this statement, there was the formula for exponential growth, $N(t) = A \exp \alpha t$, evolving directly from the differential equation $\dot{N}(t) = \alpha N(t)$ where $N(t)$ denotes the size of a population at a time $t$. This model, however, did not take into consideration the possibility of a saturation in the growth process. For this reason, Pierre François Verhulst (1804-1849) added an upper limit or saturation value $W$ to the equation mentioned above: $\dot{N}(t) = \beta N(t)(W - N(t))$. Defining $Z(t) := N(t)/W$, we obtain an equation of the form $\dot{Z}(t) = \gamma Z(t)(1 - Z(t))$ whose solution is the logistic function

$$Z(t) = \frac{\exp(\alpha + \gamma t)}{1 + \exp(\alpha + \gamma t)}\,.$$

As Cramer (1991) states, this model of human population growth was introduced independently of Verhulst's study by Raymond Pearl and Lowell Reed in an article entitled "On the rate of growth of the population of the United States since 1790 and its mathematical representation" published in 1920. Cramer continues: "The basic idea that growth is proportional both to the level already attained and to the remaining room to the saturation ceiling is simple and effective, and the logistic model is used to this day to model population growth or, in market research, to describe the diffusion or market penetration of a new product or of new technologies. For new commodities that satisfy new needs like television, compact discs or video cameras, the growth of ownership is naturally proportional both to the penetration rate already achieved and to the size of the remaining potential market, and similar arguments apply to the diffusion of new products and techniques in industry."

The application of probability models to biological experiments in the thirties of the twentieth century represents another foundation of the logistic regression model. However, it was in the first place the probit model introduced in section 2.3.3 which found its reflection in the literature. According to Cramer, economists at that time did not seem to take the logit model seriously. Only after Henri Theil 1969 generalised the bivariate or dichotomous to the multinomial logit model with more than two states of the dependent variable, the logistic regression gained its wide acceptance. In the seventies, Daniel McFadden, winner of the 2000 Nobel Prize in economics, and his collaborators finally provided a theoretical framework to the logit model linking it directly to the mathematical theory of economic choice (see McFadden 1974).

# 3 Consistency of the Maximum Likelihood Estimator

## 3.1 A Different Approach

In this section, we are going to investigate the consistency of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$. More precisely, we are going to show that $\hat{\boldsymbol{\beta}}$ converges under certain hypotheses to the real value $\boldsymbol{\beta}^0$ if the number $n$ of observations $(y_i, \boldsymbol{x}_i)$ considered in the model tends to infinity. This, of course, is a purely theoretical viewpoint since it will never be possible to include an infinity of samples in an empirical statistical analysis. Nonetheless, the consistency of an estimator is an important aspect of its nature.

Different proofs of consistency can be found in the literature (see e.g. Gourieroux and Monfort (1981) or Amemiya (1985, pp. 270-274)). All of them include or base upon the fact that the probability of the existence of the estimator $\hat{\boldsymbol{\beta}}$ approaches 1 as $n$ tends to infinity. Furthermore, they proceed on the assumption that the number $p$ of exogenous variables is fixed once and for all. In other words, $p$ is compelled to remain constant while the sample size $n$ grows.

Results presented by Mazza and Antille (1998) shall enable us to release this last aspect. We are going to assume that $p$ is variable but dependent on $n$ and to examine what relationship between $p$ and $n$ is necessary in order not to destroy the consistency of our estimator $\hat{\boldsymbol{\beta}}$. Intuition suggests that the number of observations should be larger than the number of real parameters to be estimated, i.e. $n > p$. The requirements on $p$, however, are more subtle as we will see on the following pages.

Following Mazza and Antille (1998), we now define the error function

$$E_{\boldsymbol{y},\sigma}(\boldsymbol{\beta}) := \mathbf{1}_n^{\mathrm{T}}\, H_\sigma(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}\,, \tag{3.1}$$

where $\mathbf{1}_n := (1,\ldots,1)^{\mathrm{T}} \in \mathbb{R}^n$ and $H_\sigma$ is a primitive of $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$. Considering the gradient

$$\nabla E_{\boldsymbol{y},\sigma}(\boldsymbol{\beta}) = \boldsymbol{X}^{\mathrm{T}}\sigma(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{X}^{\mathrm{T}}\big(\sigma(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{y}\big)$$

we become aware that

$$\nabla \ln L(\boldsymbol{\beta}) = -\nabla E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}(\boldsymbol{\beta})\,.$$

Any estimator $\hat{\boldsymbol{\beta}}$ which maximises the log-likelihood function $\ln L$ minimises thus simultaneously the error function $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$. This, however, is not very surprising. If we take into consideration that a possible primitive of $\sigma_{\mathrm{LR}}$ is

$H_{\sigma_{\mathrm{LR}}}(z) = \ln(1+\exp z)$, we see directly by (2.6) and (3.1) that $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}(\boldsymbol{\beta}) = -\ln L(\boldsymbol{\beta})$.

The theorem 1 provided by Mazza and Antille (1998, p. 4) based on the definition of $E_{\boldsymbol{y},\sigma}$ is therefore directly applicable to our logistic regression model. On the following pages, we shall recite this theorem and its proof in a variant slightly adjusted to our problem.[4]

**Assumptions.** We consider the following assumptions:

**M1:** $\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2 \le Cnp$ for some positive constant $C > 0$.

**M2:** If $\lambda_*$ denotes the smallest eigenvalue of the symmetric matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \in \mathbb{R}^{p\times p}$, there exists a positive constant $c > 0$ such that $\lambda_* > c\,n$ for all $n$. ◇

**Theorem 3.1.** *Assume **M1** and **M2**, let $\boldsymbol{\beta} \in \mathbb{R}^p$, and consider the random vector $\boldsymbol{y} = \sigma_{\mathrm{LR}}(\boldsymbol{X}\boldsymbol{\beta}^0) + \boldsymbol{r}$, where $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)^{\mathrm{T}}$ has mutually independent entries $r_i$ such that $\mathrm{E}(r_i) = 0$ and $\mathrm{E}(r_i^2) = s_i^2 > 0$ for all $i \in \{1,\ldots,n\}$.[5] Let $B(\boldsymbol{\beta}^0, \delta) \subset \mathbb{R}^p$ be the open ball of radius $\delta$ centered at $\boldsymbol{\beta}^0$, and let*

$$a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0) := \inf_{\boldsymbol{\zeta}\in\boldsymbol{X}B(\boldsymbol{\beta}^0,\delta)} \min_{i=1,\ldots,n} \sigma'_{\mathrm{LR}}(\zeta_i) > 0\,.$$

*Moreover, assume that $p = p(n)$ depends on $n$ such that*

$$\sqrt{\frac{p}{n}}\,\frac{1}{a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)} \xrightarrow[n\to\infty]{} 0 \qquad (3.2)$$

*for some $\delta > 0$. Then, with probability converging to 1 as $n$ tends to infinity, the absolute minimum $\hat{\boldsymbol{\beta}}$ of $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$ (i.e. the absolute maximum of $\ln L$) is the unique root of*

$$\nabla E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}(\boldsymbol{\beta}) = -\nabla \ln L(\boldsymbol{\beta})$$

*and converges in probability to the true value $\boldsymbol{\beta}^0$.* ◇

For the proof of theorem 3.1, we need the following lemma which is in fact a variant of Ortega and Rheinboldt's (1970, p. 163) lemma 6.3.4 adapted for our specific purpose.

[4] There are two main differences. First, the assumptions on the differentiability of $\sigma$ and the positivity of its derivative are omitted because both of them are automatically satisfied by $\sigma_{\mathrm{LR}}$. Furthermore, the assumption of homoscedasticity of the random variables $r_i$ is released in order to allow mutually different, but positive variances.

[5] Note that, by (2.5b), we have $s_i^2 = \sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)$. Following Mazza and Antille's original, the shorter notation $s_i^2$ shall be applied in this context.

**Lemma 3.2.** *Let $B = B(\boldsymbol{\beta}^0, \delta)$ be an open ball in $\mathbb{R}^p$ with center $\boldsymbol{\beta}^0$ and radius $\delta > 0$. Assume that $G : \bar{B} \subset \mathbb{R}^p \longrightarrow \mathbb{R}^p$ is continuous and satisfies $(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^{\mathrm{T}} G(\boldsymbol{\beta}) \leq 0$ for all $\boldsymbol{\beta} \in \partial B$, the border of $B$. Then $G$ has a root in $\bar{B}$.* $\diamond$

**Proof (of Lemma 3.2).** We consider the ball $B_0 = B(\mathbf{0}, \delta)$ and define $G_0 : \bar{B}_0 \longrightarrow \mathbb{R}^n$ by $G_0(\boldsymbol{\gamma}) := \boldsymbol{\gamma} + G(\boldsymbol{\gamma} + \boldsymbol{\beta}^0)$. Given the continuity of $G$, the function $G_0$ is also continuous. Let $\boldsymbol{\gamma} \in \partial B_0$, i.e. $\|\boldsymbol{\gamma}\|^2 = \delta^2$. Then $\boldsymbol{\gamma} + \boldsymbol{\beta}^0 \in \partial B$, and thus, for any $\lambda > 1$,

$$\begin{aligned} \boldsymbol{\gamma}^{\mathrm{T}}(\lambda\boldsymbol{\gamma} - G_0(\boldsymbol{\gamma})) &= \boldsymbol{\gamma}^{\mathrm{T}}\big((\lambda - 1)\boldsymbol{\gamma} - G(\boldsymbol{\gamma} + \boldsymbol{\beta}^0)\big) \\ &= \underbrace{(\lambda - 1)\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma}}_{=(\lambda-1)\|\boldsymbol{\gamma}\|^2>0} - \underbrace{\boldsymbol{\gamma}^{\mathrm{T}} G(\boldsymbol{\gamma} + \boldsymbol{\beta}^0)}_{\leq 0 \text{ by assumption}} > 0\,. \end{aligned} \tag{3.3}$$

We now want to show, that $G_0$ has a fixed point $\hat{\boldsymbol{\gamma}} \in \bar{B}_0$, i.e. $G_0(\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\gamma}}$. This result would finally mean that $G(\hat{\boldsymbol{\gamma}}+\boldsymbol{\beta}^0) = 0$ and, therefore, $\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\gamma}}+\boldsymbol{\beta}^0 \in \bar{B}$ would be a root of $G$.

Assume that $G_0$ has no fixed point in $\bar{B}_0$. Then the mapping $\hat{G}(\boldsymbol{\gamma}) := \delta\,\big(G_0(\boldsymbol{\gamma}) - \boldsymbol{\gamma}\big)/\big\|G_0(\boldsymbol{\gamma}) - \boldsymbol{\gamma}\big\|$ is well-defined and continuous on $\bar{B}_0$, and $\|\hat{G}(\boldsymbol{\gamma})\| = \delta$ for any $\boldsymbol{\gamma} \in \bar{B}_0$. According to the *Brouwer Fixed-Point Theorem*[6], $\hat{G}$ has a fixed point $\boldsymbol{\gamma}^*$ in $\bar{B}_0$ and $\|\boldsymbol{\gamma}^*\| = \|\hat{G}(\boldsymbol{\gamma}^*)\| = \delta$. As

$$\boldsymbol{\gamma}^* = \hat{G}(\boldsymbol{\gamma}^*) = \delta \cdot \frac{G_0(\boldsymbol{\gamma}^*) - \boldsymbol{\gamma}^*}{\big\|G_0(\boldsymbol{\gamma}^*) - \boldsymbol{\gamma}^*\big\|}\,,$$

we have

$$G_0(\boldsymbol{\gamma}^*) = \frac{\boldsymbol{\gamma}^*}{\delta}\,\big\|G_0(\boldsymbol{\gamma}^*) - \boldsymbol{\gamma}^*\big\| + \boldsymbol{\gamma}^* = \underbrace{\big(1 + \tfrac{1}{\delta}\,\big\|G_0(\boldsymbol{\gamma}^*) - \boldsymbol{\gamma}^*\big\|\big)}_{=:\lambda^*>1}\,\boldsymbol{\gamma}^* = \lambda^*\boldsymbol{\gamma}^*\,.$$

However, we thus get ${\boldsymbol{\gamma}^*}^{\mathrm{T}}(\lambda^*\boldsymbol{\gamma}^* - G_0(\boldsymbol{\gamma}^*)) = 0$ which contradicts condition (3.3). ■

**Remark.** For the proof of lemma 3.2, the condition $(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^{\mathrm{T}} G(\boldsymbol{\beta}) \leq 0$ is not essential. The same result holds when $(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^{\mathrm{T}} G(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta} \in \partial B$. In the present case, however, the previous version suits better our needs. $\diamond$

**Proof (of Theorem 3.1).** Considering the function $G(\boldsymbol{\beta}) := -\nabla E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}(\boldsymbol{\beta})$, we are going to show that there exists a ball $\bar{B}(\boldsymbol{\beta}^0, \delta) \subset \mathbb{R}^p$, which – with

[6] The Brouwer Fixed-Point Theorem reads as follows (see Ortega and Rheinboldt 1970, p. 161): *Every continuous mapping $G : \bar{C} \longrightarrow \bar{C}$, where $\bar{C}$ is a compact, convex set in $\mathbb{R}^p$, has a fixed point in $\bar{C}$.*

probability converging to 1 as $n$ tends to infinity – contains a root $\hat{\boldsymbol{\beta}}$ of $G$ for an arbitrary small $\delta > 0$.

Let $G(\boldsymbol{\beta})_j$ denote the $j$th component of the vector $G(\boldsymbol{\beta})$. We have

$$\begin{aligned} G(\boldsymbol{\beta})_j &= \Big(\boldsymbol{X}^{\mathrm{T}}\big(\boldsymbol{y} - \sigma(\boldsymbol{X}\boldsymbol{\beta})\big)\Big)_j \\ &= \sum_{i=1}^{n} x_{ij}\big(y_i - \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\big) \\ &= \sum_{i=1}^{n} x_{ij}\big(y_i - \sigma_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0 + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma})\big) \end{aligned}$$

where $\boldsymbol{\gamma} := \boldsymbol{\beta} - \boldsymbol{\beta}^0$, and by defining $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}} := \boldsymbol{X}\boldsymbol{\beta}^0$ we get with the model assumption (2.4) the expression

$$G(\boldsymbol{\beta})_j = \sum_{i=1}^{n} x_{ij}\big(\sigma_{\mathrm{LR}}(\eta_i) + r_i - \sigma_{\mathrm{LR}}(\eta_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma})\big)$$

such that, as a result of the mean value theorem, there exists some $\xi_i = \eta_i + \alpha_i\,\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}$ with $\alpha_i \in ]0,1[$ satisfying

$$G(\boldsymbol{\beta})_j = \sum_{i=1}^{n} x_{ij}\big(r_i - \sigma'_{\mathrm{LR}}(\xi_i)\cdot\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\big)\,.$$

Thanks to the previous lemma, we only need to prove that $\boldsymbol{\gamma}^{\mathrm{T}}G(\boldsymbol{\beta}) \le 0$ for all $\boldsymbol{\gamma} = \boldsymbol{\beta} - \boldsymbol{\beta}^0$ with $\|\boldsymbol{\gamma}\| = \delta$. We thus consider the expression

$$\begin{aligned} \boldsymbol{\gamma}^{\mathrm{T}}G(\boldsymbol{\beta}) &= \sum_{j=1}^{p}\gamma_j\,G(\boldsymbol{\beta})_j = \sum_{j=1}^{p}\gamma_j\left(\sum_{i=1}^{n} x_{ij}\big(r_i - \sigma'_{\mathrm{LR}}(\xi_i)\cdot\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\big)\right) \\ &= \sum_{i=1}^{n} r_i\left(\sum_{j=1}^{p}\gamma_j\,x_{ij}\right) - \sum_{i=1}^{n}\left(\sum_{j=1}^{p}\gamma_j\,x_{ij}\right)\sigma'_{\mathrm{LR}}(\xi_i)\,\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma} \\ &= \underbrace{\sum_{i=1}^{n} r_i\,\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}}_{=:A_1} - \underbrace{\sum_{i=1}^{n}\big(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\big)^2\sigma'_{\mathrm{LR}}(\xi_i)}_{=:A_2}\,. \end{aligned}$$

Using the Cauchy-Schwarz inequality[7], we get an upper boundary for $|A_1|$ by

$$|A_1| = \left|\sum_{i=1}^{n} r_i\,\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\right| = \left|\left(\sum_{i=1}^{n} r_i\,\boldsymbol{x}_i\right)^{\mathrm{T}}\boldsymbol{\gamma}\right| \le \left\|\sum_{i=1}^{n} r_i\,\boldsymbol{x}_i\right\|\cdot\underbrace{\|\boldsymbol{\gamma}\|}_{=\delta}\,. \qquad (3.4)$$

[7] See e.g. Ortega and Rheinboldt (1970, p. 39).

If we examine the second moment of $\|\sum_{i=1}^{n} r_i\, \boldsymbol{x}_i\|$, we obtain

$$\begin{aligned}
\mathrm{E}\left(\left\|\sum_{i=1}^{n} r_i\, \boldsymbol{x}_i\right\|^2\right) &= \mathrm{E}\left(\sum_{j=1}^{p}\left(\sum_{i=1}^{n} r_i\, x_{ij}\right)^2\right) \\
&= \mathrm{E}\left(\sum_{j=1}^{p}\left(\sum_{i=1}^{n} r_i^2\, x_{ij}^2 + \sum_{i=1}^{n}\sum_{\substack{k=1\\k\neq i}}^{n} r_i\, r_k\, x_{ij}\, x_{kj}\right)\right) \\
&= \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \mathrm{E}(r_i^2)\, x_{ij}^2 + \sum_{i=1}^{n}\sum_{\substack{k=1\\k\neq i}}^{n} \mathrm{E}(r_i\, r_k)\, x_{ij}\, x_{kj}\right)
\end{aligned}$$

and, as the random variables $r_i$ are mutually independent and their expectations are zero, we get $\mathrm{E}(r_i\, r_k) = \mathrm{E}(r_i)\,\mathrm{E}(r_k) = 0$ and thus, as $s_i^2 = \sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0) \leq \frac{1}{4}$ for all $i$,

$$\mathrm{E}\left(\left\|\sum_{i=1}^{n} r_i\, \boldsymbol{x}_i\right\|^2\right) = \sum_{i=1}^{n} \underbrace{\mathrm{E}(r_i^2)}_{=s_i^2} \sum_{j=1}^{p} x_{ij}^2 \leq \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2 \overset{\mathbf{M1}}{\leq} \frac{Cnp}{4} \tag{3.5}$$

for some positive constant $C > 0$. We see from (3.4) and (3.5) that $\mathrm{E}(A_1^2) \leq \delta^2 Cnp/4$. Using the Tchebychev inequality (see Feller 1968, p. 233), we have

$$\begin{aligned}
& \mathrm{P}(|A_1| \geq t) \leq t^{-2}\mathrm{E}(A_1^2) \leq t^{-2}\delta^2 Cnp/4 && \forall\, t > 0 \\
\iff\quad & \mathrm{P}(|A_1| < t) \geq 1 - \underbrace{t^{-2}\delta^2 Cnp/4}_{=:\varepsilon} && \forall\, t > 0 \\
\iff\quad & \mathrm{P}\left(|A_1| < \frac{\delta\sqrt{Cnp}}{2\sqrt{\varepsilon}}\right) \geq 1 - \varepsilon && \forall\, \varepsilon > 0\,.
\end{aligned}$$

Defining $C^* := \sqrt{C}/2$, it is obvious that

$$\mathrm{P}\left(A_1 \leq \delta\, C^*\sqrt{\frac{np}{\varepsilon}}\right) \geq \mathrm{P}\left(|A_1| < \frac{\delta\, C^*\sqrt{np}}{\sqrt{\varepsilon}}\right)\,.$$

If we set $n_\varepsilon := n/\varepsilon$ for any given $\varepsilon > 0$, we obtain

$$\mathrm{P}(A_1 \leq \delta\, C^*\sqrt{n_\varepsilon p}) \geq 1 - \varepsilon\,,$$

so for all $\varepsilon > 0$, there exists a number $n_\varepsilon \in \mathbb{N}$ such that for any $n \geq n_\varepsilon$ and any given $C^* > 0$, we have

$$\mathrm{P}(A_1 \leq \delta\, C^*\sqrt{np}) \geq 1 - \varepsilon\,. \tag{3.6}$$

Now, let us turn to the examination of $A_2$. Let $Z := \{\xi \in \mathbb{R}^n \,|\, \xi_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0 + \alpha_i\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma},\ \alpha_i \in\, ]0,1[\}$.

**Affirmation.** *For any vector* $\boldsymbol{\xi} \in Z$*, we have*

$$\sigma'_{\mathrm{LR}}(\xi_i) \geq a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0) = \inf_{\boldsymbol{\zeta} \in \boldsymbol{X}B(\boldsymbol{\beta}^0,\delta)} \min_{i=1,\ldots,n} \sigma'_{\mathrm{LR}}(\zeta_i) \qquad \forall\, i \in \{1, \ldots, n\}\,.$$

$\diamond$

**Verification.** Assume that there is some $\boldsymbol{\xi}^* \in Z$ such that for an $i^* \in \{1, \ldots, n\}$ we have

$$\sigma'_{\mathrm{LR}}(\xi^*_{i^*}) < \inf_{\boldsymbol{\zeta} \in \boldsymbol{X}B(\boldsymbol{\beta}^0,\delta)} \min_{i=1,\ldots,n} \sigma'_{\mathrm{LR}}(\zeta_i)\,.$$

As $\boldsymbol{\xi}^* \in Z$, there are numbers $\alpha_i \in ]0, 1[$ $(i = 1, \ldots, n)$ such that $\xi^*_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0 + \alpha_i\, \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}$. We define $\boldsymbol{\zeta}^* := \boldsymbol{X}\boldsymbol{\beta}^0 + \alpha_{i^*}\boldsymbol{X}\boldsymbol{\gamma}$. Then, $\boldsymbol{\zeta}^* := \boldsymbol{X}(\boldsymbol{\beta}^0 + \alpha_{i^*}\boldsymbol{\gamma})$ and, as $\|\alpha^*_i\boldsymbol{\gamma}\| < \|\boldsymbol{\gamma}\| = \delta$, we have $\boldsymbol{\zeta}^* \in \boldsymbol{X}B(\boldsymbol{\beta}^0, \delta)$. But

$$\min_{i=1,\ldots,n} \sigma'_{\mathrm{LR}}(\zeta^*_i) \leq \sigma'_{\mathrm{LR}}(\zeta^*_{i^*}) = \sigma'_{\mathrm{LR}}(\boldsymbol{x}_{i^*}^{\mathrm{T}}\boldsymbol{\beta}^0 + \alpha_{i^*}\, \boldsymbol{x}_{i^*}^{\mathrm{T}}\boldsymbol{\gamma}) = \sigma'_{\mathrm{LR}}(\xi^*_{i^*})\,,$$

which is a contradiction. $\square$

We therefore get

$$\begin{aligned} A_2 &= \sum_{i=1}^{n} \left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)^2 \sigma'_{\mathrm{LR}}(\xi_i) \\ &\geq a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0) \sum_{i=1}^{n} \left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\right)^2 = a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, \|\boldsymbol{X}\boldsymbol{\gamma}\|^2\,. \end{aligned} \tag{3.7}$$

Looking at $\|\boldsymbol{X}\boldsymbol{\gamma}\|^2 = (\boldsymbol{X}\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\gamma} = \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\gamma}$, we observe that $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is symmetric and – as a result of **M2** – positive definite. Therefore, there exists a set of orthonormal eigenvectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\}$ with attributed eigenvalues $\{\lambda_1, \ldots, \lambda_p\}$ forming a base of $\mathbb{R}^p$. If we now write $\boldsymbol{\gamma} = \sum_{i=1}^{p} \gamma^*_i \boldsymbol{v}_i$, we obtain

$$\begin{aligned} \|\boldsymbol{X}\boldsymbol{\gamma}\|^2 &= \left(\sum_{i=1}^{p} \gamma^*_i \boldsymbol{v}_i^{\mathrm{T}}\right) \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \left(\sum_{j=1}^{p} \gamma^*_j \boldsymbol{v}_j\right) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} \gamma^*_i\gamma^*_j \boldsymbol{v}_i^{\mathrm{T}} \underbrace{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{v}_j}_{=\lambda_j\boldsymbol{v}_j} = \sum_{i=1}^{p}\sum_{j=1}^{p} \gamma^*_i\gamma^*_j \lambda_j \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_j \\ &\geq \lambda_* \sum_{i=1}^{p}\sum_{j=1}^{p} \gamma^*_i\gamma^*_j \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_j = \lambda_* \sum_{i=1}^{p} \gamma^*_i \boldsymbol{v}_i^{\mathrm{T}} \sum_{j=1}^{p} \gamma^*_j \boldsymbol{v}_j \\ &= \lambda_* \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma} = \lambda_* \|\boldsymbol{\gamma}\|^2 = \lambda_*\, \delta^2\,, \end{aligned} \tag{3.8}$$

where $\lambda_*$ denotes the smallest eigenvalue of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. We thus see that $A_2 \geq a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, \lambda_*\, \delta^2$ and, because of **M2**, there is a positive constant $c > 0$ such that $A_2 \geq a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, c\, n\, \delta^2$. When combining this result with (3.6), we get that, for any $\varepsilon > 0$, there is a number $n_\varepsilon$ such that

$$\mathrm{P}(A_1 - A_2 \leq \delta\, C^* \sqrt{np} - a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, c\, n\, \delta^2) \geq 1 - \varepsilon$$

for all $n \geq n_\varepsilon$. As we are interested in the case where $\boldsymbol{\gamma}^{\mathrm{T}} G(\boldsymbol{\beta}) = A_1 - A_2 \leq 0$, we consider the inequality

$$\begin{aligned}
& \delta\, C^* \sqrt{np} - a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, c\, n\, \delta^2 \leq 0 \\
\Longleftrightarrow \quad & C^* \sqrt{np} \leq a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, c\, n\, \delta \\
\Longleftrightarrow \quad & \frac{C^*}{c} \sqrt{\frac{p}{n}} \frac{1}{a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)} \leq \delta\,. \qquad (3.9)
\end{aligned}$$

As a consequence of the assumption (3.2), there is for any $\delta > 0$ a number $n_\delta$ such that (3.9) holds for all $n \geq n_\delta$. We therefore know that, for any positive $\delta$ and $\varepsilon$, we have

$$\mathrm{P}\big(\boldsymbol{\gamma}^{\mathrm{T}} G(\boldsymbol{\beta}) \leq 0\big) \geq 1 - \varepsilon$$

for all $n \geq \max(n_\delta, n_\varepsilon)$ and $\boldsymbol{\gamma} \in \partial B(\boldsymbol{0}, \delta)$. Furthermore, we remember from (2.9) that

$$\boldsymbol{u}^{\mathrm{T}}\, \mathbf{H}_{\ln L}(\boldsymbol{\beta})\, \boldsymbol{u} = -\sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u})^2\, \sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})\,,$$

from where we obtain in analogy to (3.7) that

$$\begin{aligned}
\boldsymbol{u}^{\mathrm{T}}\, \mathbf{H}_{E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}}(\boldsymbol{\beta})\, \boldsymbol{u} &= \sum_{i=1}^{n} (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{u})^2\, \sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}) \\
&\geq a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, c\, n\, \|\boldsymbol{u}\|^2 \geq \inf_{n \in \mathbb{N}} a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)\, c\, n\, \|\boldsymbol{u}\|^2 > 0
\end{aligned}$$

if $\boldsymbol{u} \neq \boldsymbol{0}$. The last inequality is a result of the assumption (3.2): If $n\, a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)$ converged to 0 as $n$ tends to infinity, we would have

$$\sqrt{\frac{p}{n}} \frac{1}{a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)} = \frac{\sqrt{p\, n}}{n\, a^{\delta}_{pn}(\boldsymbol{X}, \boldsymbol{\beta}^0)} \xrightarrow[n \to \infty]{} \infty$$

which is a contradiction. The Hessian matrix $\mathbf{H}_{E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}}(\boldsymbol{\beta})$ is thus strictly positive definite and $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$ is strictly convex in $\bar{B}(\boldsymbol{\beta}^0, \delta)$. Therefore, under the hypotheses **M1** and **M2**, the absolute minimum $\hat{\boldsymbol{\beta}}$ of $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$ (the absolute maximum of $\ln L$) is not only unique but converges in probability to the true value $\boldsymbol{\beta}^0$ as $n$ tends to infinity. ∎

**Remark.** If we choose $\delta$ such that

$$\delta = \delta_n := \frac{C^*}{c}\sqrt{\frac{p}{n}}\frac{1}{a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)}\,,$$

we notice that, with probability converging to 1 as $n \longrightarrow \infty$,

$$\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|^2 = \mathcal{O}_p\left(\frac{p}{n}\frac{1}{\left(a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)\right)^2}\right),$$

as $\hat{\boldsymbol{\beta}} \in \bar{B}(\boldsymbol{\beta}^0, \delta_n)$. In other words, $\frac{n}{p}\left(a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)\right)^2\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|^2$ is bounded in probability. $\diamond$

## 3.2 Condition on the Relationship between $p$ and $n$

As mentioned before, Mazza and Antille (1998) originally developed this theorem for an arbitrary strictly increasing and differentiable function $\sigma$. The choice of one precise element $\sigma_{\mathrm{LR}}$ from this set of functions gives us the possibility to study more in detail what condition (3.2) really implies on the relationship between $p$ and $n$. In other words, we can attempt to give a condition on $p(n)$ such that (3.2) is a consequence.

For this purpose we assume that there is some positive constant $D > 0$ such that the true value $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ satisfies the condition

$$\sup_{p\in\mathbb{N}} \max_{j=1,\ldots,p} \left(\beta_j^0\right)^2 < D\,,$$

i.e. the norm of $\boldsymbol{\beta}^0$ is bounded for all $p$. We suppose furthermore that $\|\boldsymbol{x}_i\|^2 \le D^* p$ for some positive constant $D^* > 0$ and for all $i = 1, \ldots, n$. This requirement holds for instance when the elements of $\boldsymbol{X}$ are uniformly bounded.

For every $\delta > 0$ and all $\boldsymbol{\beta} \in B(\boldsymbol{\beta}^0, \delta)$ we thus obtain by Cauchy-Schwarz that

$$\begin{aligned}\left|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right| &\le \|\boldsymbol{x}_i\|\,\|\boldsymbol{\beta}\| < \|\boldsymbol{x}_i\|\,\left(\left\|\boldsymbol{\beta}^0\right\| + \delta\right)\\ &\le \sqrt{D^* p}\,\left(\sqrt{Dp} + \delta\right) = p\sqrt{D^* D} + \delta\sqrt{D^* p}\,.\end{aligned}$$

Hence, while $p$ and $n$ increase, there is a constant $\bar{D} := \sqrt{D^*}(\sqrt{D} + \delta) > 0$ such that $\left|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right| \le \bar{D}p$. As a result, $|\zeta_i| \le \bar{D}p$ for all $\boldsymbol{\zeta} \in \boldsymbol{X}B(\boldsymbol{\beta}^0, \delta)$ and for all $i \in \{1, \ldots, n\}$. These calculations allow us to state the inequality

$$a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0) = \inf_{\boldsymbol{\zeta}\in\boldsymbol{X}B(\boldsymbol{\beta}^0,\delta)} \min_{i=1,\ldots,n} \sigma'_{\mathrm{LR}}(\zeta_i) \ge \inf_{|z|\le\bar{D}p} \sigma'_{\mathrm{LR}}(z)\,. \tag{3.10}$$

As $\sigma'_{\mathrm{LR}}(z) = (\exp z)/(1+\exp z)^2$ is an even function having at $z=0$ its maximum (see figure 1), we get

$$\inf_{|z|\le \bar{D}p} \sigma'_{\mathrm{LR}}(z) = \sigma'_{\mathrm{LR}}(\bar{D}p) = \frac{\exp \bar{D}p}{(1+\exp \bar{D}p)^2}\,.$$

Consequently, the inequality (3.10) can be rewritten as

$$\frac{1}{\frac{\exp \bar{D}p}{(1+\exp \bar{D}p)^2}} \ge \frac{1}{a^{\delta}_{pn}(\boldsymbol{X},\boldsymbol{\beta}^0)} \quad \forall\, n\, \forall\, p$$

which gives us the implication

$$I := \sqrt{\frac{p}{n}}\,\frac{(1+\exp \bar{D}p)^2}{\exp \bar{D}p} \xrightarrow[n\to\infty]{} 0 \quad \Longrightarrow \quad \sqrt{\frac{p}{n}}\,\frac{1}{a^{\delta}_{pn}(\boldsymbol{X},\boldsymbol{\beta}^0)} \xrightarrow[n\to\infty]{} 0\,.$$

In other words, our requirement (3.2) on $p(n)$ holds when the left side of this implication is true. Let us examine

$$\begin{aligned} I &= \sqrt{\frac{p}{n}}\,\frac{(1+\exp \bar{D}p)^2}{\exp \bar{D}p} = \sqrt{\frac{p}{n}}\,\frac{1+\exp \bar{D}p}{\exp \bar{D}p}\,(1+\exp \bar{D}p) \\ &= \sqrt{\frac{p}{n}}\,\underbrace{\big(\exp(-\bar{D}p)+1\big)}_{\le \exp(-\bar{D})+1=:c_1}(1+\exp \bar{D}p) \le c_1\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{p}{n}}\,\exp \bar{D}p\right)\,. \end{aligned}$$

**Affirmation.** *Let $C$ be a constant such that $C < 1/(2\bar{D})$, and suppose that $p(n) \le C \ln n$. Then* $\lim_{n\to\infty} I = 0$. ◇

**Verification.** We have

$$\sqrt{\frac{p}{n}}\,\exp \bar{D}p \le \sqrt{\frac{C\ln n}{n}}\,\exp\big(C\bar{D}\ln n\big) = n^{C\bar{D}}\sqrt{\frac{C\ln n}{n}} = \sqrt{\frac{C\ln n}{n^{1-2C\bar{D}}}}$$

which converges to 0 as $n$ tends to infinity if and only if $1-2C\bar{D} > 0$. However, this last inequality is equivalent to $C < 1/(2\bar{D})$. Moreover, this same argument shows that $\lim_{n\to\infty}\sqrt{p/n} = 0$, and thus $\lim_{n\to\infty} I = 0$. ■

## 3.3 Reformulation, Assessment, and Comparison

We shall now resume the results of the previous two sections in the following theorem.

**Assumptions.** Consider the following assumptions:

**B1:** There exists a positive constant $D^* > 0$ such that $\|\boldsymbol{x}_i\|^2 \leq D^* p$ for all $i \in \{1, \ldots, n\}$.

**B2:** There exists a positive constant $c > 0$ such that $\lambda_* > c\,n$ for all $n$, where $\lambda_*$ denotes the smallest eigenvalue of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.

**B3:** There exists a positive constant $D > 0$ such that

$$\sup_{p \in \mathbb{N}} \max_{j=1,\ldots,p} \left(\beta_j^0\right)^2 < D\,.$$

**B4:** For an arbitrary $\delta > 0$ there exists a constant $C < 1/\big(2\sqrt{D^*}(\sqrt{D}+\delta)\big)$ such that $p(n) \leq C \ln n$. $\diamond$

**Theorem 3.3.** *Assume **B1**, **B2**, **B3**, and **B4**. Then the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ exists almost surely as $n$ tends to infinity, and $\hat{\boldsymbol{\beta}}$ converges to the true value $\boldsymbol{\beta}^0$.* $\diamond$

**Remark.** Note that **B1** implies **M1** while **B2** equals **M2**. $\diamond$

As an assessment of this result, we shall finally attempt to compare it to the theorem on the existence and strong consistency of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ formulated by Gourieroux and Monfort (1981). In contrast to other authors, they were able to prove these properties under comparatively weak assumptions, as they themselves affirm.

For the purpose of such a comparison, we return to the case where $p$ is constant. As a consequence of this restriction, the assumptions **B3** and **B4** are automatically satisfied, at least if $n$ is sufficiently large. Moreover, **B1** reduces to

**$\widetilde{\mathbf{B1}}$:** There exists a positive constant $M > 0$ such that $\|\boldsymbol{x}_i\|^2 \leq M$ for all $i \in \{1, \ldots, n\}$.

Let us now turn to the article of Gourieroux and Monfort. They make the following assumptions:

**G1:** The exogenous variables are uniformly bounded, i.e. there exists a positive constant $M_0$ such that $|x_{ij}| \leq M_0$ for all $i \in \{1, \ldots, n\}$ and all $j \in \{1, \ldots, p\}$.

**G2:** Let $\lambda_{1n}$ and $\lambda_{pn}$ be respectively the smallest and the largest eigenvalue of $-\mathbf{H}_{\ln L}(\boldsymbol{\beta}^0) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{\beta}^0)\boldsymbol{X}$, the diagonal matrix $\boldsymbol{D}(\boldsymbol{\beta}^0)$ being defined as in (2.8). There exists a constant $M_1$ such that $\lambda_{pn}/\lambda_{1n} < M_1$ for all $n$.

**Theorem 3.4 (Gourieroux and Monfort).** *If **G1** and **G2** are satisfied, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ exists almost surely as n goes to infinity, and $\hat{\boldsymbol{\beta}}$ converges almost surely to the true value $\boldsymbol{\beta}^0$ if and only if*

$$\lim_{n\to\infty} \lambda_{1n} = +\infty\,. \qquad \diamond$$

By the norm-equivalence theorem[8] with respect to the Euclidean and the maximum norm, we immediately get the equivalence of $\widehat{\mathbf{B1}}$ and **G1**. As, furthermore, **B2** implies that $\lambda_*$ tends to infinity as $n$ increases, it is interesting to check whether this fact has an impact on the limit of $\lambda_{1n}$.

**Affirmation.** *Assume $\widehat{\boldsymbol{B1}}$. Then*

$$\lim_{n\to\infty} \lambda_* = +\infty \quad\Longleftrightarrow\quad \lim_{n\to\infty} \lambda_{1n} = +\infty\,. \qquad \diamond$$

**Verification.** Let $\boldsymbol{\Sigma} = (s_{ij}) \in \mathbb{R}^{n\times n}$ denote the diagonal matrix defined by

$$s_{ij} = \begin{cases} \sqrt{\sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{3.11}$$

Note that $\boldsymbol{\Sigma}^2 = \boldsymbol{D}(\boldsymbol{\beta}^0)$ and that $s_{ii} = s_i = \sqrt{\mathrm{Var}(r_i)}$. On one hand, we have $\sigma'_{\mathrm{LR}}(z) \le 1/4\ \forall\, z$ and thus $s_i^2 \le 1/4$. On the other hand,

$$|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0| \le \|\boldsymbol{x_i}\|\,\|\boldsymbol{\beta}^0\| \overset{\widehat{\mathbf{B1}}}{\le} \sqrt{M}\,\|\boldsymbol{\beta}^0\|$$

such that $s_i^2 \ge \sigma'_{\mathrm{LR}}(\sqrt{M}\,\|\boldsymbol{\beta}^0\|) =: c_s > 0$ for all $i$. Consequently, the real-valued function $\|\cdot\|_s$ defined as

$$\|\boldsymbol{w}\|_s := \sqrt{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}^2\boldsymbol{w}} = \sqrt{\sum_{i=1}^{n} s_i^2\, w_i^2}$$

is a norm on $\mathbb{R}^n$, and for all $\boldsymbol{w} \in \mathbb{R}^n$ we have the inequality

$$c_s\,\|\boldsymbol{w}\|^2 \le \|\boldsymbol{w}\|_s^2 \le \frac{\|\boldsymbol{w}\|^2}{4}\,. \tag{3.12}$$

Let $S^{p-1}$ denote the unit sphere in $\mathbb{R}^p$, i.e. $S^{p-1} := \{\boldsymbol{v} \in \mathbb{R}^p \mid \|\boldsymbol{v}\| = 1\}$. By the Courant-Fischer-Weyl minimax principle[9] we have

$$\lambda_* = \min_{\boldsymbol{v}\in S^{p-1}} \|\boldsymbol{X}\boldsymbol{v}\|^2 \quad\text{and}\quad \lambda_{1n} = \min_{\boldsymbol{v}\in S^{p-1}} \|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{v}\|^2 = \min_{\boldsymbol{v}\in S^{p-1}} \|\boldsymbol{X}\boldsymbol{v}\|_s^2\,.$$

[8] See e.g. Ortega and Rheinboldt (1970, p. 39).

[9] See e.g. Bhatia (1997, p. 58).

Choose $\boldsymbol{v}_* \in S^{p-1}$ such that $\lambda_* = \|\boldsymbol{X}\boldsymbol{v}_*\|^2$. By (3.12) we have $\lambda_* = \|\boldsymbol{X}\boldsymbol{v}_*\|^2 \geq 4\,\|\boldsymbol{X}\boldsymbol{v}_*\|_s^2 \geq 4\,\lambda_{1n}$. Conversely, if we choose $\boldsymbol{v}_{**} \in S^{p-1}$ such that $\lambda_{1n} = \|\boldsymbol{X}\boldsymbol{v}_{**}\|_s^2$, we get $\lambda_{1n} = \|\boldsymbol{X}\boldsymbol{v}_{**}\|_s^2 \geq c_s\,\|\boldsymbol{X}\boldsymbol{v}_*\|^2 \geq c_s\,\lambda_*$. In consequence, the inequality

$$c_s\,\lambda_* \leq \lambda_{1n} \leq \tfrac{1}{4}\,\lambda_* \tag{3.13}$$

holds for all $n$, from where we get the affirmed equivalence. ■

This result shows that **B2** implies that $\lambda_{1n}$ also goes to infinity as $n$ increases. The opposite, however, is not always true: Assume for instance that $\lambda_{1n} = c_l \ln n$ where $c_l > 0$ is an arbitrary positive constant. By (3.13), we get thus

$$4\,c_l \ln n \leq \lambda_* \leq \frac{c_l}{c_s} \ln n\,.$$

So while $\lim_{n\to\infty} \lambda_* = \infty$, there is no positive constant $c$ such that $\lambda_* > c\,n$ for all $n$.

The assumption $\lim_{n\to\infty} \lambda_{1n} = \infty$ of the theorem 3.4 is therefore less restrictive than **B2**. Conversely, Gourieroux and Monfort (1981) need the supplementary hypothesis **G2** to prove the consistency of $\hat{\boldsymbol{\beta}}$. On the other hand, **G2** additionally ensures that $\lim_{n\to\infty} \lambda_{1n} = \infty$ is not only sufficient but also necessary for the consistency of $\hat{\boldsymbol{\beta}}$.

# 4 Asymptotic Normality of the Maximum Likelihood Estimator

In the previous section, we have studied the consistency of the maximum likelihood estimator in the logistic regression model. We were able to prove that, under certain hypotheses, the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} \in \mathbb{R}^p$ converges in probability to the true value $\boldsymbol{\beta}^0$.

A second important issue to be studied is the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$. More precisely, we are going to show that the normalised linear combinations of the components of $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0$ converge in distribution to random variables having a normal law with zero expectation.

The base structure of the following theorem is taken from an article by Antille and Rech (1999). Some extensions had to be made in respect of the heteroscedasticity of the components of $\boldsymbol{r}$. On the other hand, the specific knowledge of $\sigma_{\mathrm{LR}}$ gave rise to several simplifications. In contrast to section 3, we now assume that the number $p$ of exogenous variables is constant and thus independent of the sample size $n$.

**Assumptions.**

**A1:** The exogenous variables $\boldsymbol{X}$ are uniformly bounded, i.e. there exists a positive constant $M$ such that $|x_{ij}| \leq M$ for all $i \in \{1, \ldots, n\}$ and all $j \in \{1, \ldots, p\}$.

**A2:** If $\lambda_*$ denotes the smallest eigenvalue of the symmetric matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \in \mathbb{R}^{p\times p}$, there exists a positive constant $c > 0$ such that $\lambda_* > c\,n$ for all $n$. ◇

**Theorem 4.1.** *Assume **A1** and **A2** and let $\boldsymbol{D}(\boldsymbol{\beta}^0)$ and $\boldsymbol{\Sigma}$ be defined as in* (2.8) *and* (3.11) *respectively. Then*

$$\frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{\beta}^0)\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|} \xrightarrow[n\to\infty]{\mathrm{L}} N(0,1)$$

*for all $\boldsymbol{e} \in \mathbb{R}^p$ with $\|\boldsymbol{e}\| = 1$.* ◇

**Proof.** From theorem 3.1 we know that, with a probability approaching 1 as $n \to \infty$, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is the unique solution of

$$-\nabla \ln L(\boldsymbol{\beta}) = \boldsymbol{X}^{\mathrm{T}}\big(\sigma_{\mathrm{LR}}(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{y}\big) = \boldsymbol{0}\,.$$

To simplify the notation we will thus assume that, for all $n$,

$$\boldsymbol{X}^{\mathrm{T}}\big(\sigma_{\mathrm{LR}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}) - \boldsymbol{y}\big) = \boldsymbol{0}\,. \tag{4.1}$$

This condition can be reformulated as

$$\begin{aligned}
& \boldsymbol{X}^{\mathrm{T}}\Big(\sigma_{\mathrm{LR}}\big(\boldsymbol{X}(\boldsymbol{\beta}^0+\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\big)-\boldsymbol{y}\Big)=\boldsymbol{0} \\
\iff \quad & \sum_{i=1}^n x_{ij}\Big(\sigma_{\mathrm{LR}}\big(\underbrace{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0}_{=:\gamma_i^0}+\underbrace{\boldsymbol{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)}_{=:\hat{\varepsilon}_i}\big)-y_i\Big)=0 \qquad \forall\, j \\
\iff \quad & \sum_{i=1}^n x_{ij}\big(\sigma_{\mathrm{LR}}(\gamma_i^0+\hat{\varepsilon}_i)-y_i\big)=0 \qquad \forall\, j\,.
\end{aligned}$$

In application of the mean value theorem, we can write $\sigma_{\mathrm{LR}}(\gamma_i^0+\hat{\varepsilon}_i) = \sigma'_{\mathrm{LR}}(\gamma_i^0)+\hat{\varepsilon}_i\,\sigma_{\mathrm{LR}}(\xi_i)$ where $\xi_i=\gamma_i^0+\alpha_i\,\hat{\varepsilon}_i$ with $\alpha_i\in\,]0,1[$. This gives us the equivalent condition

$$\iff \quad \sum_{i=1}^n x_{ij}\big(\hat{\varepsilon}_i\,\sigma'_{\mathrm{LR}}(\xi_i)+\underbrace{\sigma_{\mathrm{LR}}(\gamma_i^0)-y_i}_{=-r_i}\big)=0 \qquad \forall\, j\,.$$

Let $\boldsymbol{e}\in\mathbb{R}^p$ with $\|\boldsymbol{e}\|=1$. We thus have

$$\begin{aligned}
0 &= \sum_{j=1}^p \frac{e_j\sum_{i=1}^n x_{ij}\big(\hat{\varepsilon}_i\,\sigma'_{\mathrm{LR}}(\xi_i)-r_i\big)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|} \\
&= \underbrace{-\sum_{j=1}^p \frac{e_j\sum_{i=1}^n x_{ij}\,r_i}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|}}_{=:N_1} + \underbrace{\sum_{j=1}^p \frac{e_j\sum_{i=1}^n x_{ij}\,\hat{\varepsilon}_i\,\big(\sigma'_{\mathrm{LR}}(\xi_i)-\sigma'_{\mathrm{LR}}(\gamma_i^0)\big)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|}}_{=:N_2} \\
&\quad \underbrace{+\sum_{j=1}^p \frac{e_j\sum_{i=1}^n x_{ij}\,\hat{\varepsilon}_i\,\sigma'_{\mathrm{LR}}(\gamma_i^0)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|}}_{=:N_3}\,.
\end{aligned} \tag{4.2}$$

Let us first show that $N_2 \xrightarrow[n\to\infty]{P} 0$: We have

$$\begin{aligned}
\max_{i\in\{1,\dots,n\}}|\xi_i-\gamma_i^0| = \max_{i\in\{1,\dots,n\}}|\gamma_i^0+\alpha_i\,\hat{\varepsilon}_i-\gamma_i^0| &= \max_{i\in\{1,\dots,n\}}|\alpha_i\,\boldsymbol{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)| \\
&\le \boldsymbol{x}_{i^*}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0) \quad \text{for an } i^*\in\{1,\dots,n\}.
\end{aligned}$$

By the inequality of Cauchy-Schwarz (C-S),

$$\begin{aligned}
\boldsymbol{x}_{i^*}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0) \le \|\boldsymbol{x}_{i^*}\|\,\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\| &= \sqrt{\sum_{j=1}^p x_{i^*j}^2}\,\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\| \\
&\overset{\mathbf{A1}}{\le} \sqrt{p}\,M\,\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|\,.
\end{aligned}$$

Moreover, as $\hat{\boldsymbol{\beta}}$ is consistent, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| \xrightarrow[n\to\infty]{P} 0$ from where it follows that

$$\max_{i\in\{1,\dots,n\}} |\xi_i - \gamma_i^0| \xrightarrow[n\to\infty]{P} 0\,. \tag{4.3}$$

Therefore,

$$\begin{aligned}
|N_2| &= \sum_{j=1}^{p} \frac{e_j \sum_{i=1}^n x_{ij}\,\hat{\varepsilon}_i \left(\sigma'_{\mathrm{LR}}(\xi_i) - \sigma'_{\mathrm{LR}}(\gamma_i^0)\right)}{\|\boldsymbol{\Sigma X e}\|} \\
&\overset{\text{C-S}}{\le} \frac{\|\boldsymbol{e}\|}{\|\boldsymbol{\Sigma X e}\|} \sqrt{\sum_{j=1}^{p} \left| \sum_{i=1}^{n} x_{ij}\,\hat{\varepsilon}_i \left(\sigma'_{\mathrm{LR}}(\xi_i) - \sigma'_{\mathrm{LR}}(\gamma_i^0)\right)\right|^2} \\
&\le \frac{1}{\|\boldsymbol{\Sigma X e}\|} \sqrt{p\, n^2\, M^2\, \hat{\varepsilon}^2 \max_{i\in\{1,\dots,n\}} \left(\sigma'_{\mathrm{LR}}(\xi_i) - \sigma'_{\mathrm{LR}}(\gamma_i^0)\right)^2} \\
&= \frac{1}{\|\boldsymbol{\Sigma X e}\|} \sqrt{p}\, n\, M\, \hat{\varepsilon} \max_{i\in\{1,\dots,n\}} \left|\sigma'_{\mathrm{LR}}(\xi_i) - \sigma'_{\mathrm{LR}}(\gamma_i^0)\right|
\end{aligned}$$

where $\hat{\varepsilon} := \max_{i\in\{1,\dots,n\}} |\hat{\varepsilon}_i|$. For the denominator term, we get

$$\|\boldsymbol{\Sigma X e}\| = \sqrt{\sum_{i=1}^{n} \sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)(\boldsymbol{X e})_i^2} \ge \|\boldsymbol{X e}\| \min_{i\in\{1,\dots,n\}} s_i$$

with $s_i = \sqrt{\sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)}$. However, on one hand,

$$|\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0| \le \|\boldsymbol{x_i}\|\, \|\boldsymbol{\beta}^0\| \overset{\mathbf{A1}}{\le} \sqrt{p}\, M\, \|\boldsymbol{\beta}^0\|$$

such that $s_i^2 \ge \sigma'_{\mathrm{LR}}(\sqrt{p}\, M\, \|\boldsymbol{\beta}^0\|) =: c_0$. As, on the other hand, $\|\boldsymbol{X e}\| \ge \sqrt{cn}$ (combine (3.8) with **A2**), it follows that

$$\|\boldsymbol{\Sigma X e}\| \ge \sqrt{c_0\, c\, n} \tag{4.4}$$

and thus

$$|N_2| \le \frac{1}{\sqrt{c_0 c}} \sqrt{p}\, M\, \sqrt{n}\, \hat{\varepsilon} \max_{i\in\{1,\dots,n\}} \left|\sigma'_{\mathrm{LR}}(\xi_i) - \sigma'_{\mathrm{LR}}(\gamma_i^0)\right| \tag{4.5}$$

By the continuity of $\sigma'_{\mathrm{LR}}$ we get with (4.3) that

$$\max_{i\in\{1,\dots,n\}} \left|\sigma'_{\mathrm{LR}}(\xi_i) - \sigma'_{\mathrm{LR}}(\gamma_i^0)\right| \xrightarrow[n\to\infty]{P} 0\,, \tag{4.6}$$

and, by Cauchy-Schwarz again,

$$\hat{\varepsilon} = \max_{i\in\{1,\dots,n\}} |\hat{\varepsilon}_i| = \max_{i\in\{1,\dots,n\}} |\boldsymbol{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)| \le \sqrt{p}\, M\, \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|$$

Since $p$ is fixed and thus $\boldsymbol{X}B(\boldsymbol{\beta}^0,\delta)$ is bounded, there is a positive constant $C>0$ such that

$$|a_{pn}^{\delta}(\boldsymbol{X},\boldsymbol{\beta}^0)| = |\inf_{\boldsymbol{\zeta}\in\boldsymbol{X}B(\boldsymbol{\beta}^0,\delta)} \min_{i\in\{1,\dots,n\}} \sigma'_{\mathrm{LR}}(\zeta_i)| \geq C\,.$$

Therefore, as a conclusion of the remark on page 22, $\sqrt{n}\,\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|$ is bounded in probability. The same holds thus for $\sqrt{n}\,\hat{\varepsilon}$ and we get with (4.5) and (4.6) that

$$N_2 \xrightarrow[n\to\infty]{P} 0\,.$$

This result yields by (4.2) the condition

$$N_1 + N_3 \xrightarrow[n\to\infty]{P} 0\,. \tag{4.7}$$

Furthermore, we have

$$\begin{aligned} N_3 &= \sum_{j=1}^{p} \frac{e_j \sum_{i=1}^{n} x_{ij}\,\hat{\varepsilon}_i\,\sigma'_{\mathrm{LR}}(\gamma_i^0)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|} = \sum_{j=1}^{p} \frac{e_j \sum_{i=1}^{n} x_{ij}\,\boldsymbol{x}_i^{\mathrm{T}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\,\sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|} \\ &= \frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{\beta}^0)\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|}\,. \end{aligned}$$

As a result, if we can show that

$$-N_1 = \sum_{j=1}^{p} \frac{e_j \sum_{i=1}^{n} x_{ij}\,r_i}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|} = \frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|} \xrightarrow[n\to\infty]{\mathrm{L}} N(0,1)\,, \tag{4.8}$$

we have by (4.7) that $N_3 \xrightarrow[n\to\infty]{\mathrm{L}} N(0,1)$. For this purpose, we write $-N_1$ as

$$-N_1 = \frac{\sum_{i=1}^{n}\rho_i}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|}$$

where $\rho_i := (\boldsymbol{X}\boldsymbol{e})_i\,r_i$. We note that

$$\sum_{i=1}^{n}\mathrm{Var}(\rho_i) = \sum_{i=1}^{n}(\boldsymbol{X}\boldsymbol{e})_i^2\,\mathrm{Var}(r_i) = \sum_{i=1}^{n}(\boldsymbol{X}\boldsymbol{e})_i^2\,s_i^2 = \sum_{i=1}^{n}(\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e})_i^2 = \|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|^2\,.$$

Therefore, as a result of Lindeberg's variant of the central limit theorem (see Feller 1971, p. 262), (4.8) holds if the variances $\mathrm{Var}(\rho_i)$ satisfy the Lindeberg condition, i.e. for all $t>0$,

$$\sum_{i=1}^{n}\frac{\mathrm{E}(\rho_i^2\mid|\rho_i|\geq t\,\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|^2} \xrightarrow[n\to\infty]{} 0\,. \tag{4.9}$$

However, we have

$$\frac{\mathrm{Var}(\rho_i)}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|^2} = \frac{s_i^2\,(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{e})^2}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|^2} \leq \frac{s_i^2\,\|\boldsymbol{x}_i\|^2\,\|\boldsymbol{e}\|^2}{\|\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{e}\|^2} \leq \frac{p\,M^2}{4\,c_0\,c\,n} \xrightarrow[n\to\infty]{} 0\,, \tag{4.10}$$

the last inequality being a consequence of **A1**, (4.4), and the fact that $s_i^2 = \sigma'_{\mathrm{LR}}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^0) \leq \frac{1}{4}$. Finally, (4.10) implies (4.9), which completes the proof. ■

# 5 Case Study: Health Enhancing Physical Activity in the Swiss Population

In this last section, we shall study an example of applied logistic regression. In 1999, the Swiss Federal Offices of Sports and Statistics carried out a nationwide survey of attitudes, knowledge and behaviour related to health enhancing physical activity (HEPA)[10]. Based on recommendations[11] of the Swiss Federal Offices of Sports and Public Health as well as the Network HEPA Switzerland they questioned 1,529 randomly chosen Swiss inhabitants of at least 15 years of age. The samples for each of the German, French and Italian speaking parts of Switzerland were designed to have an equal size in order to obtain results of comparable precision.

Two criteria were used to determine whether a person was physically active or not. First, people were asked whether they usually practised half an hour of moderately intense physical activity each day, which is considered to be the base criterion for HEPA. An activity of moderate intensity is characterised by getting somewhat out of breath without necessarily sweating. Brisk walking, hiking, dancing, gardening and sports were stated as examples for such an activity. In weighting the results accordingly to the language region, the age, the gender and the household size of the questioned persons, it was estimated that 50.6% of the Swiss population met the HEPA base recommendations.

As a second criterion, people were questioned whether they practised regularly at least three times a week during twenty minutes or more a sporting activity of vigorous intensity. This was the case for 37.1% of the Swiss population.

A person who met at least one of these criteria was designated as "physically active", a person who did not meet both of them was considered "physically inactive". This information can now be represented for each person by the binary variable $y_i$. Let $H = \{h_i\}_{i\in\{1,\dots,n\}}$ denote the set of the questioned individuals. We define

$$y_i := \begin{cases} 1 & \text{if } h_i \text{ is ``physically inactive''}, \\ 0 & \text{if } h_i \text{ is ``physically active''}. \end{cases}$$

It has to be noted that, among the 1,529 questioned Swiss inhabitants, 65 reported not to be able to walk a distance of more than 200 meters without outside help. These and five others who did not give answers to the criteria mentioned above will not be taken into consideration. Therefore, $n$ is assigned the value of 1,459.

[10] First results have been published by Martin, Mäder and Calmonte (1999).

[11] See Network HEPA Switzerland (1999).

In order to illustrate the application of the logistic regression model, we shall now analyse the impact of the linguistic regions of Switzerland on the physical activity of their inhabitants. We divide $H$ in three mutually exclusive sets $H_G$, $H_F$ and $H_I$ containing the questioned inhabitants of the German, French and Italian speaking parts of Switzerland respectively. For every individual $h_i$ we define two indicator variables $x_{Fi}$ and $x_{Ii}$ by

$$x_{Fi} := \mathbf{1}_{H_F}(h_i) \quad \text{and} \quad x_{Ii} := \mathbf{1}_{H_I}(h_i)\,,$$

where $\mathbf{1}_S$ is the indicator function respective to the set $S$. Let us now estimate the coefficients $\beta_0$, $\beta_1$, and $\beta_2$ of the logistic regression model

$$\operatorname{logit}\pi = \beta_0 + \beta_1\, x_F + \beta_2\, x_I\,.$$

For this purpose, we are going to minimise by the use of *Mathematica* the error function $E_{\boldsymbol{y},\sigma_{\text{LR}}}$ defined as in (3.1) with respect to the given data set. Note that this procedure is absolutely equivalent to maximising the log-likelihood function, as has been shown in section 3.1. While the vector $\boldsymbol{y}$ contains the variables $y_i$ defined above, the matrix $\boldsymbol{X}$ is constructed by the row vectors $\boldsymbol{x}_i^{\text{T}} := (1\ x_{Fi}\ x_{Ii})$. The explicit values of $\boldsymbol{y}$ and $\boldsymbol{X}$ are not going to be displayed at this place. We assume that they are stored in the *Mathematica* variables **y** and **X**. The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\text{T}}$ is represented by the variable **b**:

```
b = {b0, b1, b2};
```

First of all, we assign the values of $n$ and $p$:

```
{n, p} = Dimensions[X]
{1459, 3}
```

In addition, the vector $\mathbf{1}_n$, the primitive function $H_{\sigma_{\text{LR}}}$, and the error function $E_{\boldsymbol{y},\sigma_{\text{LR}}}$ are defined and stored in **ev**, **H**, and **Err** respectively:

```
ev = Table[1, {e, 1, n}];
H[z_] := Log[Exp[z] + 1]
Err[b_] := ev.H[X.b] - y.(X.b)
```

Having specified $\boldsymbol{y}$ and $\boldsymbol{X}$, it is now possible to print out the precise form of $E_{\boldsymbol{y},\sigma_{\text{LR}}}(\boldsymbol{\beta})$:

```
Err[b]
-721 b0 - 299 b1 - 261 b2 + 509 Log[1 + e^b0] +
  485 Log[1 + e^(b0+b1)] + 465 Log[1 + e^(b0+b2)]
```

Finally, starting from $\boldsymbol{\beta} = \mathbf{0}$, the function **FindMinimum** searches for a local minimum in $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$:

```
betahat = FindMinimum[Err[{b0, b1, b2}],
  {b0, 0}, {b1, 0}, {b2, 0} , WorkingPrecision- > 20]
{959.3452655931,
  {b0 → -0.770799, b1 → 1.2455, b2 → 1.0172}}
```

The estimated coefficients are thus $\hat{\boldsymbol{\beta}}^{\mathrm{T}} = (-0.770799, 1.2455, 1.0172)$, and we get the model equation

$$\operatorname{logit} \pi = -0.770799 + 1.2455\, x_F + 1.0172\, x_I \,.$$

For the interpretation of these coefficients, we take into consideration the comments given in section 2.3.2. From there we know that the odds on the event that $y$ equals 1 increase (or decrease) by the factor $\exp \beta_i$ if $x_i$ grows by one unit. For people living in the German speaking part of Switzerland, both $x_F$ and $x_I$ are zero. Consequently, $H_G$ can be viewed as a "control group" in the present investigation. Living in the French or Italian speaking part of our country means from this viewpoint an increase of $x_F$ or $x_G$ by one unit respectively. We shall thus calculate the exponentials of $\hat{\beta}_1$ and $\hat{\beta}_2$:

```
{Exp[b1], Exp[b2]} /. betahat[[2]]
{3.47466, 2.76544}
```

As a result, the odds on an inhabitant of the French part of Switzerland being physically inactive are almost 3.5 times as high as those of someone living in the German part. For the Italian in comparison to the German parts, the respective odds ratio is about 2.75.

As Martin and Mäder (2001), both members of the Swiss Federal Office of Sports, state, these huge differences between the distinct linguistic regions of Switzerland are astonishing. They assume that the perception of moderately intensive physical activity is influenced by each indiviual's social and cultural background. The desirability of physical fitness is considered to be particularly developed in the German part of Switzerland which could have lead to a considerable overreporting (overclaiming) among this group of questioned people.

# 6 Conclusion

The example provided in section 5 gives a very rudimentary impression of the power of applied logistic regression. The analysis of this specific approach for modelling binary data within the scope of this thesis, however, has illustrated its versatility.

In the case of a fixed number of explanatory variables, both the consistency and asymptotic normality of the maximum likelihood estimator are established under two comparably weak hypotheses. Both of them are technical conditions with respect to the "layout" of the independent input variables.

If, on the other hand, the number of exogenous variables is allowed to grow along with the dimension of the sample space, it has been shown that there had to be essentially a logarithmic relation between the former and the latter in order to maintain the consistency of the maximum likelihood estimator.

Even without an explicit quantification of this relation, this result provides an interesting insight into the interdependence of these two characteristics.

# Acknowledgements

I would like to thank Prof Dr André Antille for the proposal of this interesting and many-sided topic as well as for his academic assistance during the development process of this diploma thesis.

Secondly, I want to express my gratitude to Prof Dr Bernard Marti, head of the Institute of Sports Sciences within the Swiss Federal Office of Sports in Magglingen, and to his collaborator Dr Brian W. Martin for placing at my disposal the complete data set of the HEPA study used and referred to in section 5. Thanks also to PD Dr Hans Howald for arranging the necessary contacts.

Finally, my appreciation goes to Manrico Glauser and Ralf Lutz for proofreading this text and giving me precious feedback.

# References

Amemiya, T. (1985). *Advanced Econometrics*, Harvard University Press, Cambridge.

Antille, A. and Rech, M. (1999). Learning Single Layer Neural Networks: Asymptotic Normality of Flat Spot Elimination Techniques, preprint.

Bhatia, R. (1997). *Matrix Analysis*, Graduate Texts in Mathematics, Springer, New York.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, Springer Texts in Statistics, 2nd edn, Springer, New York.

Collett, D. (1991). *Modelling Binary Data*, Chapman & Hall, London.

Cramer, J. S. (1991). *The LOGIT model: an introduction for economists*, Edward Arnold, London.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn, John Wiley & Sons, New York.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd edn, John Wiley & Sons, New York.

Garson, G. D. (2001). *Statnotes: An Online Textbook* [online], available from: `http://www2.chass.ncsu.edu/garson/pa765/statnote.htm`. [Accessed 23 September 2001].

Gourieroux, C. and Monfort, A. (1981). Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models, *Journal of Econometrics* **17**: 83–97.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*, John Wiley & Sons, New York.

Kleinbaum, D. G. (1994). *Logistic Regression: A Self-Learning Text*, Statistics in the Health Sciences, Springer-Verlag, New York.

Martin, B. W. and Mäder, U. (2001). Körperliches Aktivitätsverhalten in der Schweiz, preprint.

Martin, B. W., Mäder, U. and Calmonte, R. (1999). Einstellung, Wissen und Verhalten der Schweizer Bevölkerung bezüglich körperlicher Aktivität: Resultate aus dem Bewegungssurvey 1999, *Schweizerische Zeitschrift für ≪Sportmedizin und Sporttraumatologie≫* **47**: 165–169.

Mazza, C. and Antille, A. (1998). Learning Single Layer Neural Networks: Consistency of Flat Spots Elimination Techniques, preprint.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York, pp. 105–142.

National Safety Council (2001). *What Are the Odds of Dying?* [online], available from: `http://www.nsc.org/lrs/statinfo/odds.htm`. [Accessed 15 October 2001].

Network HEPA Switzerland (1999). *Health Enhancing Physical Activity HEPA (Recommendations)* [online], Available from: `http://www.hepa.ch/english/pdf-hepa/Empf_e.pdf`. [Accessed 10 November 2001].

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York.

Santner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*, Springer Texts in Statistics, Springer, New York.