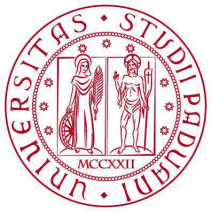

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione finale

Akaike's Information Criterion in Generalized Estimating Equations

Relatore: Prof. Alessandra Salvan

Dipartimento di Scienze Statistiche

Laureando: Francesco Ignazio Re
Matricola: 1149556

Contents

Introduction

Overview

Statistical analysis is a process that can be broken into different steps. From data collection, through data analysis, up to the yielding of consistent results, statisticians are continuously asked to come down to compromises in the attempt of tackling the underlying trends of their object of study. Among these steps, the greatest controversy is probably bound to model selection: a bitter truth known to every statistician is that there is no such thing as the best model. With that said, it is still reasonable to search - if not for the best - for a *better* model and, in this respect, several indexes were built for comparing different models with each other. A particularly powerful index is the Akaike's information criterion; it is based on the likelihood and asymptotic properties of the maximum likelihood estimator and allows model comparison in terms of predictability and parsimony. Despite being a powerful tool, its strict dependence on the likelihood implies the model distribution to be fully known: a requirement that cannot always be fulfilled. In this context, this work sets its aim at assessing methods to widen the AIC usage to those models for which there is no likelihood defined. We will specifically focus our attention on the Akaike's information criterion for models estimated through the generalized estimating equation (GEE) approach, very useful for working with correlated data, but based on the quasi-likelihood estimation, and hence, unconstrained by any exact specification of the distribution.

Summary

In order to achieve the aim we set out for, we proceed by firstly introducing some core concepts of inferential statistics, to establish the notation used when referring to them and to outline the relationship that these have with our specific topic of study. For this reason, the first chapter will cover the theory of likelihood and the most-frequently-used models where it is applied. The second chapter will instead focus on the quasi-likelihood theory, on its connection with the least-square method and on the accuracy of its estimates. In the last part of the chapter, we will focus on generalized estimating equations, as a method to produce efficient estimates when working with correlated data. In the last chapter, we will provide some insight on the derivation of AIC, on its relation with information entropy and we will try to assess its performance when the quasi-likelihood is used instead. We will eventually draw our conclusions by testing our theoretical results on a real data set.

Chapter 1

Parametric models and likelihood

1.1 Introduction

In this chapter, we will first introduce the likelihood function along with its main properties. We will then briefly discuss Linear Models (LM) and Generalized Linear Models (GLM), as being two classes of models that use the likelihood function for the estimation of their parameters of interest. The information herein provided is referenced from Pace, Salvan (2001, §1-3), V. Hogg, Tanis, Zimmerman (2014, §6.4) and Salvan, Sartori, Pace (2018, §1-2).

1.2 Likelihood

1.2.1 Model specification

The aim of statistical inference is to gain insight regarding the underlying distribution of a phenomenon of interest Y, given that we have access to a limited sample of observations of Y, $(y_1, y_2, ..., y_n)$. Assuming that Y is defined by the parametric density function $f(y, \theta_0)$, with θ_0 being the only unknown component of $f(\cdot)$, then our goal is to draw conclusions regarding the value θ_0, using the information embedded in the sample $(y_1, y_2, ..., y_n)$. In this way, we restrict our interest on a precise family of distributions to which we refer as our model of interest. Formally, we define a parametric model $\mathcal{F}$ as

$$\mathcal{F} = \{f(y; \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\} \tag{1.1}$$

with $p \in \mathbb{N}^+$ and Θ being the parametric space, namely the space containing all the possible values of θ and, indeed, θ_0 itself.

1.2.2 Likelihood function

The concept of likelihood is at the very core of traditional statistical inference. The term was firstly used by Fisher, in 1921, and defined as follows:

The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.

In other words, it is a method to discriminate among all the possible values of θ, considering for each $\theta \in \Theta$ the values assumed by the density function when conditioned to the sample $(y_1, y_2, ..., y_n)$: the higher the density for a given θ_1, the more likely for θ_1 to be the real θ_0. Assuming the model $\mathcal{F}$ with density function $f(y, \theta)$ to be correct for the sample $(y_1, y_2, ..., y_n)$, we can then define the likelihood function $L : \Theta \to \mathbb{R}^p$ as

$$L(\theta) = L(\theta; y) = c(y)f(y; \theta), \tag{1.2}$$

with $c(y)$ being a function of the data, independent from the parameter. With respect to the model $\mathcal{F}$, the likelihood is a class of functions equivalent to each other, and differing only for the component $c(y)$. If the observations $(y_1, ..., y_n)$ are independent and identically distributed, then the likelihood function is simply the product of the individual densities, thus can be expressed as

$$L(\theta) = \prod_{i=1}^{i=n} f_{Y_i}(y_i, \theta),$$

with $f_{Y_i}(y_i, \theta)$ being the density function of the random variable Y_i, generator of the i-th observation, y_i, of the sample $(y_1, ..., y_n)$.

For a more straightforward approach in calculations, we usually operate with the natural logarithm of the likelihood function: being the natural logarithm a monotonically increasing transformation, it does not alter the information embedded in the data, while still providing a much more manageable form. We then define the log-likelihood as

$$l(\theta) = l(\theta; y) = \log L(\theta; y). \tag{1.3}$$

In the case of independent and identically distributed observations, the log-likelihood would be

$$l(\theta) = \sum_{i=1}^{i=n} \log f(y_i, \theta). \tag{1.4}$$

1.2.3 Maximum likelihood estimation

Maximum Likelihood Estimate (MLE)

Given a sample of observations $(y_1, y_2, ..., y_n)$, any estimate $\hat{\theta} \in \Theta$ that maximizes $L(\theta)$ over Θ is called a maximum likelihood estimate (MLE) of the unknown true parameter θ_0. We should note that this definition by itself does not assume either the existence or uniqueness of the MLE. If $\hat{\theta} = \hat{\theta}(y)$ exists and it is unique with probability equal to one, then the random variable $\hat{\theta}(Y)$ is called Maximum Likelihood estimator. The ML estimator is obtained by replacing the observations $(y_1, y_2, ..., y_n)$ with the random vector $Y = (Y_1, ..., Y_n)$.

Regular models

In order to find the MLE through the method we are about to discuss, we require some regularity conditions on the model under consideration. A model that conforms to these conditions it is called a regular model. The conditions are:

- Θ to be an open subset of $\mathbb{R}^d$.
- the log-likelihood function to be differentiable at least three times.
- the model to be identifiable.
- the support of the density of the model to be independent from the parameter.

In the case of regular models, the partial derivatives of the log-likelihood function are zero when evaluated at any local extreme value. These points correspond to the solution of the so-called likelihood equation(s) - also known as *score*.

Score function

Given the parameter $\theta = (\theta_1, \theta_2, ..., \theta_p)$, the vector of partial derivatives corresponding to the d-dimensional set of likelihood equations

$$l_*(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, ..., \frac{\partial l(\theta)}{\partial \theta_p} \right) = \left[\frac{\partial l(\theta)}{\partial \theta_r} \right] = [l_r(\theta)], \tag{1.5}$$

is called *score*.

Observed and expected information

To make sure that a solution to the score corresponds to a maximum it is necessary to check the Hessian matrix containing the second-order partial derivatives of the log-likelihood. This matrix provides some insightful information regarding the curvature of the function, giving an hint on how steeply the function approaches its maximum, and

hence, on how choosing $\hat{\theta}$ differs from choosing any other θ in the surroundings of $\hat{\theta}$. We make the most of this clue by defining

$$j(\theta) = -l_{**}(\theta) = -\left[\frac{\partial^2 l(\theta)}{\partial\theta_r \partial\theta_s}\right] = [j_{rs}(\theta)], \tag{1.6}$$

as the observed information matrix. The expected value under θ of the observed information matrix is the expected information matrix

$$i(\theta) = E_\theta[j(\theta)] = [i_{rs}(\theta)]. \tag{1.7}$$

1.2.4 Likelihood properties

Likelihood exact properties

For any regular model, satisfying as well some other further regulatory conditions, we have that

$$E_\theta(l_*(\theta; Y)) = 0 \qquad \forall\, \theta \in \Theta, \tag{1.8}$$

$$E_\theta(l_*(\theta; Y)(l_*(\theta; Y))^\top) = i(\theta) \qquad \forall\, \theta \in \Theta. \tag{1.9}$$

Likelihood asymptotic properties

We already defined in section 1.2.3 the maximum likelihood estimator. Under regulatory conditions, and specifically when the dimension of Θ is independent from n, the maximum likelihood estimator is consistent, that is,

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

Furthermore, for θ being the real value of the parameter, we also have that

$$l_*(\theta) \dot{\sim} N_d(0, i(\theta)), \tag{1.10}$$

$$\hat{\theta} - \theta \dot{\sim} N_d(0, i(\theta)^{-1}), \tag{1.11}$$

or, as well

$$\hat{\theta} - \theta \dot{\sim} N_d(0, j(\theta)^{-1}). \tag{1.12}$$

Moreover,

$$W_e(\theta) = (\hat{\theta} - \theta)^\top j(\hat{\theta})(\hat{\theta} - \theta) \dot{\sim} \chi^2_d, \tag{1.13}$$

$$W_u(\theta) = (\hat{\theta} - \theta)^\top i(\hat{\theta})(\hat{\theta} - \theta) \dot{\sim} \chi^2_d, \tag{1.14}$$

$$W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\} \dot{\sim} \chi^2_d. \tag{1.15}$$

The three quantities W_e, W_u and W are asymptotically equivalent. Through these quantities, confidence intervals can be built and hypothesis can be tested. We call them, respectively, Wald quantity, score and likelihood ratio.

1.3 Linear regression models

1.3.1 Assumptions

Linear regression models are used for modeling the relationship between a response variable Y, and one - or a set - of independent variables $x_1, ..., x_p$, assuming this relationship to be linear. We decide to discuss these models in this chapter because the assumptions on which they are built allow for a straightforward use of the Maximum Likelihood Estimation for estimating the parameters of interest. These assumptions are:

1. $Y = X\beta + \varepsilon = \eta + \varepsilon$, with

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & ... & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & ... & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

2. X matrix of $n \times p$ constants, with rows $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and full rank p.

3. $\varepsilon \sim N_n(0, \sigma^2 I_n)$, with $\sigma^2 > 0$

The notation $N_m(\mu, \Sigma)$ refers to a m-dimensional normal distribution with vector of means μ, and covariance matrix Σ. Hence, given the independence between different observations of Y - as implicitly stated by the assumptions - and thanks to the a-priori settled distribution, we can easily compute the likelihood function for estimating the β coefficients and σ^2.

1.3.2 MLE of the parameters

In a linear regression model we can compute the likelihood as the joint probability density function of the vector Y. The log-likelihood, with parameters (β, σ^2), is defined on the parametric space $\mathbb{R}^p \times (0, +\infty)$ as

$$\begin{aligned} l(\beta, \sigma^2; y) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^\top(y - X\beta) \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\|y - X\beta\right\|^2, \end{aligned}$$

where, given a vector $u \in \mathbb{R}^p$, $\|u\|^2 = u^\top u$ is the squared norm of u. The function depends on the data through

$$\sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 = \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{p}\beta_j \sum_{i=1}^{n} x_{ij} y_i + \sum_{i=1}^{n}(\boldsymbol{x}_i\beta)^2.$$

Hence, the minimal sufficient statistic is

$$s = (\sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} x_{i1} y_i, ..., \sum_{i=1}^{n} x_{ip} y_i), \tag{1.16}$$

for all the information needed for estimating (β, σ^2) are the components of s.

The Maximum Likelihood Estimate is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \tag{1.17}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^\top(y - X\hat{\beta}) = \frac{1}{n}\left\|y - X\beta\right\|^2. \tag{1.18}$$

Since the estimator $\hat{\beta}$ is a linear transformation of the normally distributed Y, it is itself a normally distributed random variable. Given (β, σ^2) to be the true value of the parameter, then

$$\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1}). \tag{1.19}$$

The estimator $\hat{\sigma}^2$ follows instead a χ^2_{n-p} distribution with $n - p$ degrees of freedom, for

$$n\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}. \tag{1.20}$$

1.3.3 Predicted values and deviance

Once the MLE $\hat{\beta}$ is obtained, it is possible to compute the vector of the values predicted by the model, given by

$$\hat{y} = \hat{\mu} = X(X^\top X)^{-1}X^\top y = \hat{\beta}X, \tag{1.21}$$

We can further define the residuals' vector of the model as

$$e = y - \hat{y} = y - X\beta, \tag{1.22}$$

that contains information regarding how close every predicted value is from its correspondent observed one. Every individual numeric difference - namely the residual e_i - can be considered as an estimate of the casual error ε_i, as also the MLE $\hat{\sigma}^2$ suggests:

$$\hat{\sigma}^2 = \frac{1}{n}\|y - X\beta\|^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{1.23}$$

A quantity that well explains the connection between the residuals and the variability of the response variable Y, is the equality

$$y^\top y = \hat{y}^\top \hat{y} + e^\top e, \tag{1.24}$$

which is equal to

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{1.25}$$

As we can see, the smaller the residuals squared sum, the higher the variability we are able to tackle through the model. For this reason, we name each component

- $\sum_{i=1}^{n}(y_i - \bar{y})^2 = SQ_{tot}$ = Total deviance,
- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = SQ_{reg}$ = Explained deviance,
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SQ_{res}$ = Residual deviance.

An index that exploits the information embedded in the above formula and that can be used for assessing the goodness of fit of a linear model is the R^2 coefficient. It is defined as follows:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{res}}{SQ_{tot}}, \tag{1.26}$$

and represents the portion of variability of Y explained by the regression model.

For the purpose of this work, we make a further consideration about the deviance: as being a quantity implicitly dependent on the likelihood, the more the likelihood

increases, the more the residual deviance decreases, hence the explained deviance approaches the total deviance. When adding a new independent variable x_{p+1} - with coefficient β_{p+1} - to the model, the likelihood will always increase, independently from the actual influence that x_{p+1} has on Y. This means that the R^2 coefficient as well will increase whenever adding a new variable, even if the variable itself had little influence on the response. This aspect represents a limitation of R^2, and as we will see, whenever building an index to assess the goodness of fit of a model, we will also need to count in a penalization linked to the numbers of parameters used. As long as R^2 is concerned, the adjusted version that takes into account the number of parameters used by the model is

$$R^2_{adj} = 1 - (1 - R^2)\frac{(N-1)}{(N-p-1)}. \tag{1.27}$$

1.3.4 Least square prediction: good enough?

Having talked about likelihood and linear models, one assumption that we made was about the normality in distribution of the response. This assumption significantly restricts the class of treatable problems, ruling out any other variable that is not normally distributed. However, how much does it change to not assume the distribution of Y? Well, even if we were not to use the likelihood to find the estimate of the coefficients, we could still use the least square prediction method, that - given the other assumptions to still hold true - would result into the same outcome of the former case.

The method consists in finding the minimum

$$argmin\{(y - X\beta)^\top (y - X\beta)\},$$

and the result is, as said,

$$\hat{\beta} = (X^\top X)^{-1} X^\top y. \tag{1.28}$$

Regarding the validity of this method, the Gauss-Markov theorem states that in a linear regression model in which the errors have expectation zero, are uncorrelated and have equal variances, the best linear unbiased estimator of the coefficients is given by the ordinary least squares estimator, provided it exists - namely (**??**). What this method lacks, however, is a way to assess statistical uncertainty through exact confidence intervals and statistical tests, since we do not have a settled distribution to start from.

1.4 Generalized linear models

1.4.1 Exponential families

Let us consider now a single random variable Y_i, $i = 1, ..., n$, whose probability distribution depends on the parameter (θ_i, ϕ). The distribution belongs to the exponential family if it can be written in the form

$$p(y_i; \theta_i, \phi) = \exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{1.29}$$

with $y_i \in S \subseteq \mathbb{R}, a_i(\phi) > 0$. The parameter θ_i is called natural parameter, whereas ϕ is called dispersion parameter. When $a_i(\phi) = 1$ and $c(y_i, \phi) = c(y_i)$, then the family of distribution is the natural exponential one, with density

$$p(y_i; \theta_i) = \exp\{\theta_i y_i - b(\theta_i) + c(y_i)\}.$$

If we lay out the functions $a(\cdot)$ and $b(\cdot)$ we obtain a specific parametric model. Among the parametric distribution belonging to the exponential family there are the the normal distribution, the Poisson, gamma, Binomial etc.

1.4.2 Moment and cumulant generating function, variance and expected value

The moment generating function is defined as

$$M_{Y_i}(t; \theta_i, \phi) = E(e^{tY_i}) = \int_S e^{ty_i} p(y_i; \theta_i, \phi) dy_i \tag{1.30}$$

$$= \exp\{[b(\theta_i + ta_i(\phi)) - b(\theta_i)]/a_i(\phi)\}. \tag{1.31}$$

$$\tag{1.32}$$

Hence, the cumulant generating function of Y_i is

$$K_{Y_i}(t; \theta_i, \phi) = [b(\theta_i + ta_i(\phi)) - b(\theta_i)]/a_i(\phi), \tag{1.33}$$

and the cumulant of of the r-th order of Y_i is

$$k_r(Y_i) = \left.\frac{\partial^r K_{Y_i}(t; \theta_i, \phi)}{\partial t^r}\right|_{t=0} = a_i(\phi)^{r-1} b^{(r)}(\theta_i),$$

where $b^{(r)}(\theta_i)$ is the r-th derivative of $b(\theta_i)$, $\quad r = 1, 2, \ldots$.

We can use the cumulant generating function to compute $E(Y)$ and $Var(Y)$, knowing that

$$E(Y) = \frac{d}{dt}K_y(t)\Big|_{t=0}, \quad Var(Y) = \frac{d^2}{dt^2}K_y(t)\Big|_{t=0},$$

then

$$E(Y) = E_{\theta_i,\phi}(Y_i) = b^{'}(\theta_i)Var(Y_i) = Var_{\theta_i,\phi}(Y_i) = a_i(\phi)b^{''}(\theta_i). \tag{1.34}$$

1.4.3 Reparameterization with mean and variance function

If we set

$$\mu_i(\theta_i) = E_{\theta_i,\phi}(Y_i) = b^{'}(\theta_i), \tag{1.35}$$

then we have

$$Var_{\theta_i,\phi}(Y_i) = a_i(\phi)\frac{d}{d\theta_i}\mu_i(\theta_i) = a_i(\phi)\mu_i^{'}(\theta_i), \tag{1.36}$$

with $\mu_i(\theta_i)$ defining a reparameterization (μ_i, ϕ) of an exponential family distribution. We shall note that $Var_{\theta_i,\phi}(Y_i) > 0$ for any θ_i, hence $\mu_i(\cdot)$ is monotonically increasing with domain Θ and co-domain the mean space $M = \mu_i(\text{int}\Theta)$, where int Θ is the set of inner points of Θ, independent from i. Indicating with $\theta_i(\mu_i)$ the inverse that asserts θ_i as a function of μ_i, then the variance expressed through the reparameterization (μ_i, ϕ) is

$$Var_{\theta_i,\phi}(Y_i) = a_i(\phi)\ b_i^{''}(\theta_i)\Big|_{\theta_i=\theta_i(\mu_i)} = a_i(\phi)v(\mu_i),$$

where the function defined on the mean space M

$$v(\mu_i) = b_i^{''}(\theta_i)\Big|_{\theta_i=\theta_i(\mu_i)},$$

is called variance function. If we know the mean space M and $v(\mu_i)$, we can find $b(\theta_i)$ that - known $a_i(\phi)$ - defines a parametric model with density belonging to the exponential family. We can then define the singular random variable Y_i as

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \quad \mu_i \in M. \tag{1.37}$$

1.4.4 Assumptions in generalized linear models

In defining a generalized linear model (g.l.m.) we make the following assumptions regarding the random component of the model, the linear predictor and the link function:

- $Y_1, ..., Y_n$ independent random variables .
- $g(E(Y_i)) = g(\mu_i) = \eta_i = \boldsymbol{x}_i\beta$,
- $Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i))$,

where $g(\cdot)$ is a known invertible function, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ is the known row-vector of covariates and $\beta = (\beta_{i1}, ..., \beta_{ip})^\top$ is the vector of regression coefficients. For every specification of the statistical model regarding the response variable Y, among all the possible $g(\cdot)$, the one such that

$$g(\mu_i) = \theta_i(\mu_i)$$

grants the natural parameter of the exponential family θ_i to be a linear combination of the covariates with coefficients β, $\theta_i = \boldsymbol{x}_i\beta$, $i = 1, ..., n$. In this case, the link function is called canonic link function.

1.4.5 Likelihood estimation in generalized linear models

Having $Y_1, ..., Y_n$ random variables distributed following the assumptions enlisted above, then $Y = (Y_1, ...Y_n)$ has the joint density equal to the product of the marginal ones, hence the log-likelihood is

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi), \tag{1.38}$$

with $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\boldsymbol{x}_i\beta))$.

Sufficient Statistic

Even if we suppose ϕ to be fully known, generally there is not a sufficient statistic for estimating β of dimension inferior to n. However, in the specific case where $g(\mu_i) = \theta_i(\mu_i)$, such that $\theta_i = (\boldsymbol{x}_i\beta)$, then it exists a p-dimensional statistic for β. Then the log-likelihood can be written as

$$l(\beta,\phi) = \sum_{i=1}^{n} \frac{y_i \boldsymbol{x}_i\beta - b(\boldsymbol{x}_i\beta)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi)$$
$$= \beta_1 \sum_{i=1}^{n} \frac{1}{a_i(\phi)} x_{i1}y_i + ... + \beta_p \sum_{i=1}^{n} \frac{1}{a_i(\phi)} x_{ip}y_i - \sum_{i=1}^{n} \frac{b(\boldsymbol{x}_i\beta)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i,\phi),$$

and the p-dimensional statistic, defined as

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)} \boldsymbol{x}_i y_i = \Big(\sum_{i=1}^{n} \frac{1}{a_i(\phi)} x_{i1}y_i, ..., \sum_{i=1}^{n} \frac{1}{a_i(\phi)} x_{ip}y_i \Big), \tag{1.39}$$

is a minimal sufficient statistic for β given any fixed value of ϕ.

Score function

Taking into consideration that

- $\dfrac{\partial\theta_i}{\partial\mu_i} = \theta_i'(\mu_i) = \dfrac{1}{\mu_i'(\theta_i)}\Big|_{\theta_i=\theta_i(\mu_i)} = \dfrac{1}{b_i''(\theta_i)}\Big|_{\theta_i=\theta_i(\mu_i)} = \dfrac{1}{v(\mu_i)},$
- $\dfrac{\partial\mu_i}{\partial\eta_i} = \dfrac{1}{g'(\mu_i)}\Big|_{\mu_i=g^{-1}(\eta_i)},$

We then have

$$\frac{\partial\theta_i}{\partial\beta_r} = \frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_r} = \frac{1}{v(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r} = \frac{1}{v(\mu_i)}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_r} = \frac{1}{v(\mu_i)}\frac{1}{g'(\mu_i)}x_{ir}, \quad r = 1, ..., p.$$

If $g(\cdot)$ is the canonic link function, $g(\mu_i) = \theta_i(\mu_i)$, then $g'(\mu_i) = 1/v(\mu_i)$, for

$$\frac{\partial\theta_i}{\partial\beta_r} = x_{ir}, \qquad r = 1, ..., p$$

We can then define the score vector with components

$$l_r = \frac{\partial l(\beta,\phi)}{\partial\beta_r} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \Big(y_i \frac{\partial\theta_i}{\partial\beta_r} - \frac{\partial b(\theta_i)}{\partial\beta_r} \Big), \qquad r = 1, ..., p \tag{1.40}$$

$$l_\phi = \frac{\partial l(\beta,\phi)}{\partial\phi} = -\sum_{i=1}^{n} \frac{a_i'(\phi)}{(a_i(\phi))^2}(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c'(y_i,\phi). \tag{1.41}$$

Since

$$\frac{\partial b(\theta_i)}{\partial\beta_r} = b'(\theta_i)\frac{\partial\theta_i}{\partial\beta_r} = \mu_i\frac{\partial\theta_i}{\partial\beta_r},$$

we can write the *score* component related to β as

$$l_r = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}(y_i - \mu_i)\frac{\partial \theta_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)}\frac{\partial \mu_i}{\partial \beta_r} = 0, \qquad r = 1, ..., p \tag{1.42}$$

assuming ϕ to be fixed. If the link function is the canonic one, then $g'(\mu_i) = 1/v(\mu_i)$, so the likelihood equations can be written as

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)} y_i x_{ir} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \mu_i x_{ir}, \qquad r = 1, ..., p \tag{1.43}$$

The *score* components of β can also be expressed as

$$(y - \mu)^\top V^{-1} D = 0, \tag{1.44}$$

with $(y - \mu)^\top = (y_1 - \mu_1, ..., y_n - \mu_n)$,

$$V = \text{diag}(Var(Y_i)), \qquad i = 1, ..., n$$

and D being a matrix $n \times p$ with generic element

$$d_{ir} = \frac{\partial \mu_i}{\partial \beta_r} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{1}{g'(\mu_i)} x_{ir}, \qquad i = 1, ..., n,\ r = 1, ..., p$$

Fisher information

We shall first prove that β and ϕ are orthogonal, that is the $i_{\beta\phi}$ component of the expected information matrix has all its elements equal to 0. In fact, we have

$$E[-l_{\beta\phi}] = E\Big[-\frac{\partial^2 l(\beta, \phi)}{\partial \beta_r \partial \phi}\Big] = E\Big[\sum_{i=1}^{n} \frac{a_i'(\phi)}{(a_i(\phi))^2}(y_i - \mu_i)\frac{\partial \theta_i}{\partial \beta_r}\Big] = 0, \qquad i = 1, ..., n \tag{1.45}$$

considering that $E[Y_i] = \mu_i$. Hence, as a consequence of orthogonality, the MLE of β and of ϕ are asymptotically independent. Hence, if interested in estimating β, we can solely focus on the component related to β of the information matrix. We then obtain

$$j_{rs} = -l_{rs} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\Big[\frac{\partial \mu_i}{\partial \beta_s}\frac{\partial \theta_i}{\partial \beta_r} - (y_i - \mu_i)\frac{\partial^2 \theta_i}{\partial \beta_r \partial \beta_s}\Big], \tag{1.46}$$

and, concerning the expected information component,

$$i_{rs} = E[j_{rs}] = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\frac{\partial \mu_i}{\partial \beta_s}\frac{\partial \theta_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\frac{x_{ir}x_{is}}{(g'(\mu_i))^2 v(\mu_i)}, \tag{1.47}$$

that is usually reported as

$$i_{\beta\beta} = X^\top W X, \tag{1.48}$$

where

$$W = \text{diag}(w_i), \qquad \text{with } w_i = \frac{1}{(g'(\mu_i))^2 Var(Y_i)}, \qquad i = 1, ..., n.$$

Defined these quantities, the property of asymptotic normality of the MLE grants the approximation, for n diverging,

$$\hat{\beta} \dot{\sim} N(\beta, (X^\top W X)^{-1}). \tag{1.49}$$

Iteratively re-weighted least squares

Regarding the estimation of the coefficients of the model, the likelihood equations do not usually have an explicit solution. Iterative methods such as Newton-Raphson come into play to help us compute a good approximation of the solution. Defining l_* as the vector of elements l_r and $j = j_{\beta\beta}$ as the observed information matrix with elements $-l_{rs}$, the $(m+1)-$th iteration provides us with the following approximation:

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \left(j_{\beta\beta}(\hat{\beta}^{(m)})\right)^{-1} l_*(\hat{\beta}^{(m)}). \tag{1.50}$$

As we mentioned above, the expected information matrix figures in a much simpler expression, for it is convenient to switch the quantity $j_{\beta\beta}(\hat{\beta}^{(m)})$ with $i_{\beta\beta}(\hat{\beta}^{(m)})$, knowing that the algorithm would still converge. We, thus, obtain

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \left(i_{\beta\beta}(\hat{\beta}^{(m)})\right)^{-1} l_*(\hat{\beta}^{(m)}),$$

that can also be written as

$$i_{\beta\beta}(\hat{\beta}^{(m)})\hat{\beta}^{(m+1)} = i_{\beta\beta}(\hat{\beta}^{(m)})\hat{\beta}^{(m)} + l_*(\hat{\beta}^{(m)}).$$

Knowing that we can express l_* and $i_{\beta\beta}$ as

$$l_* = X^\top W u, \tag{1.51}$$

$$i_{\beta\beta} = X^\top W X, \tag{1.52}$$

with $u = ((y_1 - \mu_1)g'(\mu_1), ..., (y_n - \mu_n)g'(\mu_n))^\top$ and $W = \text{diag}(w_i)$, with $w_i = \dfrac{1}{(g'(\mu_i))^2 Var(Y_i)}, i = 1, ..., n$, then we can write

$$X^\top W X \hat{\beta}^{(m+1)} = X^\top W z^{(m)}, \tag{1.53}$$

where

$$z^{(m)} = X\hat{\beta}^{(m)} + u.$$

The generic component of $z^{(m)}$ is

$$z_i^{(m)} = \boldsymbol{x}_i \hat{\beta}^{(m)} + (y_i - \mu_i)g'(\mu_i), \qquad i = 1, ..., n, \tag{1.54}$$

and can be seen as a linear approximation of $g(y_i)$. In fact

$$g(y_i) \doteq g(\mu_i) + (y_i - \mu_i)g'(\mu_i) = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}. \tag{1.55}$$

Then, we may set the initializing value of $z_i^{(0)} = g(y_i)$ and $W^{(0)} = I_n$. Reached the convergence of the algorithm, we will have

$$\hat{\beta} = (X^\top \hat{W} X)^{-1} X^\top \hat{W} \hat{z}, \tag{1.56}$$

with $\hat{z} = X\hat{\beta} + \hat{u}$.

Dispersion parameter estimation

In those models where ϕ is not fixed, it is usually estimated with the momentum method: knowing that $Var(Y_i) = \phi v(\mu_i)$ and given β to be known, we would have

$$\tilde{\phi} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{v(\mu_i)}. \tag{1.57}$$

However, when using $\hat{\beta}$ to estimate ϕ, in the attempt of taking account for the variability of $\hat{\beta}$, we prefer to use

$$\tilde{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{1.58}$$

Chapter 2

Semiparametric models and Quasi Likelihood

2.1 Introduction

In chapter 1, we introduced the linear regression model and the generalized linear models as examples of statistical parametric models. In that context, we presented the likelihood as an approach for estimating the models' parameters. However, concerning linear regression models only, we also briefly presented the least square method, an alternative approach that, while yielding to the same estimate of the likelihood, does not need to assume the normality of the response. The optimality of the least-squares estimate, as a matter of fact, depends on assumptions regarding solely the first two moments of the distribution, rather than the entire distributional form - as stated by the Gauss-Markov theorem.

Shifting to generalized linear models, historically, Wedderburn (1974) was the first to notice that the score function of these models - that again depends on the parameters only through mean and variance - can also be written as a weighted least-squares estimating function. So, to estimate the parameters, he suggested to use the score function of the exponential family even when the model distribution was not specified. This intuition gave birth to the quasi-likelihood, which we will be treating in this chapter. The content hereinafter is referenced from Salvan, Sartori, Pace (2018,§6-7), Liang, L. Zeger (1986), Godambe, Heyde (1987), McCullagh (1983).

2.2 Quasi-likelihood function

2.2.1 Framework of estimating function theory

We firstly intend to better establish the connection between the Gauss-Markov theorem and - as we will call it - the quasi-likelihood function.

Let us consider $Y_1, ..., Y_n$ independent random variables and $\mathcal{F} = \{f(y; \theta, \sigma^2); \theta \in \Theta, \sigma^2 > 0\}$. The parameters θ and σ^2 are such that, for $i = 1, ..., n$

$$E_\theta(Y_i - \theta) = 0, \qquad E_\theta(Y_i - \theta)^2 = \sigma^2, \qquad \theta \in \Theta.$$

Now, for any specified numbers $\alpha_i \in \mathbb{R}$, $i = 1, ..., n$, we define the function

$$q = \sum_{i=1}^{n} (Y_i - \theta)\alpha_i,$$

that is both linear - since it is linear in $(Y_i - \theta)$ - and unbiased, for $E_\theta(q) = 0$. Thus, we call q a linear unbiased estimating function, for the equation $q = 0$, when solved for θ, provides an estimate $\hat{\theta}$ for θ.

Let us now restrict on $\mathcal{G}$, the class of all linear unbiased estimating functions subject to the condition $\sum_{i=1}^{n} \alpha_i = c$, with c being a constant. In this context, an estimating function q^* is said to be optimal in $\mathcal{G}$ if $q^* \in \mathcal{G}$ and if, for all $q \in \mathcal{G}$,

$$E_\theta\{(q^*)^2\} < E_\theta(q^2), \qquad \theta \in \Theta \tag{2.1}$$

Hence, we are looking for the estimator obtained from q that has minimal variance among all the estimators obtained from estimating functions in $\mathcal{G}$.

It is easy to see, that up to a constant multiple, the estimating function q^* is given by

$$q^* = \bar{y} - \theta \tag{2.2}$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$.

This is essentially what the Gauss-Markov theorem states and the reason why the least square method is such a good estimation method.

Shifting to a more general framework, let us now find a way to extend what said to semiparametric models, with $f(y; \theta)$ being not known. For this class of problems, we define our model as $\mathcal{F} = \{f\} \times \{\theta\}$. $\mathcal{F}$ is a union of families of parametric distributions, each family being indexed by the same parameter. This means that, for each f, the parameter range is the same, namely $\{\theta\}$. We can then assume the existence of the *score* function related to $\mathcal{F}$, namely $\partial \log f_\theta / \partial\theta$, for all θ and f in $\mathcal{F}$. It can be shown

that the optimality of function q^*, such that $E_\theta\{(q^*)^2\} < E_\theta(q^2)$, is equivalent to each of:

- $E_\theta\Big(q^* - \dfrac{\partial \log f_\theta}{\partial \theta}\Big)^2 < E_\theta\Big(q - \dfrac{\partial \log f_\theta}{\partial \theta}\Big)^2$
- $\text{corr}\left(q^*, \dfrac{\partial \log f_\theta}{\partial \theta}\right) > \text{corr}\left(q, \dfrac{\partial \log f_\theta}{\partial \theta}\right)$

for all $q \in \mathcal{G}$ and $f \in \mathcal{F}$. Furthermore, under suitable conditions, for large samples the confidence interval for θ associated with q^* is smaller or equal to that associated with q for all $q \in \mathcal{F}$ and $f \in \mathcal{F}$. Because of these properties, that as we have seen resemble the ones of the score function in a parametric family of distributions, we define q^* as the quasi score function, which yields to the quasi-likelihood by an appropriate integration. In the specific case where the condition of homoscedasticity of the observations holds true, then, as before, $q^* = \bar{y} - \theta$. In a more general setting, especially when there is a given relation between the variance and the mean of the observations, q^* is defined as

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i}{V(\mu_i)} \tag{2.3}$$

with $Y_1, ..., Y_n$ independent random variables, $E_{\theta_i}(Y_i) = \theta_i = \mu_i$ and $Var(Y_i) = V(\mu_i)$, function of μ_i.

2.2.2 Inference on unbiased estimating equations

After defining q as a generic linear unbiased estimating function, we shall define the vector

$$q(y;\theta) = (q_1(y;\theta), ..., q_d(y;\theta))^\top, \tag{2.4}$$

as a d-dimensional vector of linear unbiased estimating functions, with $y = (y_1, ..., y_n)$ e $\theta \in \Theta \subset \mathbb{R}^d$. Knowing that, in parametric models, the correctness of the score function is the main assumption in showing that the MLE is consistent, it can be proven that this result still holds true for any unbiased estimating function. Specifically, if, for a law of large numbers,

$$\frac{1}{n} q(Y;\theta) \xrightarrow{p} 0 \text{ for } n \to +\infty,$$

when θ is the true value of the parameter, and if, given y fixed, $q(y;\theta)$ is a bijective and continuous function for θ in a neighborhood of the real value of the parameter, then the estimator defined by the solution for θ of $q(y;\theta) = 0$ is consistent, converging in probability to the true value of the parameter, for n diverging. Under the same regularity conditions required for a score function, the estimators based on $q(y;\theta)$ have an approximated normal distribution for n diverging. Let, then, $\tilde{\theta}$ be the solution of

$q(y;\theta) = 0$. Let be, also,

$$J(\theta) = E_\theta(q(Y;\theta)q(Y;\theta)^\top), \tag{2.5}$$

and

$$H(\theta) = -E_\theta\Big(\frac{\partial q(Y;\theta)}{\partial \theta^\top}\Big), \tag{2.6}$$

with $H(\theta)$ assumed symmetric. We note that, if $q(y;\theta)$ is the *score*, then $J(\theta) = H(\theta) = i(\theta)$. Through a local linearization we get

$$\theta = \tilde{\theta} + \Big(\frac{\partial}{\partial \theta^\top} g(y;\theta)\Big)^{-1} g(y;\theta) + ...$$
$$\tilde{\theta} - \theta = -\Big(\frac{\partial}{\partial \theta^\top} g(y;\theta)\Big)^{-1} g(y;\theta) + ...$$

If, with n diverging, for a central limit theorem,

$$q(Y;\theta) \dot{\sim} N_d(0, J(\theta)), \tag{2.7}$$

and, for a law of large numbers,

$$-\frac{\partial q(Y;\theta)}{\partial \theta^\top} \doteq H(\theta), \tag{2.8}$$

then

$$\tilde{\theta} - \theta \dot{\sim} N_d(0, H(\theta)^{-1} J(\theta) H(\theta)^{-1}), \tag{2.9}$$

with

$$Var(\tilde{\theta}) \doteq H(\theta)^{-1} J(\theta) H(\theta)^{-1}. \tag{2.10}$$

2.2.3 Definition of the quasi-likelihood function

With the aim of better formalizing what stated in section 2.2.1, we now are going to define the quasi-likelihood function. Supposing that we have a sample of independent observations $y_1, ..., y_n$ with expectation μ_i and variances $V(\mu_i)$, where V is a function of μ, then for each observation we define the quasi-likelihood function $K(y_i, \mu_i)$ by the relation

$$\frac{\partial K(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}, \tag{2.11}$$

or equivalently

$$K(y_i, \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i'}{V(\mu_i')} d\mu_i' + \text{function of } y_i \tag{2.12}$$

We wish, now, to express the function as if the mean of each observation was some known function of a set of parameters $\beta_1, ..., \beta_p$, such that the following assumptions of generalized linear models hold true:

$$E(Y_i) = \mu_i(\boldsymbol{x}_i\beta) = g^{-1}(\boldsymbol{x}_i\beta),$$

$$Var(Y_i) = \phi v(\mu_i),$$

$$\text{cov}(Y_i, Y_j) = 0 \text{ if } i \neq j,$$

where $\phi > 0$ is an unknown parameter, called dispersion parameter. In this scenario, the quasi-likelihood score function for each observation would be

$$\frac{\partial K(y_i, \beta)}{\partial \beta_r} = \frac{y_i - \mu_i}{Var(Y_i)} \frac{\partial \mu_i}{\partial \beta_r} \qquad r = 1, ..., p \tag{2.13}$$

Hence, the quasi-likelihood function is identical to the log-likelihood function if the the observations come from a one-parameter exponential family.

2.2.4 Properties of quasi-likelihood

The equation

$$\frac{\partial K(y_i, \beta)}{\partial \beta_r} = 0, \qquad r = 1, ..., p,$$

is easy to show to be unbiased, given $E(Y_i) = \mu_i = g^{-1}(\boldsymbol{x}_i\beta)$. Even when the observations do not come from an exponential family, the function K has still properties similar to those of log-likelihoods. In fact, also the following identity is maintained:

$$E\Big(\frac{\partial K}{\partial \beta_r}\frac{\partial K}{\partial \beta_s}\Big) = E\Big(\frac{\partial K^2}{\partial \beta_r \partial \beta_s}\Big). \tag{2.14}$$

We can easily demonstrate (**??**), showing that

$$\begin{aligned} E\Big(\frac{\partial K}{\partial \beta_r}\frac{\partial K}{\partial \beta_s}\Big) &= \frac{1}{\phi^2} E\Big(\sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{v(\mu_i)} \frac{\partial \mu_i}{\beta_r} \sum_{j=1}^{n} \frac{(Y_j - \mu_j)}{v(\mu_j)} \frac{\partial \mu_j}{\beta_s}\Big) \\ &= \frac{1}{\phi^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial \mu_i}{\beta_r}\frac{\partial \mu_j}{\beta_s} \frac{1}{v(\mu_i)}\frac{1}{v(\mu_j)} E[(Y_i - \mu_i)(Y_j - \mu_j)] \\ &= \frac{1}{\phi} \sum_{i=1}^{n} \frac{1}{v(\mu_i)} \frac{\partial \mu_i}{\partial \beta_r}\frac{\partial \mu_i}{\partial \beta_s}. \end{aligned}$$

On the other hand, we have

$$E\Big(\frac{\partial K^2}{\partial\beta_r\partial\beta_s}\Big) = \frac{1}{\phi}E\bigg(\sum_{i=1}^{n}\Big[(y_i-\mu_i)\frac{\partial}{\partial\beta_s}\Big(\frac{1}{v(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\Big) - \frac{1}{v(\mu_i)}\frac{1}{v(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s}\Big]\bigg)$$
$$= \frac{1}{\phi}\sum_{i=1}^{n}\frac{1}{v(\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s}.$$

Because of what said in the previous paragraph, and thanks to these properties, the consistency of the estimate $\hat{\beta}$ is maintained, along with the approximation

$$\hat{\beta} \dot{\sim} N(\beta, (X^\top W X)^{-1}), \tag{2.15}$$

where

$$W = \text{diag}(w_i), \qquad \text{with } w_i = \frac{1}{(g'(\mu_i))^2 Var(Y_i)}, \qquad i = 1, ..., n. \tag{2.16}$$

Furthermore, as already shown in section 2.2.1, the estimating equation (**??**) is the one with minimum variance among all the unbiased generating equations - we can see this result as a generalization of the Gauss-Markov Theorem. As McCullagh (1983) proved, this property still holds true for the estimator $\hat{\beta}$ given by the solution of (**??**) ; in fact, for any linear function of $a^\top\beta$ - with $a \in \mathbb{R}^p$- $a^\top\hat{\beta}$ will have asymptotically the smallest possible variance.

Focusing, now, on estimating the variance of the estimator $\hat{\beta}$, we can proceed in two ways:

i) we can estimate

$$\widehat{Var}(\hat{\beta}) = (X^\top \widehat{W} X)^{-1}, \tag{2.17}$$

where $\hat{W}$ is the W matrix defined in (**??**) with $\mu_i = \hat{\mu_i} = \boldsymbol{x}_i\beta$ and $\phi = \tilde{\phi}$. Given that the assumption related to the relationship between mean and variance holds true, the estimate (**??**) is consistent, meaning that - if multiplied by n - converges in probability, for n diverging, to the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$.

ii) The other way around for estimating the covariance matrix of $\hat{\beta}$ is to define another estimate of $Var(\hat{\beta})$, robust with the respect to the assumption regarding the specific relation between mean and variance. In order to do that, let us consider that, not taking into account ϕ, the sum of (**??**) over all the observations can be written as

$$q^*(y;\beta) = X^\top V_0^{-1}(y-\mu) = 0, \tag{2.18}$$

where $(y-\mu)^\top = (y_1-\mu_1, ..., y_n-\mu_n), V_0 = \text{diag}(g'(\mu_i)v(\mu_i))$.

Then, let us consider the variance of $\hat{\beta}$ as in (**??**), considering that $\hat{\beta}$ is the estimator of an unbiased estimating equation. Hence,

$$Var(\hat{\beta}) = H(\beta)^{-1} J(\beta) H(\beta)^{-1}, \tag{2.19}$$

with

$$H(\beta) = -E\Big(\frac{\partial q^*(Y;\beta)}{\partial \beta^\top}\Big) = -\frac{\partial (X^\top V_0^{-1})}{\partial \beta^\top} E(Y-\mu) + X^\top V_0^{-1}\Big(\frac{\partial \mu(\beta)}{\partial \beta^\top}\Big) = X^\top V_0^{-1} D,$$

where D is the matrix with generic element

$$d_{ir} = \frac{\partial \mu_i}{\partial \beta_r} = \frac{\partial \mu_i}{\partial \eta_i} x_{ir} = \frac{1}{g'(\mu_i)} x_{ir},$$

and

$$J(\beta) = Var(q^*(Y;\beta)) = X^\top V_0^{-1} Var(Y) V_0^{-1} X. \tag{2.20}$$

Hence, noting that $V_0^{-1} D = \phi W X$, we have that

$$Var(\hat{\beta}) \doteq \Big(\frac{1}{\phi}\Big)^2 (X^\top W X)^{-1} X^\top V_0^{-1} Var(Y) V_0^{-1} X (X^\top W X)^{-1}. \tag{2.21}$$

If $Var(Y) = \phi \,\mathrm{diag}(v(\mu_i))$, according to the assumptions regarding the relationship between mean and variance, we have that $V_0^{-1} Var(Y) V_0^{-1} = \phi^2 W$, and the whole expression turns out to be the former $Var(\hat{\beta}) \doteq (X^\top W X)^{-1}$. A robust estimate can be obtained estimating $Var(Y)$ with $\mathrm{diag}((y_i - \hat{\mu}_i)^2)$, and it would result in

$$\widehat{Var_R(\hat{\beta})} = \Big(\frac{1}{\tilde{\phi}^2}\Big)^2 (X^\top \hat{W} X)^{-1} X^\top \hat{V}_0^{-1} \,\mathrm{diag}((y_i - \hat{\mu}_i)^2) \hat{V}_0^{-1} X (X^\top \hat{W} X)^{-1}, \tag{2.22}$$

with $\hat{V}_0 = \mathrm{diag}(g'(\hat{\mu}_i) v(\hat{\mu}_i))$.

2.3 Generalized estimating equations

2.3.1 Longitudinal data

For all the models that we have encountered so far, we never debated on the assumption of independence of the response variable Y: we always assumed that the sample of study had independent observations. Independence however, is an assumption that does restrict our realm of interest, ruling out several phenomena where instead the observations are, by structure, strictly dependent on each other. Longitudinal data are a blatant case where this condition subsists. Longitudinal data track the same type of

information on the same subjects at multiple points in time. An example may regard patients that are monitored several times a year. Intuitively, the observations regarding the same patient are likely to be dependent on each other, whereas it is still plausible to assume independence among observations belonging to different patients. In order to find a method that accounts for this condition, we firstly set out our notation, defining

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{im_i})^\top,$$

the vector of observations related to the i-th subject - Y_{ij} being a scalar - and

$$X_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, ..., \boldsymbol{x}_{im_i})^\top,$$

the matrix of covariates related to the $i-$th subject, with $\boldsymbol{x}_{ij}$ being the p-dimensional vector of covariates related to the Y_{ij} observation. To deal with these type of data, there is a class of estimating equations that yields to consistent estimates of the parameters and of their variance, under mild assumptions on the dependence among observations belonging to the same subject.

2.3.2 Independence estimating equations

The problem of dealing with correlated data results from the difficulty of defining a likelihood function, and hence, outline a parametric model, when the response variable is not normal in its distribution. For this reason, it is plausible to focus our interest on the quasi-likelihood, being it a method that does not require the specification of the entire joint distribution of the observations. Then, if we start from considering the observations as independent also within the same subject, we can use a quasi-likelihood model, setting the required assumptions on the mean and variance - $E(Y_{ij}) = \mu_{ij}, \quad Var(Y_{ij}) = v(\mu_{ij})$- to estimate the parameters β. Then, we would have, setting for simplicity $m_1 = m_2 = ... = m_n = m$, the quasi-likelihood score to be

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{y_{ij} - \mu_{ij}}{\phi v(\mu_{ij})} \frac{\partial \mu_{ij}}{\partial \beta_r}, \qquad r = 1, ..., p. \tag{2.23}$$

The variance of the estimator would have the form of (**??**). Adapting it to the current notation of longitudinal data, we have

$$Var(\hat{\beta}) \doteq \Big(\frac{1}{\phi}\Big)^2 \Big(\sum_{i=1}^{n} X_i^\top W_i X_i\Big)^{-1} \Big(\sum_{i=1}^{n} X_i^\top V_{0i}^{-1} Var(\mathbf{Y}_i) V_{0i}^{-1} X_i\Big) \Big(\sum_{i=1}^{n} X_i^\top W_i X_i\Big)^{-1},$$

where

$$W_i = \text{diag}(w_{ij}), \qquad \text{with } w_{ij} = \frac{1}{(g'(\mu_{ij}))^2 v(\mu_{ij})\phi}, \qquad j = 1, ..., m, i = 1, ..., n,$$

and

$$V_{0i} = \text{diag}(g'(\mu_{ij})v(\mu_{ij})) = \text{diag}\left(\left(\frac{\partial\theta_{ij}}{\partial\eta_{ij}}\right)^{-1}\right), \qquad j = 1, ..., m, i = 1, ..., n,$$

with $Var(\mathbf{Y}_i)$ estimable, for example, as $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$.

Using this method implies assuming the data to be uncorrelated. This assumption, when not reflecting the reality of the data, does not effect the consistency of $\hat{\beta}$ and of $\widehat{Var(\hat{\beta})}$, even though it does affect their efficiency.

2.3.3 Generalized estimating equations

In order to increase the efficiency of the estimator, we now present a class of estimating equations that take the correlation into account in such a way that

- the estimator of $\hat{\beta}$ is still consistent.
- the variance estimate is consistent under the assumption that a weighted average of the estimated correlation matrices converges to a fixed matrix. So, in order to explain the correlation within the single subject, we define the matrix

$$V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}} \phi, \tag{2.24}$$

  where $A_i = \text{diag}(v(\mu_{ij}))$, $R(\alpha)$ is a $n \times n$ symmetric matrix that fulfill the requirement of being a correlation matrix, with α being a $s \times 1$ vector that fully characterizes $R(\alpha)$. If $R(\alpha)$ is the real correlation matrix of the data, then V_0 corresponds to $\text{cov}(Y_i)$.

In this way, we can define the general estimating equations as

$$\sum_{i=1}^{n} D_i^\top V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.25}$$

with D an $m \times p$ matrix with generic element

$$d_{ijr} = \frac{\partial\mu_{ij}}{\partial\beta_r}, \quad j = 1, ..., m, \ r = 1, ..., p,$$

that can also be written as $D_i = A_i \Delta_i X_i$, where $\Delta_i = \text{diag}(\partial\theta_{ij}/\partial\eta_{ij})$. We shall note that when we specify $R(\alpha)$ as the identity matrix, we obtain an independence estimating equation, namely the one discussed in the previous section.

The above equation can be expressed as a function of β alone by replacing α by $\hat{\alpha}(Y, \beta, \phi)$, an asymptotically consistent estimator of α given β and ϕ to be known, and ϕ by $\tilde{\phi}(Y, \beta)$, an asymptotically consistent estimator of ϕ when β is known. Consequently, we can write the equation as

$$\sum_{i=1}^{n} U_i[\beta, \hat{\alpha}\{\beta, \tilde{\phi}(\beta)\}] = 0, \tag{2.26}$$

and call $\hat{\beta}_G$ the solution of the equation, to be solved through an iterative procedure.

Large-sample property for $\hat{\beta}_G$

Under mild regularity conditions and given that:

- $\hat{\alpha}$ is asymptotically consistent given β and ϕ,
- $\tilde{\phi}$ is asymptotically consistent given β,
- $|\partial\hat{\alpha}(\beta, \phi)/\partial\phi| \leqslant H(Y, \beta)$, with $H(Y, \beta)$ being a consistent estimator for β,

then $\sqrt{n}(\hat{\beta}_G - \beta)$ is asymptotically multivariate Gaussian with zero mean and covariance matrix asymptotically equivalent to

$$V_G = n\left(\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} D_i^\top V_i^{-1} Var(\mathbf{Y}_i) V_i^{-1} D_i\right) \left(\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\right)^{-1}$$

The estimator of this matrix is obtained replacing β with $\hat{\beta}_G$, α with $\hat{\alpha}$ and ϕ with $\tilde{\phi}$. Furthermore, $Var(\mathbf{Y}_i)$ is estimated through $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top$. If the covariance matrix is correctly specified, then $Var(\mathbf{Y}_i) = V_i$ for $V_G = (\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i)^{-1}$.

As in the independence case - and in quasi-likelihood models - the consistency of $\hat{\beta}_G$ and $\hat{V}_G$ depends solely on the correct specification of the mean, not on the correct choice of R. However, the correct specification of R effects efficiency.

2.3.4 Parameters estimation in GEE

To compute the equation that solves for $\hat{\beta}_G$, we use an iterative method similar to Fisher's scoring - the one used in Generalized Linear Models (and hence in quasi-likelihood models too) - replacing α and ϕ with the estimates $\hat{\alpha}$ and $\tilde{\phi}$. The procedure is the following:

$$\hat{\beta}_{j+1} = \hat{\beta}_j - \left(\sum_{i=1}^{n} D_i^\top(\hat{\beta}_j) \hat{V}_i^{-1} D_j(\hat{\beta}_j) \right)^{-1} \left(\sum_{i=1}^{n} D_i^\top(\hat{\beta}_j) \hat{V}_i^{-1}(\hat{\beta}_j)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \right)$$

where $\hat{\mu}_{ij} = g^{-1}(\boldsymbol{x}_i \hat{\beta}_j)$ and $\hat{V}_i(\beta) = V_i[\beta, \{\hat{\alpha}, \tilde{\phi}(\beta)\}]$.

Now, let us define $D = (D_1^\top, ..., D_n^\top)^\top, S = ((\boldsymbol{y}_1 - \boldsymbol{\mu}_1)^\top, ..., (\boldsymbol{y}_i - \boldsymbol{\mu}_n)^\top)$ and let $\hat{V}$ be a $nm \times nm$ block diagonal matrix with $\hat{V}_i$'s as the diagonal elements. If we further define the modified dependent variable

$$Z = D\beta - S, \tag{2.27}$$

then, the iterative procedure above for calculating $\hat{\beta}_G$ is equivalent to performing an iteratively reweighted linear regression of Z on D with weight $\hat{V}^{-1}$.

Estimators of α and ϕ

At a given iteration the correlation parameters α and scale parameter ϕ can be estimated from the current Pearson residuals defined by

$$\hat{r}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{v(\hat{\mu}_{ij})^{\frac{1}{2}}}, \tag{2.28}$$

with $\hat{\mu}_{ij}$ depending by the current value for β. We can estimate ϕ as a generalized form of the (1.58), by

$$\tilde{\phi} = \sum_{i=1}^{n} \sum_{i=1}^{m} \frac{\hat{r}_{ij}^2}{(N-p)}, \tag{2.29}$$

with $N = m \cdot n$. It is easily shown that $\tilde{\phi}$ is a consistent estimator given the fourth moment of the response variable to be finite. The specific estimator of α depends upon the choice of $R(\alpha)$. The general approach is to estimate α by a simple function of

$$\hat{R}_{jh} = \sum_{i=1}^{n} \frac{\hat{r}_{ij} \hat{r}_{ih}}{(N-p)}. \tag{2.30}$$

We shall still keep in mind that the distribution of $\hat{\beta}_G$ does not depend on the specific choice of α and ϕ among those that are consistent.

2.3.5 Examples of correlation matrices

Above we treated the case where $R = I$ and we noted that, for any R, the estimators $\hat{\beta}_G$ and $\hat{V}_G$ will be consistent. Choosing R closer to the true correlation increases the

efficiency. We now enlist some possible correlation structures:

Case 1: one-by-one dependence

Let us suppose $\alpha = (\rho_1, ..., \rho_{m-1})^\top$, where $\rho_j = \text{corr}(Y_{ij}, Y_{i(j+1)})$ for $j = 1, ..., m-1$. Given β and ϕ, then we can estimate ρ_j by

$$\begin{aligned}\hat{\rho}_j &= \sum_{i=1}^{n} \frac{(y_{ij} - \hat{\mu}_{ij})(y_{ij} - \hat{\mu}_{i(j+1)})}{\sqrt{\phi v(\hat{\mu}_{ij})}\sqrt{\phi v(\hat{\mu}_{i(j+1)})}} \frac{1}{(K-p)} \\ &= \frac{1}{\phi} \sum_{i=1}^{n} \frac{\hat{r}_{ij}\hat{r}_{i(j+1)}}{(K-p)} \qquad j = 1, ..., m-1\end{aligned}$$

Let us now define $R(\alpha)$ as the tridiagonal matrix with $R_{i(i+1)} = \rho_j$. So

$$R(\alpha) = \begin{pmatrix} 1 & \rho_1 & 0 & 0 & ... & 0 & 0 \\ \rho_1 & 1 & \rho_2 & 0 & ... & 0 & 0 \\ 0 & \rho_2 & 1 & \rho_3 & ... & 0 & 0 \\ \vdots & & & \vdots & & & \vdots \\ 0 & 0 & 0 & ... & \rho_{m-2} & 1 & \rho_{m-1} \\ 0 & 0 & 0 & 0 & ... & ...\rho_{m-1} & 1 \end{pmatrix}$$

This is equivalent to the one-dependent model. We shall note that an estimator of ϕ is unnecessary for calculating $\hat{\beta}_G$ and $\hat{V}_G$ when using this structure as the ϕ in the formula cancels when calculating V_i. For the special case where $\alpha = \rho_1 = \rho_2 = ... = \rho_{m-1}$, we can estimate the common α by

$$\hat{\alpha} = \sum_{j=1}^{n} \frac{\hat{\rho}_j}{(n-1)}.$$

Case 2: Exchangeable correlation structure

Let us assume that $\text{corr}(y_{ij}, y_{ij'}) = \rho = \alpha$ for all $j \neq j'$. Given ϕ, α can be estimated by

$$\hat{\alpha} = \frac{1}{\phi} \frac{\sum_{i=1}^{n} \sum_{j>j'}^{m} \hat{r}_{ij}\hat{r}_{ij'}}{\frac{1}{2}m(m-1)n - p}, \tag{2.31}$$

We shall note that with this assumption, a differing number of observations for each subject is possible. In that case, the estimator would be

$$\hat{\alpha} = \frac{1}{\phi} \frac{\sum_{i=1}^{n} \sum_{j>j'}^{m} \hat{r}_{ij}\hat{r}_{ij'}}{\sum_{i=1}^{n} \frac{1}{2}m_i(m_i-1) - p}. \tag{2.32}$$

Case 3: Autoregressive correlation structure

Let $\text{corr}(y_{ij}, y_{ij'}) = \alpha^{|j-j'|}$. If y_{ij} is Gaussian, this is the correlation structure of the continuous time analogue of the first-order autoregressive process, $AR-1$.

Case 4: Unspecified correlation

Let us now take $R(\alpha)$ as completely unspecified. The number of parameters to be estimated would be $\frac{1}{2}n(n-1)$. In this case, R can be estimated by

$$R = \frac{\phi}{n}\sum_{i=1}^{n} A_i^{-\frac{1}{2}}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top A_i^{-\frac{1}{2}}. \tag{2.33}$$

We shall note that, when considering this correlation structure, the G.E.E. would be identical to the likelihood score if the $\boldsymbol{Y}_i$'s followed a multivariate Gaussian distribution. Furthermore, the asymptotic variance V_G reduces to

$$V_G = \Big(\sum_{i=1}^{n} D_i^\top Var(\boldsymbol{Y}_i)^{-1} D_i/n\Big)^{-1}, \tag{2.34}$$

with $Var(\boldsymbol{Y}_i)$ being the covariance matrix of $\boldsymbol{Y}_i$. This is because R is the true correlation matrix.

Chapter 3

Akaike's information Criterion

3.1 Introduction

In this chapter, we introduce the Akaike's information criterion along with its main properties. We then look for a way to widen its use over the generalized estimating equations, a method that, as we have seen in chapter 2, does not need a fully-known distributional form to be used.

3.2 Kullback Leibler divergence

In order to fully understand the Akaike's information criterion, we firstly shall introduce the Kullback-Leibler divergence. In fact, considering that the AIC aims at providing us with a means to enhance model selection, it would be logical to think of it as somehow related to a measure that assesses the 'distance' between two statistical models. This (non-symmetric) measure is called the Kullback-Leibler divergence. Letting $p(x)$ and $p_0(x)$ be two probability distribution, if X is a discrete random variable, then the KL divergence of $p(x)$ from $p_0(x)$ is defined as

$$\Delta\big[p(x), p_0(x)\big] = \sum_{x \in X} p_0(x) \log \frac{p_0(x)}{p(x)} = E_{p_0}\Big[\log \frac{p_0(x)}{p(x)}\Big]. \tag{3.1}$$

In the continuous case, it would be

$$\Delta\big[p(x), p_0(x)\big] = \int_{x \in X} p_0(x) \log \frac{p_0(x)}{p(x)} = E_{p_0}\Big[\log \frac{p_0(x)}{p(x)}\Big]. \tag{3.2}$$

It can be viewed as a measure of the information lost when $p(x)$ is used to approximate $p_0(x)$. Conceptually, it is strictly related to information entropy. In fact

$$\begin{aligned}\Delta\big[p(x), p_0(x)\big] &= \sum_{x\in X} p_0(x) \log \frac{p_0(x)}{p(x)} \\ &= \sum_{x\in X} p_0(x) \log p_0(x) - \sum_{x\in X} p_0(x) \log p(x) \\ &= -\sum_{x\in X} p_0(x) \log p(x) - H(p_0(x))\end{aligned}$$

where $H(p_0(x))$ is the entropy of the random variable $X \sim p_0(x)$. Then, substituting in (3.1) the natural logarithm with the logarithm base 2, we can view $\Delta\big[p(x), p_0(x)\big]$ as the expected number of extra bits required to code samples from $p_0(x)$ when using a code based on $p(x)$. Hence, $p_0(x)$ corresponds to the true distribution of the data, and $p(x)$ to its approximation. Restricting the notation to our context, we will from now on refer to $p_0(x)$ as $p(x;\theta_0)$ and to $p(x)$ as $p(x;\theta)$. Hence, the two distributions differ only for the value of θ. Anyhow, we should note that the KL divergence is not really a distance measure. In fact, the following properties do not make it so:

- it is not symmetric. The divergence from $p_0(x)$ to $p(x)$ is generally different from the one from $p(x)$ to $p_0(x)$, and
- it does not satisfy the triangular inequality.

Still, as any distance measure, it is always a positive value or 0 when $p(x) = p_0(x)$.

3.3 AIC

In introducing an index useful for comparing models with each other, let us first determine what we intend by model comparison in the first place. Let us suppose to be wanting to choose the best model for a sample of data y, among k models such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_k$ with their correspondent parametric spaces being

$$\Theta_1 \subset \Theta_2 \subset ... \subset \Theta_k \subset \mathbb{R}^k$$

We refer to the k models parameters as $\theta^{(1)}, ..., \theta^{(k)}$. For simplicity, we suppose that the parametric space $\Theta_k = \{\theta^{(k)} = (\theta_1, ..., \theta_k)\}$ is equivalent to Θ_{k-1} under the hypothesis $\theta_k = 0$, and that this is true for all models when other components are made null. Let then $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, ..., \hat{\theta}^{(k)}$ be the MLE estimates for each different model. Then, the models' maximized log-likelihoods are $l(\hat{\theta}^{(1)}; y), ..., l(\hat{\theta}^{(k)}; y)$. Wanting to choose the model that

fits the data the best, we should note that the likelihoods are of no use by themselves, because, necessarily,

$$l(\hat{\theta}^{(1)}; y) < l(\hat{\theta}^{(2)}; y) < ... < l(\hat{\theta}^{(k)}; y),$$

for the best model would always result to be the k-dimensional one. Be θ^0 the real value of the parameter. We then assume that $\theta^0 \in \Theta_k$, and that $\theta^0 \in \Theta_{d^*}$, with $d^* \in \{1, ..., k\}$ such that $\theta^0 \notin \Theta_{d*-1}$. d^* is the dimension of the most parsimonious model among those that are correctly specified. Let us now further consider y^*, a set of observations that are yet to be sampled and that are generated by the same process that generated y. Then y^* comes from the random variable Y^*, with probability function $p_Y(y^*; \theta^0)$. This density represents the best means for making predictions about y^*. Being θ^0 unknown, we do not know $p_Y(y^*; \theta^0)$; however, if $\mathcal{F}_d$ is a correctly specified model with parameter $\theta^{(d)}$, we may estimate $p_Y(y^*; \theta^0)$ with $p_Y(y^*; \hat{\theta}^{(d)})$, where $\hat{\theta}^{(d)} = \hat{\theta}^{(d)}(y)$. We would typically have

$$p_Y(y^*; \hat{\theta}^{(d)}) < p_Y(y; \hat{\theta}^{(d)})$$

being y the sample that estimated $\hat{\theta}^{(d)}$. Furthermore, we would have that

$$E_0\Big[\log p_Y(Y^*; \hat{\theta}^{(d)}(Y))\Big] = E_0\Big[l(\hat{\theta}^{(d)}(Y); Y^*)\Big]. \tag{3.3}$$

Hence, we may consider (3.3) as a means to assess model prediction, when the number of parameters is d. Our aim, then, translates in finding a way to maximize (3.3). At this point, we can rephrase our problem if we consider that finding the maximum of the above expected value equals finding the minimum of the KL divergence from $p_Y(Y^*; \theta_0)$ to $p_Y(Y^*; \hat{\theta}^{(d)}(Y))$, defined as

$$\Delta[\theta_0, \theta] = E_0\Big[\log \frac{p_Y(Y^*; \theta_0)}{p_Y(Y^*; \hat{\theta}^{(d)}(Y))}\Big]. \tag{3.4}$$

In finding a way to compute the minimum of (3.4), we first intend to highlight some important aspects of the above formula that would help us better understand the logic on which the AIC is built:

i) We should firstly note that $p_Y(Y^*; \theta_0)$ does not need to be estimated. For our purpose, we need just a measure of the relative distance between the two distributions: we could still measure the KL divergence up to a constant C - namely $p_Y(Y^*; \theta_0)$ - that - being the same for every model - would not hinder us from using the result

as a means for model comparison. In fact, given $\mathcal{F}_{\theta_1}$ and $\mathcal{F}_{\theta_2}$ two models, if

$$\Delta[\theta_0, \theta_1] < \Delta[\theta_0, \theta_2]$$

with $\mathcal{F}_{\theta_0}$ being the true model, so that $\mathcal{F}_{\theta_1}$ is best, then

$$\Delta[\theta_0, \theta_1] - C < \Delta[\theta_0, \theta_2] - C$$

. Moreover,

$$\Delta[\theta_0, \theta_2] - \Delta[\theta_0, \theta_1] \equiv -E_0\big[\log(p_2(y;\theta_2))\big] + E_0\big[\log(p_1(y;\theta_1))\big]$$

so we know how much better $\mathcal{F}_{\theta_1}$ is with the respect to $\mathcal{F}_{\theta_2}$.

ii) Secondly, we should keep in mind that we are working with two "levels of uncertainty"; what we mean, is that $\Delta[\theta_0, \theta]$ depends on d, as well as on the value of the estimator $\hat{\theta}^{(d)}(Y)$. Hence, even under the assumption of $d = d^*$, $\hat{\theta}^{(d)}(Y)$ would generally produce estimates not equal to θ_0. In this context, we may readjust our problem, aiming for the value that minimizes

$$E_Y\big[\Delta[\theta_0, \theta_1]\big], \tag{3.5}$$

knowing that

$$E_Y\big[\Delta[\theta_0, \theta_1]\big] > \Delta[\theta_0, \theta_1].$$

Akaike showed that, for n diverging, if $\mathcal{F}_d$ is correctly specified, then

$$E_0\Big[\log p_y(Y^*; \hat{\theta}^{(d)}(Y))\Big] \doteq E_0\Big[l(\hat{\theta}^{(d)}; Y) - d\Big].$$

From this result, we define the Akaike Information Criterion with the respect to model $\mathcal{F}_d$ as

$$AIC(\mathcal{F}_d) = 2d - 2l(\hat{\theta}^{(d)}; y). \tag{3.6}$$

At this point, we intend to provide the reader with a spark of the intuition residing behind (3.6), in order to understand the role each component plays in relation to the information provided by the index.

3.3.1 Insight on AIC derivation

In the previous section we stated that our "readjusted" aim, was to minimize the quantity (3.5), that is

$$E_Y^*\Big[E_0\Big[\log\frac{p_Y(Y^*;\theta_0)}{p_Y(Y^*;\hat{\theta}^{(d)}(Y))}\Big]\Big].$$

..... to finish.....

3.3.2 Selection Bias in AIC

The AIC tends to select models slightly over parameterised, as, for $d > d^*$

$$Pr_{\theta^0}(AIC(\mathcal{F}_d) < AIC(\mathcal{F}_d^*)) = Pr_{\theta_0}\Big(2\Big(l(\hat{\theta}^{(d)};y) - l(\hat{\theta}^{(d^*)};y)\Big) > 2(d-d^*)\Big) \tag{3.7}$$

$$\doteq Pr(\chi^2_{d-d^*} > 2(d-d^*)) \tag{3.8}$$

for n large. Hence, the probability of selecting the over parameterised model $\mathcal{F}_d$ does not converge to 0 for n diverging. Thus, the criteria is not consistent. Specifically, $Pr_{\theta_0}(AIC(\mathcal{F}_d) < AIC(\mathcal{F}_d^*))$ is approximately equal to 0.16 when $d - d^* = 1$, to 0.07 when $d - d^* = 2$, to 0.03 when $d - d^* = 3$. As a consistent alternative, the BIC criterion can be used: based on a bayesian approach, it is defined as

$$BIC(\mathcal{F}_d) = d\log n - 2l(\hat{\theta}^{(d)};y).$$

Even though BIC results to be consistent, for large n the $\log n$ penalization weights too much, causing the criterion to select underparameterised models.

3.3.3 Title of subsection

?

3.3.4 Title of subsection

?

3.3.5 Title of subsection

?

3.4 AIC with quasi-likelihood function

Appendix

Bibliography

Azzalini, A. (2001) *Inferenza Statistica. Una Presentazione Basata sul Concetto di Verosimiglianza*. Milano: Springer-Verlag Italia.

Bartlett, M. S. (1953) Approximate confidence intervals. II. More than one unknown parameter. *Biometrika* **40**, 306–317.

DiCiccio, T. J. and Stern, S. E. (1993) An adjustment to profile likelihood based on observed information. Technical report, Department of Statistics, Stanford University.

Kosmidis, I. (2016) `brglm2`*: Estimation and inference for generalized linear models using explicit and implicit methods for bias reduction.*
`https://github.com/ikosmidis/brglm2`.

Stafford, J. E. (1992) *Symbolic Computation and the Comparison of Traditional and Robust Test Statistics* (unpublished doctoral dissertation). University of Toronto, Canada.