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.ssa 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

Chapter 1

Models based on Maximum Likelihood Estimation

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 ... ...

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 to as our model of interest. Formally, we define a parametric model $\mathcal{F}$ as

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

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 $L : \Theta \rightarrow \mathbb{R}^p$ as

$$L(\theta) = L(\theta; y) = c(y)f(y; \theta),$$

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

$$l(\theta) = l(\theta; y) = \log L(\theta; y)$$

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

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

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 $\mathbf{Y} = (Y_1, ..., Y_n)$.

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

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

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 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

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

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

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

1.2.4 Important Likelihood properties and theorems

-GAUSS MARKOV

-RAO-CRAMER

...

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 $u \in \mathbb{R}^p$, $\left\|u\right\|^2 = u^\top u$ 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

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

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

The Maximum Likelihood Estimate is

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

and

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

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

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

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

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

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

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

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

$$e = y - \hat{y} = y - X\beta$$

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:

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

A quantity that well explains in which terms the residuals are connected to the variability of the response variable Y, is the following equality

$$y^\top y = \hat{y}^\top\hat{y} + e^\top e$$

which is equal to

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

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:

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

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 to 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 the 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 coefficient R^2 will always increase whenever adding a new variable, even if the variable itself was to be meaningless with the respect to the explanation of 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, a corrected version that takes into account the number of parameters used is the adjusted version

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

1.3.4 Least square prediction: good enough?

1.4 Generalized Linear Models

Table 1.1: ML fit of the Gamma regression model with log-link and Wald 0.95 confidence intervals for the parameters.

	Estimate	Estimated Standard Error	0.95 Confidence Interval
β_1	0.361	0.250	(-0.128, 0.851)
β_2	1.507	0.170	(1.174, 1.839)
β_3	1.859	0.165	(1.535, 2.183)
ϕ	0.223	0.079	(0.069, 0.377)

Chapter 2

Models based on Quasi-Likelihood Estimation

2.1 Quasi-likelihood inference

2.2 Quasi-likelihood function

2.2.1

2.2.2 Title of subsection

2.2.3 Title of subsection

2.3 Generalized Estimating Equations

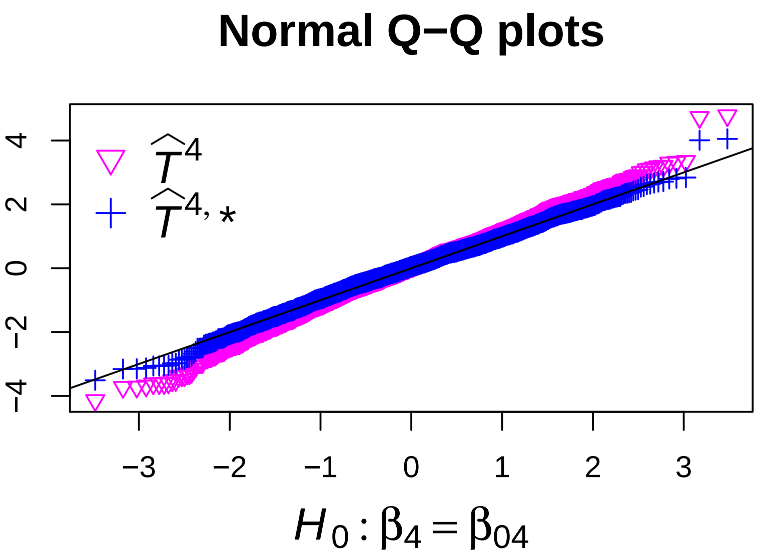

FIGURE 2.1: Normal Q-Q plots based on 2000 values of $\widehat{T}^4$ and $\widehat{T}^{4,*}$ computed under the null hypothesis $H_0 \colon \beta_4 = \beta_{04}$ in the *clotting* example.

Chapter 3

Akaike's information Criterion

3.1 Kullback Leibler divergence

Azzalini (2001)

3.2 AIC

Bartlett (1953)

3.2.1 Title of subsection

Kosmidis (2016)

3.2.2 Title of subsection

Stafford (1992)

3.2.3 Title of subsection

DiCiccio and Stern (1993)

3.3 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.