Linear Models and Regression II
Introductory Lecture

ginsbourger@stat.unibe.ch

Spring Semester 2017

Foreword

Welcome to the first lecture of the course “Linear Models and Regression II”!

The course will take place from today until the end of the semester (ending on Friday June 2nd 2017). Note that the Easter holidays will take place from Friday 14th to Sunday 23rd April 2017.

Usual times for the lecture and exercises will be respectively Mondays from 8.15AM to 10AM and Fridays from 2.15PM to 4PM in room B1 (here). An exception is March 13-17 for which lecture/exercise days will be swapped.

The (oral) exam will take place a few weeks after the end of the semester, around end of June/beginning of July 2017. Regarding the exercises, given by Dr. Dario Azzimonti, things will work similarly as for LMRI.

Beyond classical linear modelling

In the last semester we have focused on statistical models of the form

$$Y = \sum_{j=1}^{p} \theta_j f_j(X) + \epsilon,$$

where $Y \in \mathcal{Y}, X \in \mathcal{X}$ are respectively a response and a covariable (both possibly multivariate), $p \in \mathbb{N}\backslash\{0\}$, $f_j : \mathcal{X} \to \mathcal{Y}$ are basis functions with coefficients $\theta_j \in \mathbb{R}$ $(1 \leq j \leq p)$, and ϵ is a random error in $\mathcal{Y}$. We have typically assumed $\mathcal{Y} = \mathbb{R}, \mathcal{X} = \mathbb{R}^d$ and ϵ centred, Gaussian or at least L^2.

We have seen that this formulation encompasses a number of particular models (Simple and Multiple Linear Regression, Polynomial Regression ANOVA, ANCOVA, etc.) and also that under assumptions it is possible based on observed data $(X_i, Y_i)_{i=1,\ldots,n}$ $(n \in \mathbb{N}\backslash\{0\})$ to estimate $\boldsymbol{\theta}$ and also to perform tests, derive exact and approximate confidence regions, etc.

Beyond classical linear modelling

In a number of situations however, classical models and/or methods from LMRI are **not suitable**. This concerns notably the following cases:

- It is not reasonable to assume ϵ Gaussian or related as
  - Y is binary, categorical, integer-valued, or other scenarios for which an ϵ with distribution supported by the real-line does not make sense,
  - or Y and ϵ may well be real-valued but another distribution is assumed (then it might be that linear models are suitable but not least-squares)
- The relationship between Y and $\boldsymbol{\theta}$ is non-linear,
- Beyond usual assumptions but possibly still in the linear set up, it might be that ϵ_i's corresponding to different observations are dependent, that the design matrix does not have full column-rank, etc.

A detour through non-linear least squares

In cases where responses $\boldsymbol{Y} \in \mathbb{R}^n$ stem from a non-linear model of the form

$$Y = f(X, \boldsymbol{\theta}) + \epsilon$$

or in other words when $\boldsymbol{Y} = \boldsymbol{v}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$ where $\boldsymbol{v}(\boldsymbol{\theta})$ has components $f(X_i, \boldsymbol{\theta})$ and $\boldsymbol{\epsilon}$ is a centred squared-integrable noise, one common approach is to keep using the least-squares procedure.

A *Least Squares Estimator* $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is then defined by its realizations —assuming their existence— minimizing given $\boldsymbol{y}$ the loss function

$$Q : \boldsymbol{\eta} \in \mathbb{R}^p \mapsto Q(\boldsymbol{\eta}) = ||\mathbf{y} - \boldsymbol{v}(\boldsymbol{\eta})|||^2$$

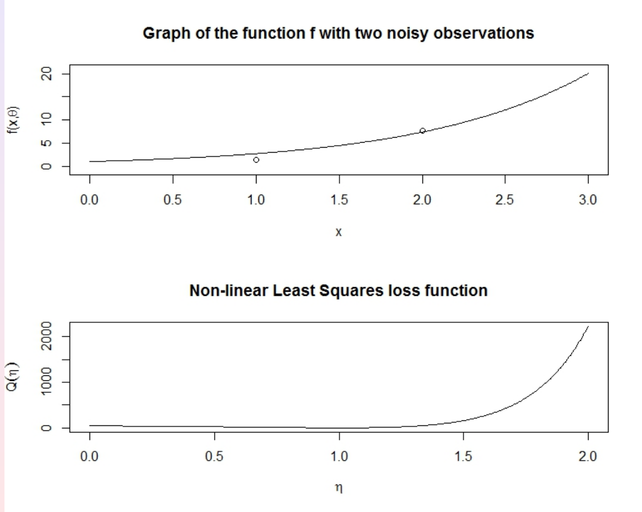
Conditions to ensure existence of the LS estimator —not discussed in detail here— classically involve Θ (e.g. assumed to be a subset of $\mathbb{R}^p$ with non-empty interior; compactness simplifies things but is not always assumed), $\boldsymbol{v}$ (typically assumed injective and twice continuously differentiable), and the ϵ_i's (typically assumed centred, squared-integrable and uncorrelated).

A detour through non-linear least squares

A simple and classical example follows (from “NonLinear Regression Modeling” (1983) by D.A. Ratkowsky, cited in “Regression non linéaire et applications” (1992) by A. Antoniadis, J. Berruyer and R. Carmona, a book that has largely inspired the first lectures of the current course).

Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and $Y_i = f(X_i, \theta) + \epsilon_i$ with $f(x; \theta) = \exp(\theta x)$, ϵ_i realizations of a centred Gaussian noise with given variance σ^2, and X_1, X_2 are also given. Here we take $\sigma = 2$, $X_1 = 1$ and $X_2 = 2$.

A detour through non-linear least squares



A detour through non-linear least squares

In this example the loss function is quite simple and it is even possible to calculate the (non-linear) LS estimator in closed form. In fact, we have

$$Q(\eta) = (y_1 - \exp(\eta))^2 + (y_2 - \exp(2\eta))^2$$

Noting that Q is here convex differentiable, we get by annulling its derivative

$$2\exp(3\widehat{\theta}) - (2y_2 - 1)\exp(\widehat{\theta}) - y_1 = 0.$$

Posing $\gamma = \exp(\widehat{\theta})$ we can solve for γ as real root of the third order polynomial $2\gamma^3 - (2y_2 - 1)\gamma - y_1$. With the (simulated) values $y_1 = 4.556237$ and $y_2 = 8.953329$, we get $\gamma = 2.7428935$ whereof $\widehat{\theta} = 1.009013$.

For more complicated cases (with larger n and/or p, more complex functions, etc.), solving non-linear LS analytically is out of reach. Most of the time, non-linear LS calls for **iterative fitting procedures**. Similar procedures will also be useful in more general frameworks as discussed in the next sections.

A detour through non-linear least squares

Assuming Θ with a non-empty interior and uncorrelated centred noise with constant variance, most iterative procedures for non-linear least squares are based on the following result:

Proposition

Any LS estimator $\widehat{\boldsymbol{\theta}}$ belonging to $\mathring{\Theta}$ is solution to the *normal equations*:

$$(\boldsymbol{Y}-\boldsymbol{v}(\boldsymbol{\eta}))^T\nabla\boldsymbol{v}(\boldsymbol{\eta})=0.$$

The *Gauss-Newton* algorithm addresses the issue of approaching solutions to this by relying on Taylor expansions of $\boldsymbol{v}$. Assuming that the current candidate parameter vector is $\boldsymbol{\theta}^{(i)}$, we have at order one

$$\boldsymbol{v}(\boldsymbol{\eta})\approx\boldsymbol{v}(\boldsymbol{\theta}^{(i)})+\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})^T(\boldsymbol{\eta}-\boldsymbol{\theta}^{(i)})$$

Replacing $\boldsymbol{v}$ by its local approximation in the normal equation, one gets

$$(\boldsymbol{y}-\boldsymbol{v}(\boldsymbol{\theta}^{(i)})-\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})^T(\boldsymbol{\eta}-\boldsymbol{\theta}^{(i)}))^T\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})=0$$

A detour through non-linear least squares

Expanding the former equation and assuming that the symmetric matrix $\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})^T\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})$ is positive definite (i.e., invertible, here), the principle of the Gauss-Newton algorithm is to define the next point as

$$\boldsymbol{\theta}^{(i+1)}=\boldsymbol{\theta}^{(i)}+\left(\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})^T\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})\right)^{-1}\nabla\boldsymbol{v}(\boldsymbol{\theta}^{(i)})\left(\boldsymbol{y}-\boldsymbol{v}(\boldsymbol{\theta}^{(i)})\right).$$

Note that another popular approach consists in directly taking a Taylor expansion of Q's gradient so as to get

$$\nabla Q(\boldsymbol{\eta})\approx\nabla Q(\boldsymbol{\theta}^{(i)})+\nabla^2 Q(\boldsymbol{\theta}^{(i)})(\boldsymbol{\eta}-\boldsymbol{\theta}^{(i)})$$

A solution to the normal equation is then sought by iteratively setting

$$\boldsymbol{\theta}^{(i+1)}=\boldsymbol{\theta}^{(i)}-\nabla^2 Q(\boldsymbol{\theta}^{(i)})^{-1}\nabla Q(\boldsymbol{\theta}^{(i)}).$$

This method is known as the *Newton-Raphson* algorithm. Note that in both cases the choice of a good starting point may be crucial.

On the way to Generalized Linear Models

In a number of situations, relations between observations and parameters $\boldsymbol{\theta}$ are not as direct as previously.

This includes notably cases where Y is binary, with a distribution $B(\rho)$ where the probability ρ (say of Y taking the value 1) is a function of X and $\boldsymbol{\theta}$. **Logistic regression** corresponds to a particular case of this.

This includes also cases where Y is integer-valued, for example with a Poisson distribution $\mathcal{P}(\lambda)$ where the intensity λ is a function of X and $\boldsymbol{\theta}$. **Poisson regression** corresponds to a particular case of this.

Before going into some overarching principles and investigating these models in more detail, let us first get a brief overview.

Beyond linear models: overview of logistic regression

In logistic regression, the output Y is binary, and has a probability ρ of taking the values 0 or 1 that depends on covariables X through a function parametrized by a vector $\boldsymbol{\theta}\in\mathbb{R}^p$.

In the logistic regression model, the following form is assumed for ρ:

$$\rho(x)=\frac{\exp(\sum_{j=1}^p\theta_j f_j(x))}{1+\exp(\sum_{j=1}^p\theta_j f_j(x))}$$

The function $h: t\in\mathbb{R}\rightarrow\exp(t)/(1+\exp(t))$ is the *standard logistic function*. Its reciprocal bijection is $\mathrm{logit}:\rho\in(0,1)\rightarrow\mathrm{logit}(\rho)=\log(\frac{\rho}{1-\rho})$. Reformulating the model in terms of the logit function results in

$$\mathrm{logit}(\rho(x))=\sum_{j=1}^p\theta_j f_j(x)$$

It may seem quite close to a linear model but it is important to stress that least squares estimation does not apply here. Estimating $\boldsymbol{\theta}$ based on binary observations hence calls for some specific procedure(s).

Beyond linear models: overview of logistic regression

A standard approach to estimate $\boldsymbol{\theta}$ in logistic regression is to maximize the so-called **likelihood function** of parameters given observations, i.e. to maximize the probability of observing $\boldsymbol{Y}$ seen as a function of $\boldsymbol{\theta}$.

Each Y_i is assumed to have been generated independently from a Bernoulli distribution with parameter $\rho(X_i)$, i.e. has probability

$$\rho(X_i)^y(1-\rho(X_i))^{1-y} = \frac{\exp(y\sum_{j=1}^{p}\theta_j f_j(X_i))}{1+\exp(\sum_{j=1}^{p}\theta_j f_j(X_i))}$$

of being equal to $y \in \{0,1\}$. The likelihood of $\boldsymbol{\eta}$ given a vector of observations $(y_1,\ldots,y_n)$ is then defined as

$$L(\boldsymbol{\eta};(y_1,\ldots,y_n)) = \prod_{i=1}^{n}\rho(X_i)^{y_i}(1-\rho(X_i))^{1-y_i} = \prod_{i=1}^{n}\frac{\exp(y_i\sum_{j=1}^{p}\eta_j f_j(X_i))}{1+\exp(\sum_{j=1}^{p}\eta_j f_j(X_i))}$$

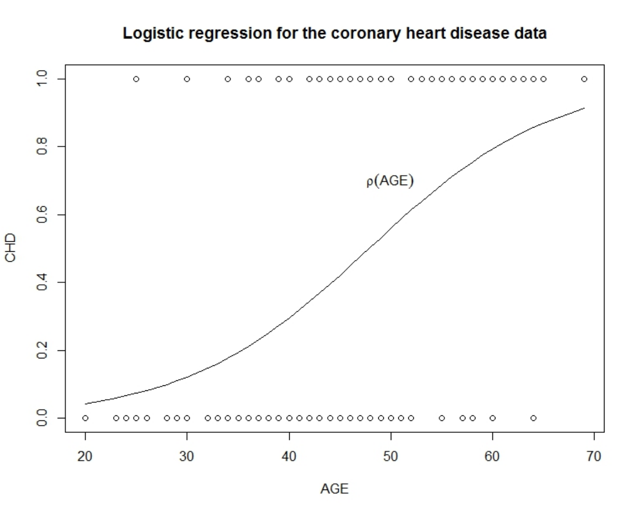
$\widehat{\boldsymbol{\theta}}$ is then found by numerically maximizing L. We will study that in more detail in next lectures, but let us now illustrate the model with a first example.

Beyond linear models: overview of logistic regression

Here we consider an example from a study on “Coronary Heart Disease” (Source: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013). Applied Logistic Regression: Third Edition. See `https://www.umass.edu/statdata/statdata/data/chdage.txt`).

The data set consists of $n = 100$ observations, with three variables: for each observation, an identification code, the age of the patient, and a binary variable indicating whether there is a coronary heart disease or not. Let us refer to the last one as Y, to the age as X and investigate the relationship between Y and X assuming a simple logistic regression model ($d = 1$).

Beyond linear models: overview of logistic regression



Beyond linear models: overview of logistic regression

For the sake of brevity we do not dive into the details of the fitted model here and much remains to be said about model interpretation, parameter estimation, confidence intervals and beyond. Along the next lectures, we will have the occasion to analyse the logistic regression model in more detail.

For now we however to a second model, Poisson regression, where the response is this time integer valued. As we will see later, these two seemingly distant examples actually share some essential properties.

Beyond linear models: overview of Poisson regression

Let us now focus on *count data*, and more particularly on regression when the response of interest is integer valued, i.e. $Y \in \mathbb{N}$.

The Poisson distribution is a standard distribution over $\mathbb{N}$. A r.v. Y is Poisson-distributed whenever its probability mass function is given by

$$P(Y = y) = \frac{\exp(-\lambda)}{y!}\lambda^y \text{ for } y \in \mathbb{N},$$

where the parameter $\lambda > 0$ is called *intensity*.
It is a well known fact that for $Y \sim \mathcal{P}(\lambda)$, $\mathbb{E}[Y] = \text{Var}[Y] = \lambda$.

In Poisson regression, Y is assumed to follow a Poisson distribution which intensity depends on the covariables X. More precisely, the logarithm of the intensity function is supposed to depend linearly on basis functions of X:

$$\log(\lambda(x)) = \sum_{j=1}^{p} \theta_j f_j(x).$$

Beyond linear models: overview of Poisson regression

Now, given observed $y_1, \dots, y_n$ from a Poisson regression model with corresponding covariable instances $X_1, \dots, X_n$, the parameters $\boldsymbol{\theta}$ are typically estimated by maximizing the likelihood function

$$L(\boldsymbol{\eta}; (y_1, \dots, y_n)) = \prod_{i=1}^{n} \left\{ \exp\left(-\exp\left(\sum_{j=1}^{p} \eta_j f_j(X_i) \right) \right) \frac{\left(\exp(\sum_{j=1}^{p} \eta_j f_j(X_i))\right)^{y_i}}{y_i!} \right\}$$

As we will see in the following, the maximization of L is not as hard as it may seem since rewriting the problem and digging into some more theory allow practical simplifications as well as theoretical guarantees.

Practical examples will be provided at a later stage, when covering more specifically the topic of Poisson regression. For now let us focus on Maximum Likelihood Estimation in a much more general framework.

About Maximum Likelihood Estimation

In the last two examples, the approach to fit model parameters has been to maximize so-called likelihood functions. We now review a few selected facts about this approach that is central in statistics and which domain of relevance goes well beyond logistic and Poisson regression.

Assume that (E, ν) is a measured space (often in our context $E = \mathbb{R}^n$ equipped with its Lebesgue measure) and that $(P_{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta}$ is a family of probability measures dominated by ν. Let us denote by $p(\cdot; \boldsymbol{\theta})$ the corresponding probability density/mass functions, and assume further that $\boldsymbol{Y}$ is a random element with distribution $P_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Theta$.

The Maximum Likelihood Estimation (MLE) approach consists, given a realization $\boldsymbol{y}$ of $\boldsymbol{Y}$, in estimating θ by maximizing (*when possible*) the function

$$L(\cdot; \boldsymbol{y}) : \boldsymbol{\eta} \in \Theta \to L(\boldsymbol{\eta}; \boldsymbol{y}) := p(\boldsymbol{y}; \boldsymbol{\eta}).$$

About Maximum Likelihood Estimation

As a motivating example, let us consider the particular case where there are only two alternative values for $\boldsymbol{\theta}$, i.e. $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$.

Assuming that one observes $\boldsymbol{Y}$ which is assumed to be drawn from $P_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \in \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, estimating $\boldsymbol{\theta}$ then amounts to dividing the set E into two disjoint subsets: the set E_1, with $\boldsymbol{\theta}$ estimated to be $\boldsymbol{\theta}_1$ when $\boldsymbol{Y} \in E_1$, and E_2 with $\boldsymbol{\theta}$ estimated to be $\boldsymbol{\theta}_2$ when $\boldsymbol{Y} \in E_2$. In other words,

$$\widehat{\boldsymbol{\theta}} = \begin{cases} \boldsymbol{\theta}_1 \text{ in case } \boldsymbol{Y} \in E_1, \\ \boldsymbol{\theta}_2 \text{ else.} \end{cases}$$

Now, nothing has been said yet on how to choose E_1.

The approach that we will follow here is to consider two kinds of error: the probability that $\boldsymbol{Y} \in E_1$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_2$ and the probability that $Y \in E_2$ when $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, that is to say respectively $P_{\boldsymbol{\theta}_2}(E_1)$ and $P_{\boldsymbol{\theta}_1}(E_2)$.

About Maximum Likelihood Estimation

Define now the function $C : E_1 \subset E \to C(E_1) = P_{\boldsymbol{\theta}_2}(E_1) + P_{\boldsymbol{\theta}_1}(E_2)$.

Noting that $P_{\boldsymbol{\theta}_1}(E_2) = 1 - P_{\boldsymbol{\theta}_1}(E_1)$ and remembering that $P_{\boldsymbol{\theta}_i}$ are dominated by ν with densities $p(\cdot;\boldsymbol{\theta}_i)$, we get that

$$C(E_1) = 1 + \int_{E_1} p(\mathbf{y};\boldsymbol{\theta}_2) - p(\mathbf{y};\boldsymbol{\theta}_1)\mathrm{d}\nu(\mathbf{y}).$$

From there, it becomes apparent that C can be minimized by taking

$$E_1 = \{\mathbf{y} \in E : p(\mathbf{y};\boldsymbol{\theta}_2) \leq p(\mathbf{y};\boldsymbol{\theta}_1)\},$$

or in other words, by estimating $\boldsymbol{\theta}$ to be $\boldsymbol{\theta}_1$ for $\mathbf{y}$ such that $p(\mathbf{y};\boldsymbol{\theta}_1) \geq p(\mathbf{y};\boldsymbol{\theta}_2)$ (note that what is done in the equality case does not change the value of C).

About Maximum Likelihood Estimation

MLE is used well beyond the previous (didactic) case where possible values for the parameter $\boldsymbol{\theta}$ were restricted to a set of two.

In a more general framework, another fact that supports MLE is the following. Assume that the Y_i's are i.i.d. with a distribution possessing a density $p_0(\cdot;\boldsymbol{\theta})$ with respect to some measure ν_o on $\mathcal{Y}$, where $\boldsymbol{\theta} \in \Theta$, and take $E = \mathcal{Y}^n$ equipped with the product measure $\nu = \nu_o^{\otimes n} = \nu_o \otimes \cdots \otimes \nu_o$.

Then the likelihood of $\boldsymbol{\eta}$ given $\mathbf{y}$ is $L(\boldsymbol{\eta};\mathbf{y}) = p(\mathbf{y};\boldsymbol{\eta}) = \prod_{i=1}^n p_0(y_i;\boldsymbol{\eta})$. Now we observe that the so-called *log-likelihood*

$$\ell(\boldsymbol{\eta};\mathbf{y}) = \log\left(\prod_{i=1}^n p_0(y_i;\boldsymbol{\eta})\right) = \sum_{i=1}^n \log(p_0(y_i;\boldsymbol{\theta}))$$

is —up to a normalization term $1/n$— a “sample analogue” of the function

$$\bar{\ell}_o : \boldsymbol{\eta} \in \Theta \to \bar{\ell}_o(\boldsymbol{\eta}) = \mathbb{E}[\log(p_0(Y;\boldsymbol{\eta}))],$$

where Y is assumed to follow $P_{\boldsymbol{\theta}}$.

About Maximum Likelihood Estimation

The crux is then that under regularity conditions $\bar{\ell}$ is actually maximal at the true parameter $\boldsymbol{\theta}$. Indeed, considering $\boldsymbol{\eta} \in \Theta$, we have formally

$$\begin{aligned}\bar{\ell}_o(\boldsymbol{\theta}) - \bar{\ell}_o(\boldsymbol{\eta}) &= \mathbb{E}[\log(p_0(Y;\boldsymbol{\theta}))] - \mathbb{E}[\log(p_0(Y;\boldsymbol{\eta}))] \\ &= \mathbb{E}\left[-\log\left(\frac{p_0(Y;\boldsymbol{\eta})}{p_0(Y;\boldsymbol{\theta})}\right)\right] \\ &\geq -\log\left(\mathbb{E}\left[\frac{p_0(Y;\boldsymbol{\eta})}{p_0(Y;\boldsymbol{\theta})}\right]\right) = -\log(1) = 0.\end{aligned}$$

This is an additional motivation for using MLE in the case of i.i.d. observations, especially when the number of replications is large.

About Maximum Likelihood Estimation

Major questions that arise regarding MLE notably include

- for a given model, the existence of some likelihood maximizer(s),
- and, when existing, how to compute the MLE in practice.

Moreover, questions such as the consistency and further asymptotic properties of MLE are of utmost importance both in theory and in practice.

When the likelihood function is sufficiently regular (twice continuously differentiable over $\Theta \subset \mathbb{R}^p$ is often assumed/met here) and the true parameter is assumed to lie on the interior of the domain, the MLE is searched among the roots of the so-called *score function*:

$$U : \boldsymbol{\theta} \in \Theta \to U(\boldsymbol{\theta}) = \nabla\ell(\boldsymbol{\theta};\mathbf{y}) = \left(\frac{\delta}{\delta\theta_j}\ell(\boldsymbol{\theta};\mathbf{y})\right)_{i=1,\ldots,p}.$$

Now, finding solutions to $U(\boldsymbol{\theta}) = \mathbf{0}$ (called “normal equation”) with a non-linear U is often non-trivial and call for iterative procedures.

About Maximum Likelihood Estimation

Another function that comes into play in maximum likelihood estimation is

$$\mathcal{J} : \boldsymbol{\theta} \in \Theta \to \mathcal{J}(\boldsymbol{\theta}) = -\nabla^2 \ell(\boldsymbol{\theta}; \boldsymbol{y}),$$

the so-called "observed information matrix".

$\mathcal{J}$ can notably be used to check if a solution to the normal equations actually is a local maximizer of ℓ. Also, $\mathcal{J}$ is instrumental in numerical MLE procedures (Cf. Newton-Raphson algorithms). The "expected version" of $\mathcal{J}$ further plays a key role in asymptotic results regarding MLE's variance.

About Maximum Likelihood Estimation

Note finally that the existence of solutions, the convergence of algorithms, asymptotic properties of MLE are model-dependent and require serious examination. Next, we will present a distribution class, Natural Exponential Families, for which MLE enjoys convenient properties and that will turn out to be relevant for logistic and Poisson regression.