Part II: Generalized Linear Models

Chapter II.2

Theory of Generalized Linear Models

Maximum Likelihood Estimation

Maximum Likelihood Estimation

Since GLMs assume a specific probability distribution for the responses that belongs to the exponential dispersion family (EDF), maximum likelihood estimation (MLE) procedures are used for parameter estimation, and a general formulation can be developed within the GLM set-up.

Log-Likelihood for GLMs' Parameters Vector $\boldsymbol{\beta}$

Consider n independent responses $Y_i \sim \mathrm{EDF}(\vartheta_i,\phi)$, with the associated function a possibly dependent on the observations, i.e. $a(\phi) = a(\phi; i)$, $i = 1, \ldots, n$. Then the corresponding log-likelihood function (s. Definition II.2.3) is:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i = \sum_{i=1}^{n} \log f(y_i; \vartheta_i,\phi) = \sum_{i=1}^{n} \frac{y_i\vartheta_i - b(\vartheta_i)}{a(\phi; i)} + \sum_{i=1}^{n} c(y_i,\phi) \; .$$

It is a function of $\boldsymbol{\beta}$ due to $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\mu} = E(\mathbf{Y}) = b'(\boldsymbol{\vartheta})$.

The first derivative of ℓ is the *score function* (s. Definition II.1.8):

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \ldots \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p} \right)' \; ,$$

where $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j}$, $j = 1, \ldots, p$. By the chain rule we further have

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \vartheta_i} \cdot \frac{\partial \vartheta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \; .$$

II.2.16 Likelihood equations for parameters vector β

Equating the components of the score function $S(\boldsymbol{\beta})$ to zero, the *likelihood equations* are obtained

$$\sum_{i=1}^{n} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot x_{ik} \right) = 0 \;, \quad k = 1, \ldots, p \;,$$

where $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = g(\mu_i)$, when the link function is g.

II.2.17 Remark

The likelihood equations above have the matrix form

$$\mathbf{X}'\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \;,$$

where $\mathbf{V} = \mathrm{diag}\,(\mathrm{Var}(Y_1), \ldots, \mathrm{Var}(Y_n))$, $\mathbf{D} = \mathrm{diag}\left(\frac{\partial \mu_1}{\partial \eta_1}, \ldots, \frac{\partial \mu_n}{\partial \eta_n}\right)$ and 0 is an n-dimensional column vector of 0s. Although $\boldsymbol{\beta}$ does not appear in the equation above, it is there implicitly through $\boldsymbol{\mu}$, since $\mu_i = g^{-1}\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$.

II.2.18 Remark (the mean-variance relation in EDF)

The likelihood equations of a GLM depend on the distribution of Y_i only through $E(Y_i)$ and $\mathrm{Var}(Y_i)$, $i = 1, \ldots, n$.

Furthermore,

- $\mathrm{Var}(Y_i)$ depends on $\mu_i = E(Y_i)$ through a functional form $\mathrm{Var}(Y_i) = v(\mu_i)$.
  For example,
  - $Y_i \sim \mathcal{P}(\mu_i)$: $\mathrm{Var}(Y_i) = \mu_i$ (identity function),
  - $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$: $\mathrm{Var}(Y_i) = \sigma^2$ (constant function),
  - $Y_i \sim \mathcal{B}(m_i, \pi_i)/m_i$ with $\mu_i = E(Y_i) = \pi_i$: $\mathrm{Var}(Y_i) = \frac{\mu_i(1-\mu_i)}{m_i}$.
- In EDF, this mean-variance relation *characterizes* the distribution.
  Example: if $\mathrm{Var}(Y_i) = E(Y_i)$, then Y_i has to be Poisson distributed.

II.2.19 Likelihood equations for GLMs with canonical link

Consider n independent responses $Y_i \sim \text{EDF}(\vartheta_i,\phi)$ with $a(\phi) = a(\phi;i)$ and $\text{E}(Y_i) = \mu_i$ $i = 1,\ldots,n$. For a GLM with *canonical link*, the likelihood equations II.2.16 are simplified to

$$\sum_{i=1}^{n} \frac{1}{a(\phi;i)} (y_i - \mu_i)\, x_{ik} = 0\ , \quad k = 1,\ldots,p\ .$$

Proof

Due to the canonical link, we have $\eta_i = g(\mu_i) = \vartheta_i$ and hence $\mu_i = g^{-1}(\vartheta_i) = g^{-1}(\eta_i)$, $i = 1,\ldots,n$. Thus,

$$\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = \frac{\partial g^{-1}(\vartheta_i)}{\partial \vartheta_i} = \frac{\partial \mu_i}{\partial \vartheta_i} = \frac{\partial b'(\vartheta_i)}{\partial \vartheta_i} = b''(\vartheta_i), \quad i = 1,\ldots,n.$$

Furthermore, it holds $\text{Var}(Y_i) = a(\phi;i)b''(\vartheta_i)$ (see Proposition II.2.4).

The result is derived by substituting $\text{Var}(Y_i)$ and $\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i}$ in the likelihood equations II.2.16 by the expressions above.

II.2.20 Remark (likelihood equations for Poisson GLMs)

The Poisson GLM is the 'standard' GLM for count data (s. Chapters II.5 and II.6). It assumes that the random component is Poisson distributed and adopts the associated canonical link, i.e. the log-link $\log(\mu_i)$ (s. Example II.2.15). In this case, considering n independent responses $Y_i \sim \mathcal{P}(\mu_i)$, $i = 1, \ldots, n$, the likelihood equations in II.2.19 become

$$\sum_{i=1}^{n} (y_i - \mu_i) x_{ik} = 0 \ , \quad k = 1, \ldots, p \ ,$$

or, in matrix notation,

$$\mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}) = 0_{n \times 1} \ ,$$

since $a(\phi; i) = \phi / w_i$ with $\phi = 1$, $w_i = 1$ (s. Remark II.2.8) and $E(Y_i) = \mu_i$, $i = 1, \ldots, n$.

Note that $\mathbf{X}'\mathbf{Y}$ is a *sufficient* statistic for β and the likelihood equations equate every β_k's sufficient statistic value $\sum_{i=1}^{n} x_{ik} y_i$ to its expected value.

☞ If all explanatory variables are categorical, then the model is the *Poisson log-linear model* (s. Chapter II.5) while in the presence of continuous explanatory variables, the model is the so called *Poisson regression* (s. Chapter II.6).

II.2.21 Remark (likelihood equations for binomial logit models)

If the responses to be modeled are proportions of success, then assuming a sample of n independent random proportions, i.e. $Y_i \sim \mathcal{B}(m_i,\pi_i)/m_i$, $i = 1,\dots,n$, the associate canonical link is the *logit* link $\log(\frac{\pi_i}{1-\pi_i})$ (s. Example II.2.15) and the likelihood equations in II.2.19 become

$$\sum_{i=1}^{n} m_i(y_i - \pi_i)x_{ik} = 0 \ , \quad k = 1,\dots,p \ ,$$

since $a(\phi;i) = \phi/w_i$ with $\phi = 1$, $w_i = m_i$ (s. Remark II.2.8) and $E(Y_i) = \mu_i = \pi_i$, $i = 1,\dots,n$.

II.2.22 Theorem (large sample Normal distribution of MLE $\hat{\beta}$)

Under standard *regularity conditions* [a], the maximum likelihood estimator (MLE) of β, $\hat{\beta}$, has an approximate normal distribution. Thus, for large n it holds

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta,\ \mathrm{Cov}(\hat{\beta})\right) .$$

The **asymptotic covariance matrix** of $\hat{\beta}$ is $\mathrm{Cov}(\hat{\beta}) = \mathcal{I}_F^{-1}$, where $\mathcal{I}_F$ is the *expected Fisher information matrix* [b]

$$\mathcal{I}_F = \mathrm{E}\left(\frac{\partial \ell}{\partial \beta}\frac{\partial \ell}{\partial \beta'}\right) = \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \beta \partial \beta'}\right) = \mathbf{X}'\mathbf{W}\mathbf{X} ,$$

with $\mathbf{W} = \mathrm{diag}\left(\frac{(\partial \mu_1/\partial \eta_1)^2}{\mathrm{Var}(Y_1)}, \ldots, \frac{(\partial \mu_n/\partial \eta_n)^2}{\mathrm{Var}(Y_n)}\right)$.

[a] They mainly require that β is in the interior of the parameter space and has fixed dimension as n increases (s. II.1.15 or Cox & Hinkley (2000, *Theoretical Statistics*, Chapman & Hall/CRC, Section 9.1).

[b] see Definition II.1.13

The asymptotic covariance matrix is estimated as

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \mathcal{I}_F^{-1}(\hat{\beta}) = \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1} ,$$

where $\widehat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\hat{\beta}$.

II.2.23 Definition (observed information matrix)

Let $\ell(\boldsymbol{\beta})$ be the log-likelihood of a GLM with associated parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$. The observed information matrix is the following $p \times p$ matrix

$$\mathcal{J}_F^{obs} = \left(-\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_r}\right) = -\mathcal{H} \, ,$$

where $\mathcal{H}$ is known as the *Hessian matrix*. It holds:

$$\mathcal{J}_F = \mathsf{E}\left(\mathcal{J}_F^{obs}\right) = \mathsf{E}\left(-\mathcal{H}\right).$$

this parameter vector corresponds to each single parameter

for example, gamma(a, b), there can be a vector that corresponds to alpha

II.2.24 Information matrix for GLMs with canonical link

For a GLM with canonical link function, since $\eta_i = \vartheta_i$, it follows that (s. Proof of II.2.19) $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial b'(\vartheta_i)}{\partial \vartheta_i} = b''(\vartheta_i)$, $i = 1, \dots, n$, leading to

$$\mathcal{H} = -\mathbf{X}'\mathbf{W}_c\mathbf{X},$$

where $\mathbf{W}_c = \mathrm{diag}(w_1, \dots, w_n)$ with $w_i = \frac{b''(\vartheta_i)}{a(\phi;i)}$, **independent** of y. Hence

$$\boxed{\mathcal{J}_F = \mathsf{E}\left(-\mathcal{H}\right) = -\mathcal{H} = \mathcal{J}_F^{obs}}$$

Existence of MLEs

II.2.25 Remark

A necessary **but** not sufficient condition for the existence and uniqueness of the MLEs is that the model matrix $\mathbf{X}$ is of full rank.

For many GLMs with full-rank $\mathbf{X}$ (including Poisson log-linear and binomial logit models), the Hessian matrix is negative definite and the log-likelihood strictly concave. In such cases, the MLEs of the model parameters exist uniquely under quite general conditions [a].

[a] Wedderburn (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, *Biometrika*, 63, 27-32.

Fitting GLMs

The likelihood equations of a GLM are usually nonlinear in $\boldsymbol{\beta}$ and do not lead to closed form expressions for the MLE $\hat{\boldsymbol{\beta}}$. The two methods usually applied for solving the likelihood equations iteratively are

1. the *Newton-Raphson*, and
2. the *Fisher scoring*.

II.2.26 Algorithm (Newton-Raphson)

The Newton-Raphson is an iterative algorithm having following steps, adjusted in our estimation problem of β:

1. Initial guess $(t = 0)$: Set $\beta^{(t)} = \beta^{(0)}$.
2. For stage $t + 1$, $t = 0, 1, 2, \ldots$:
   - If $\beta^{(t)}$ is the value assigned to $\hat{\beta}$ at the previous stage t of the iterative procedure, obtain the next guess $\beta^{(t+1)}$ through the updating equations of the Newton-Raphson algorithm

     $$\beta^{(t+1)} = \beta^{(t)} - \left(\mathcal{H}^{(t)}\right)^{-1} S(\beta^{(t)}) \,, \tag{II.4}$$

     where $S(\beta^{(t)})$ and $\mathcal{H}^{(t)}$ are the score function $S(\beta)$ and the Hessian matrix $\mathcal{H}$ evaluated at $\beta^{(t)}$. For matrix inversion to be possible, $\mathcal{H}^{(t)}$ has to be non-singular.
   - Check for convergence: The algorithm converges and stops, say after t_c iterations, when a termination criterion is met, leading to $\hat{\beta} = \beta^{(t_c)}$. A termination criterion checks whether $\beta^{(t)}$ and $\beta^{(t+1)}$ are sufficient close, e.g $\sum_{j=1}^{p} |\beta_j^{t+1} - \beta_j^{t}| \leqslant c$, for some prespecified small c.

II.2.27 Remark

The updating equation (II.4) is derived by determining the point $\hat{\beta}$ at which $\ell(\beta)$ is maximized, when $\ell(\beta)$ is approximated near $\beta^{(t)}$ by the terms up to the second order in the Taylor expansion.

II.2.28 Algorithm (Fisher Scoring)

The *Fisher scoring* is similar to the *Newton-Raphson* algorithm with the only difference being that it is based on the **expected information matrix** $\mathfrak{J}_F$, instead of the observed information matrix $\mathfrak{J}_F^{obs} = -\mathcal{H}$.

In particular, the updating equations for the Fisher scoring algorithm are

$$\beta^{(t+1)} = \beta^{(t)} + \left(\mathfrak{J}_F^{(t)}\right)^{-1} S(\beta^{(t)}) \ ,$$

where $\mathfrak{J}_F^{(t)}$ is $\mathfrak{J}_F$ evaluated at $\beta^{(t)}$.

II.2.29 Remark

For GLMs with *canonical link*, the Newton-Raphson and Fischer scoring algorithms coincide, since the observed and expected information matrices are equal (see Remark II.2.24).

II.2.30 Remark (initial values)

The data $\mathbf{y}$ can be used as the initial estimates of $\boldsymbol{\mu}$ (for both algorithms). This determines the first estimate of $\mathbf{W}$ and hence $\boldsymbol{\beta}$.

II.2.31 Corollary (covariance matrix of fitted values)

For large sample sizes, since $\hat{\boldsymbol{\eta}} = g(\hat{\boldsymbol{\mu}}) = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\text{Cov}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$ (see Theorem II.2.22), it holds

$$\text{Cov}(\hat{\boldsymbol{\eta}}) = \mathbf{X}\text{Cov}(\hat{\boldsymbol{\beta}})\mathbf{X}' = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}' \,.$$

Furthermore, since $\hat{\boldsymbol{\mu}} = g^{-1}(\hat{\boldsymbol{\eta}})$, the asymptotic covariance matrix $\text{Cov}(\hat{\boldsymbol{\mu}})$ can be obtained from $\text{Cov}(\hat{\boldsymbol{\eta}})$ by the *delta method*

$$\text{Cov}(\hat{\boldsymbol{\mu}}) = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)\text{Cov}(\hat{\boldsymbol{\eta}})\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)' = \mathbf{D}\left(\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{D} \,,$$

where $\mathbf{D} = \text{diag}\left(\frac{\partial \mu_1}{\partial \eta_1}, \ldots, \frac{\partial \mu_n}{\partial \eta_n}\right)$ (see Remark II.2.17).