Generalized Linear Models

Lecture 2. Exponential family of distributions
Score function. Fisher information

Background. Linear model

$$\boldsymbol{y} = \mathbf{X}\beta + \varepsilon, \quad \mathbf{E}\boldsymbol{y} = \boldsymbol{\mu} = \mathbf{X}\beta$$

$\boldsymbol{y} = (y_1, \ldots, y_n)^T$ – dependent variable, response
$\mathbf{X} = (\mathbf{1}, \boldsymbol{x}_1^c, \ldots, \boldsymbol{x}_k^c)$ – design matrix $n \times p$, $\mathbf{1}$ – vector of 1s
$\beta = (\beta_0, \ldots, \beta_k)^T$ – vector of unknown parameters
k – number of explanatory variables (number of unknown parameters $p = k + 1$)
$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ – vector of random errors

Assumptions:
$\mathbf{E}\varepsilon_i = 0,\ \varepsilon_i \sim N(0, \sigma^2),\ \mathbf{D}\varepsilon_i = \sigma^2.$
$\Rightarrow$ **Model has normal errors, constant variance and the mean of response equals to the linear combination of explanatory variables.**

There are some situations where general linear models are not appropriate

- The range of Y is restricted (e.g. binary, count)
- The variance of Y depends on the mean.

Generalized linear models extend the general linear model framework to address these issues.

Data structure

In cross-sectional analysis the data (y_i, x_i) consists of observations for each individual i $(i = 1, \ldots, n)$.
Ungrouped data – the common situation, one row of data matrix corresponds to exactly one individual.
Grouped data – only rows with different combinations of covariate values appear in data matrix, together with the number of repetitions and the arithmetic mean of the individual responses

Notation

- n – number of groups (or individuals in ungrouped case)
- n_i – number of individuals (repetitions) corresponding to row i
- N – total number of measurements (sample size):

$$N = \sum_{i=1}^{n} n_i$$

Ungrouped data is a special case of grouped data with $n_1 = \ldots = n_N = 1$ (and $n = N$).

Source: Fahrmeir, Tutz (2001); Tutz (2012)

Example. Grouped data

TABLE 4.3: Short- and long-term unemployment depending on age

Observ.		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Age		16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	1	2	11	31	42	50	54	43	35	25	27	21	21	19	22	17	14
	0	3	8	9	20	17	26	16	16	12	8	10	10	7	8	10	11

Observ.		17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Age		32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
	1	8	14	11	13	15	6	5	11	9	14	7	13	4	10	9	9
	0	3	7	8	9	7	7	6	7	4	5	6	8	2	6	4	6

Observ.		33	34	35	36	37	38	39	40	41	42	43	44	45	46
Age		48	49	50	51	52	53	54	55	56	57	58	59	60	61
	1	10	7	3	8	3	1	2	2	2	3	2	2	3	0
	0	3	4	5	5	2	6	4	3	1	6	4	7	5	1

Table taken from Tutz (2012)

Generalized linear model

Conditional distribution of response variable Y belongs to exponential family, $Y_i \sim \mathcal{E}$ with mean μ_i.

Generalized design matrix

- $\mathbf{X} = (x_{ij}), \quad i = 1, \ldots, n; \; j = 1, \ldots, p$
- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, where $\mathbf{x}_i$ is the covariate vector for row i
- NB! Columns of $\mathbf{X}$ can be transformed from original data using certain design functions

Generalized linear model

Conditional mean of response: $\mu_i = h(\mathbf{x}_i^T \beta), \quad g(\mu_i) = \mathbf{x}_i^T \beta$
g – link function, h – response function, $h = g^{-1}$

The generalized linear model expands the general linear model so that the dependent variable is linearly related to the explanatory variables via a specified link function.

Three components of GLM

There are three components to any GLM:

- Random component: identifies dependent variable Y, its (conditional) probability distribution belongs to exponential family, $Y_i \sim \mathcal{E}$ with mean μ_i.
- Systematic component: identifies the set of explanatory variables in the model, more specifically their linear combination in creating the so called linear predictor $\eta_i \doteq g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$.
- Link Function: identifies a function of the mean that is a linear function of the explanatory variables (specifies the link between random and systematic components $g(\mu_i) = \eta_i$)

Exponential family of distributions

Exponential Family (Pitman, Koopman, Darmois, 1935–1936)
Natural Exponential Family (Morris, 1982)
Simple Exponential Family (Tutz, 2012)
Exponential Dispersion Family (Bent Jorgensen, 1987) $\varphi \to a(\varphi)$

Definition. Natural Exponential family of distributions

Exponential family is a set of probability distributions whose probability density function (or probability mass function in case of a discrete distribution) can be expressed in the form

$$f(y_i;\theta_i,\varphi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\varphi_i} + c(y_i,\varphi_i)\right\},$$

- θ_i – natural or canonical parameter of the family
- φ_i – scale or dispersion parameter (known), often $\varphi_i = \varphi \cdot a_i$,
  for grouped data: $a_i = 1/n_i$, for ungrouped data: $a_i = 1, i = 1, \ldots, n$.
- $b(\cdot)$ – real-valued differentiable function of parameter θ_i
- $c(\cdot)$ – known function, independent of θ_i

Regular distribution

Distributions belonging to the exponential family of distributions are regular, i.e.

- Integration and differentiation can be exchanged.
- The support of the distribution ($\{x : f(x, \theta) > 0\}$) cannot depend on parameter θ.

For a regular distribution:

$$\mathbf{E}\frac{\partial l(\theta; Y)}{\partial \theta} = 0$$

$$\mathbf{E}\frac{\partial^2 l(\theta; Y)}{\partial \theta \partial \theta^T} = -\mathbf{E}(\frac{\partial l(\theta; Y)}{\partial \theta}\frac{\partial l(\theta; Y)}{\partial \theta^T})$$

Mean and variance in the exponential family, 1

Lemma

If the distribution of r.v. Y belongs to exponential family, $Y \sim \mathcal{E}$, it can be shown that:

- expected value of Y is equal to the first derivative of b with respect to θ, i.e. $\mathbf{E}Y = b'(\theta)$
- variance of Y is the product of the second derivative of $b(\theta)$ and the scale parameter φ, i.e. $\mathbf{D}Y = \varphi b''(\theta)$

Note that both mean and variance are determined by the function $b(\cdot)$

The second derivative of function b is called the variance function $\nu(\theta) := b''(\theta)$

The variance function ν describes how the variance depends on the mean.

Mean and variance in the exponential family, 2

In a GLM, we assume that the conditional distribution of responses $Y_i = Y|\mathbf{x}_i$ is in the exponential family, i.e.

$$f(y_i; \theta_i, \varphi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\varphi_i} + c(y_i, \varphi_i)\right\}$$

Then, by previous lemma:

- $\mu_i = \mathbf{E}(Y_i) = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta}$
- $\sigma_i^2 = \mathbf{D}(Y_i) = \varphi_i \nu(\theta_i) = \varphi_i b''(\theta_i) = \varphi_i \frac{\partial^2 b(\theta_i)}{\partial \theta^2}$

Since θ_i depends on μ_i, it can be written as $\theta_i = \theta(\mu_i)$ and thus ν is usually considered as a function of μ_i

Link function

Definition. Link function

Link function $g(\cdot)$ is a function that relates the mean value of response to the linear predictor (linear combination of predictors).

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \beta$$

Link function must be monotone and differentiable.

For a monotone function g we can define the inverse called the **response function** ($h = g^{-1}$).

The choice of the link function depends on the type of data.

Certain link functions are natural for certain distributions and are called canonical (natural) links.
For each member of the exponential family, there exists a natural or canonical link function.
The canonical link function relates the **canonical parameter** θ_i directly to the linear predictor $\theta_i = g(\mu_i) = \eta_i = \mathbf{x}_i^T \beta$
Canonical links yield 'nice' mathematical properties in the estimation process.

Setup

$$\boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i \quad \underset{g}{\overset{h=g^{-1}}{\rightleftarrows}} \quad \mu_i \quad \underset{\mu}{\overset{\theta}{\rightleftarrows}} \quad \theta_i$$

- Functions μ and θ are specified by a particular distribution
- We want to predict μ_i using $\boldsymbol{x}_i \Rightarrow$ we need to find best $\boldsymbol{\beta}$ for our model
- MLE goes through θ_i, the natural parameter of exponential family
- With canonical link $\eta_i = \theta_i$ and the calculations are simpler (yet this is not the only criterion for choosing link function g)

Canonical (natural) links

Normal distribution $N(\mu, \sigma^2)$

Canonical parameter: $\theta = \mu \Rightarrow$ canonical link: $g(\mu) = \mu$, link **identity**

Bernoulli distribution $B(1, \pi)$, $\mu = \pi$

Canonical parameter: $\theta = \ln \frac{\pi}{1-\pi}$, canonical link: $g(\mu) = \ln \frac{\mu}{1-\mu}$, **Logit**-link

Poisson Distribution $Po(\lambda)$, $\mu = \lambda$

Canonical parameter: $\theta = \ln \lambda$, canonical link: $g(\mu) = \ln \mu$, **Log**-link

Gamma distribution

Canonical parameter: $\theta = (-)\mu^{-1}$, canonical link: $g(\mu) = (-)\mu^{-1}$, **inverse**-link

Inverse Gaussian *IG*

Canonical parameter: $\theta = (-)\frac{1}{2\mu^2}$, canonical link: $g(\mu) = (-)\frac{1}{2\mu^2}$, **squared inverse** link

Example: Normal distribution $N(\mu_i, \sigma^2)$

Probability density function

$$f(y_i; \mu_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right),$$

$$-\frac{(y_i - \mu_i)^2}{2\sigma^2} = -\frac{y_i^2}{2\sigma^2} + \frac{2y_i\mu_i - \mu_i^2}{2\sigma^2}$$

$\Rightarrow$ Normal distribution belongs to the exponential family:

- canonical (natural) parameter $\theta_i = \mu_i$ (identity link is canonical)
- $b(\theta_i) = \frac{1}{2}\mu_i^2 \ (= \frac{1}{2}\theta_i^2)$
- mean: $b'(\theta_i) = \mu_i \ (= \theta_i)$
- variance: $\varphi_i \cdot b''(\theta_i)$, $b''(\theta_i) = 1$, $\varphi_i = \sigma^2$

Example: Poisson distribution $Po(\mu_i)$ $(Po(\lambda_i))$

Probability mass function

$$p(y_i; \mu_i) = exp(-\mu_i)\frac{\mu_i^{y_i}}{y_i!} = \exp(-\mu_i + y_i \ln \mu_i - \ln(y_i!))$$

Let $\theta_i = \ln \mu_i$, then

$$p(y_i; \theta_i) = \exp(y_i\theta_i - \exp(\theta_i) - c(y_i))$$

$\Rightarrow$ Poisson distribution belongs to the exponential family:

- canonical (natural) parameter $\theta_i = \ln \mu_i$ (log-link is canonical)
- $b(\theta_i) = \exp(\theta_i)$
- mean: $b'(\theta_i) = \exp(\theta_i) = \mu_i$
- variance: $\varphi_i \cdot b''(\theta_i)$, $b''(\theta_i) = \mu_i$, $\varphi_i = 1$

Example: Bernoulli distribution $B(1, \mu_i)$ $(B(1, \pi_i))$

Probability mass function

$$p(y_i; \mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \left(\frac{\mu_i}{1-\mu_i}\right)^{y_i}(1-\mu_i)$$

$$p(y_i; \mu_i) = \exp\left(y_i \ln(\frac{\mu_i}{1-\mu_i}) + \ln(1-\mu_i)\right)$$

$\Rightarrow$ Bernoulli distribution belongs to the exponential family:

- canonical (natural) parameter is log-odds, which yields logistic canonical link:

$$\theta_i = \ln(\frac{\mu_i}{1-\mu_i})$$

- $b(\theta_i) = \ln(1 + e^{\theta_i})$
- dispersion (scale) parameter $\varphi_i = 1$

Likelihood and log-likelihood

Estimation of the parameters of a GLM is done using the maximum likelihood method.
Parameter values that maximize the log-likelihood are chosen as the estimates for parameters.

Likelihood function

$$L(\theta_i; y_i) = f(y_i; \theta_i, \varphi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\varphi_i} + c(y_i, \varphi_i)\right\}$$

Log-likelihood function

$$l_i(\theta_i) \doteq \ln L(\theta_i; y_i) = \frac{y_i\theta_i - b(\theta_i)}{\varphi_i} + c(y_i, \varphi_i)$$

Likelihood of a sample

Let us have a sample $\mathbf{y}$ consisting of n independent units and let $\boldsymbol{\theta}$ be the vector of corresponding natural parameters. Then we have

Likelihood of a sample

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{\varphi_i} + c(y_i, \varphi_i) \right\}$$

Log-likelihood of a sample

$$l(\boldsymbol{\theta}; \mathbf{y}) \doteq \ln L(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} l_i(\theta_i) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\varphi_i} + c(y_i, \varphi_i) \right\}$$

Maximum Likelihood estimation

Note than we are actually interested in estimating the parameters β not $\boldsymbol{\theta}$

Let us also recall that μ_i and θ_i depend on each other, i.e.

- $\mu_i = \mu(\theta_i)$
- $\theta_i = \theta(\mu_i)$

Further, since our model is $g(\mu(\theta_i)) = \mathbf{x}_i^T \beta = \eta_i$ or, equivalently, $\mu(\theta_i) = h(\mathbf{x}_i^T \beta) = h(\eta_i)$, we also have

$$\theta_i = \theta(\mu_i) = \theta(h(\mathbf{x}_i^T \beta))$$

$\Rightarrow$ we can express θ_i through the unknown parameter β

This allows to write us the likelihood w.r.t. β as follows:

$$l(\beta; \mathbf{y}) = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} \frac{y_i \theta(h(\mathbf{x}_i^T \beta)) - b(\theta(h(\mathbf{x}_i^T \beta)))}{\varphi_i} + \dots$$

Score function

Definition. Score function

Score function $s(\cdot)$ is the derivative of log-likelihood function with respect to the parameters

Score function for parameters β:

$$s(\beta) = \sum_{i=1}^{n} s_i(\beta), \quad s_i(\beta) = \frac{\partial l_i(\beta)}{\partial \beta} = \left(\frac{\partial l_i(\beta)}{\partial \beta_1}, \ldots, \frac{\partial l_i(\beta)}{\partial \beta_p} \right)^T$$

Calculation of the score function, 1

Recall that $\theta_i = \theta(\mu_i) = \theta(h(\eta_i)) = \theta(h(\mathbf{x}_i^T \beta))$

Thus we can calculate the score function as follows:

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial l_i(\theta_i)}{\partial \theta} \cdot \frac{\partial \theta(\mu_i)}{\partial \mu} \cdot \frac{\partial h(\eta_i)}{\partial \eta} \cdot \frac{\partial \eta_i}{\partial \beta},$$

where

1. $\frac{\partial l_i(\theta_i)}{\partial \theta} = \frac{y_i - b'(\theta_i)}{\varphi_i} = \frac{y_i - \mu_i}{\varphi_i}$
2. $\frac{\partial \theta(\mu_i)}{\partial \mu} = \left(\frac{\partial \mu(\theta_i)}{\partial \theta} \right)^{-1} = \left(\frac{\partial^2 b(\theta_i)}{\partial \theta^2} \right)^{-1} = \frac{1}{\nu(\theta_i)} = \frac{\varphi_i}{\sigma_i^2}$
3. $\frac{\partial \eta_i}{\partial \beta} = \mathbf{x}_i$

Calculation of the score function, 2

Now, putting the obtained derivatives together, we get

$$s(\beta) = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial h(\eta_i)}{\partial \eta} \frac{(y_i - \mu_i)}{\sigma_i^2}$$

Calculation of the score function, 2

Now, putting the obtained derivatives together, we get

$$s(\beta) = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial h(\eta_i)}{\partial \eta} \frac{(y_i - \mu_i)}{\sigma_i^2}$$

Question

What happens if the link function is canonical, i.e. $\theta_i = \eta_i$?

Calculation of the score function, 2

Now, putting the obtained derivatives together, we get

$$s(\beta) = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial h(\eta_i)}{\partial \eta} \frac{(y_i - \mu_i)}{\sigma_i^2}$$

Question

What happens if the link function is canonical, i.e. $\theta_i = \eta_i$?

$$s(\beta) = \sum_{i=1}^{n} \frac{\partial l_i(\theta_i)}{\partial \theta} \frac{\partial \eta_i}{\partial \beta} = \sum_{i=1}^{n} \mathbf{x}_i \frac{(y_i - \mu_i)}{\varphi_i}$$

Calculation of the score function, 3

Let us denote (and recall):

- $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ – observed sample
- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ – vector of means
- $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$ – the covariance matrix
- $\mathbf{D} = \mathrm{Diag}(\frac{\partial h(\eta_1)}{\partial \eta}, \ldots, \frac{\partial h(\eta_n)}{\partial \eta})$ – diagonal matrix of derivatives
- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ – the design matrix
- $\mathbf{W} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^T$ – the weight matrix

These notations allow us to write the score function in matrix notation:

$$s(\beta) = \mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{X}^T\mathbf{W}\mathbf{D}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

Empirical (observed) Fisher information matrix $\tilde{\mathbf{F}}$ $(\mathcal{I})$

As our aim is to estimate parameters β, the obvious way is to solve the system $s(\beta) = 0$.

Since the equations are non-linear, we need additional tools before proceeding.

Definition. Empirical Fisher information matrix

Empirical Fisher information matrix is the negative of the second derivative of the log-likelihood function.

$$\tilde{\mathbf{F}}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T} = \left(-\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} \right)_{j,k=1,\dots p}$$

Recall that in mathematics, the second order partial derivative matrix is called the Hessian matrix $\mathbf{H}$, so $\tilde{\mathbf{F}} = -\mathbf{H}$

NB! As $\tilde{\mathbf{F}}(\beta)$ depends on the observations, it is therefore random.

Theoretical Fisher information matrix $\mathbf{F}$ $(\mathcal{J})$, 1

Definition. Theoretical (expected) Fisher information matrix

The theoretical (expected) Fisher information matrix is the mean of the observed Fisher information matrix

$$\mathbf{F}(\beta) = \mathbf{E}(\tilde{\mathbf{F}}(\beta))$$

Because of the regularity ($\mathbf{E}(s_i(\beta)) = 0$ and $\mathbf{E}(-\frac{\partial^2 l_i(\beta)}{\partial\beta\partial\beta^T}) = \mathbf{E}(\frac{\partial l_i(\beta)}{\partial\beta}\frac{\partial l_i(\beta)}{\partial\beta^T})$), the expected Fisher information matrix is also the variance-covariance matrix of the score vector:

$$\mathbf{F}(\beta) = \mathbf{\Sigma}_{s(\beta)} = \mathbf{E}(s(\beta)s(\beta)^T)$$

Now, applying the form of $s(\beta)$, we can derive

$$\mathbf{F}(\beta) = \mathbf{E}\left(\sum_{i=1}^{n} s_i(\beta)s_i(\beta)^T\right) = \mathbf{E}\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \left(\frac{\partial h(\eta_i)}{\partial\eta}\right)^2 \frac{(y_i-\mu_i)^2}{\sigma_i^4}\right)$$
$$= \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \left(\frac{\partial h(\eta_i)}{\partial\eta}\right)^2 \frac{1}{\sigma_i^2}$$

Theoretical (expected) Fisher information matrix $\mathbf{F}$, 2

Using the matrix notation, we can also write

$$\begin{aligned}\mathbf{F}(\beta) &= \mathbf{E}(s(\beta)s(\beta)^T) \\ &= \mathbf{E}(\mathbf{X}^T\mathbf{D}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{D}^T\mathbf{X}) = \mathbf{X}^T\mathbf{W}\mathbf{X},\end{aligned}$$

where the weight matrix W is

$$\mathbf{W} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^T = \text{Diag}\left(\left(\frac{\partial h(\eta_1)}{\partial \eta}\right)^2/\sigma_1^2, \ldots, \left(\frac{\partial h(\eta_n)}{\partial \eta}\right)^2/\sigma_n^2\right)$$

Theoretical (expected) Fisher information matrix **F**, 3

Question

What happens if the link function is canonical, i.e. $\theta_i = \eta_i$?

Theoretical (expected) Fisher information matrix **F**, 3

Question

What happens if the link function is canonical, i.e. $\theta_i = \eta_i$?

$$\mathbf{W} = \mathrm{Diag}\left(\frac{\sigma_1^2}{\varphi_1^2}, \ldots, \frac{\sigma_n^2}{\varphi_n^2}\right)$$

$$\mathbf{F}(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \frac{\sigma_i^2}{\varphi_i^2}$$

NB! Think what happens in case of normal distribution!

MLE equations

Idea

Equate the score function to zero: $s(\beta) = 0$ and solve simultaneously for parameters β to get ML estimates $\hat{\beta}$

Since β is a p-dimensional vector, we get p equations

$$s(\beta_1) = 0,\ s(\beta_2) = 0, \ldots, s(\beta_p) = 0$$

As the equations are non-linear, iterative methods are needed. General setup:

1. Choose the starting value $\hat{\beta}^{(0)}$
2. Adjust the value in each step r by $\Delta\hat{\beta}^{(r)}$: $\quad \hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + \Delta\hat{\beta}^{(r)}$
3. If the process converges, $\Delta\hat{\beta}^{(r)} \to 0$, proceed until $\Delta\hat{\beta}^{(r)}$ is 'small enough' and choose $\hat{\beta}^{(r)}$ as the ML estimate of parameter

The commonly used procedures are:

- Newton-Raphson method
- Fisher's method of scoring

Newton-Raphson method

The Newton-Raphson method is an iterative method for solving non-linear equations.

1. Start with initial guess $\hat{\beta}^{(0)}$
2. Consider the linear Taylor approximation $s_{\text{lin}}(\beta) \approx s(\beta)$ at $\hat{\beta}^{(r)}$, where $\hat{\beta}^{(r)}$ is the estimate in the rth step:

$$s(\beta) \approx s_{\text{lin}}(\beta) = s(\hat{\beta}^{(r)}) + \frac{\partial s(\hat{\beta}^{(r)})}{\partial \beta}(\beta - \hat{\beta}^{(r)})$$

   Now, $s_{\text{lin}}(\beta) = 0$ gives us the next estimate $\hat{\beta}^{(r+1)}$

$$\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + \tilde{\mathbf{F}}^{-1}(\hat{\beta}^{(r)})s(\hat{\beta}^{(r)}),$$

   since $\frac{\partial s(\hat{\beta}^{(r)})}{\partial \beta} = \mathbf{H}(\hat{\beta}^{(r)}) = -\tilde{\mathbf{F}}(\hat{\beta}^{(r)})$

3. The iterations are stopped when

$$\frac{\|\hat{\beta}^{(r+1)} - \hat{\beta}^{(r)}\|}{\|\hat{\beta}^{(r)}\|} \leq \varepsilon, \quad \varepsilon > 0$$

Fisher's method of scoring

An alternative method is the Newton method with Fisher scoring.
The essential difference is that the observed information matrix is replaced by expected.

$$\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)} + \mathbf{F}^{-1}(\hat{\beta}^{(r)})s(\hat{\beta}^{(r)})$$

Advantages:

- To calculate the expected information matrix we do not need to evaluate the second order derivatives.
- The expected information matrix is usually positive definite, so some non-convergence problems do not occur.

Properties of ML estimator

Asymptotic properties

- Consistency: ML estimator is consistent
- Distribution: under certain regularity conditions, MLE has asymptotically normal distribution

$$\hat{\beta} \overset{a}{\sim} N(\beta, \mathbf{F}^{-1}(\hat{\beta})),$$

where

$$\mathbf{F}(\hat{\beta}) = \mathbf{X}^T \hat{\mathbf{W}} \mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \left(\frac{\partial h(\hat{\eta}_i)}{\partial \eta} \right)^2 \frac{1}{\hat{\sigma}_i^2}$$

with $\hat{\mathbf{W}}$ being the matrix $\mathbf{W}$ evaluated at $\hat{\beta}$ and $\hat{\sigma}_i^2 = \hat{\mathbf{D}}(Y|\mathbf{x}_i) = \hat{\varphi}_i \nu(\hat{\mu}_i)$

- Efficiency: ML estimator is consistent and has the smallest asymptotic variance

The ML estimator is asymptotically normally distributed, asymptotically unbiased and efficient.

Estimate of the dispersion (scale) parameter φ

If the dispersion parameter φ is not known, it also needs to be estimated

Recall that $\sigma_i^2 = \varphi_i \nu(\mu_i) = \varphi \frac{\nu(\mu_i)}{n_i}$ and $\varphi = \frac{\sigma_i^2}{\nu(\mu_i)/n_i}$,
where n_i is the group size for grouped data and $n_i = 1$ for ungrouped data

Now the moment based (Pearson) estimate for φ is

$$\hat{\varphi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)/n_i}$$

Here p is the number of parameters and the model thus has $n - p$ degrees of freedom

NB! For a classical linear model we get the well known result that the variance is estimated by the sum of squares of the residuals.