last change: 22.nov.2011 @18.00

2 Maximum likelihood

Summary 1 *Asymptotic normality of mle, Fisher information, Wald test, score function, score test, LR-test, Newton-method, Hessian, observed information, Fisher scoring, Poisson regression, IRLS*

2.1 Asymptotic maximum likelihood theory

sample $Y_1, \ldots, Y_n$ iid $f(y|\ \theta) \implies$ log likelihood function:

$$l\left(\theta \mid y_1, \ldots, y_n\right) = l(\theta|\ \mathbf{y}) = \sum_i \log f(y_i|\ \theta) = \sum_i l(\theta|\ y_i)$$

score function ($p-$vector):

$$\frac{\partial}{\partial \theta} l(\theta|\ \mathbf{y}) = \left[\frac{\partial}{\partial \theta_j} l(\theta|\ \mathbf{y})\right] = \sum_i^n \frac{\partial}{\partial \theta} l(\theta|\ y_i)$$

mle ($p-$vector)

$$\frac{\partial}{\partial \theta} l(\widehat{\theta}|\ \mathbf{y}) = \left[\frac{\partial}{\partial \theta_j} l(\widehat{\theta}|\ \mathbf{y})\right] = \mathbf{0}$$

Fisher information ($p \times p-$matrix) of n observations ($\mathbb{I}_n$) resp of 1 observation ($\mathbb{I}_1$).

$$\mathbb{I}_n(\theta) = -\mathbb{E}\frac{\partial^2}{\partial \theta^2} l(\theta|\ \mathbf{y}) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} l(\theta|\ \mathbf{y})\right] = -\left[\sum_i^n \mathbb{E}\frac{\partial^2}{\partial \theta_j \partial \theta_k} l(\theta|\ y)\right]$$

Notes:

1. In the iid case $\mathbb{I}_n = n\mathbb{I}_1$
2. since $\mathbb{I}_n(\theta)$ is a positive definite symmetric matrix it can be factored into the product of two matrices $\left(\sqrt{\mathbb{I}_n(\theta)}\right)' \sqrt{\mathbb{I}_n(\theta)} = \mathbb{I}_n(\theta)$.

Theorem: under certain regularity conditions (existence of derivatives of log likelihoods etc, true value parameter not boundary value etc)

$$\sqrt{\mathbb{I}_n}\left(\widehat{\theta} - \theta\right) \to \mathcal{N}_p(0, \mathbf{I}) \qquad \text{in the independent, general case}$$

$(\sqrt{n}\left(\widehat{\theta}-\theta\right) \to \mathcal{N}_p\left(0, \mathbb{I}_1^{-1}\right)$ in the iid case)

Consequence:
$\mathbb{I}_n^{-1}(\theta)$ is the (asymptotic) variance-covariance matrix of $\widehat{\theta}$. ($\mathbb{I}_1^{-1}(\theta)$ is the variance-covariance matrix of $\sqrt{n}(\widehat{\theta}-\theta)$ for the iid case)

2.2 Example: Poisson regression

$$\mathbb{P}(Y=y\ |x) = \frac{\mu^y}{y!}\exp(-\mu)$$
$$\mu = \exp(\alpha+\beta x)$$

likelihood[1]:

$$L(\alpha,\beta|\ \mathbf{y}) = \prod_i \mathbb{P}\left(Y=y_i\ |\alpha,\beta,x_i\right)$$

log likelihood:

$$l(\alpha,\beta|\ \mathbf{y}) = \sum_i \left((\alpha+\beta x_i)y_i - \exp(\alpha+\beta x_i)\right)$$

score function:

$$\frac{\partial}{\partial\alpha} l\left(\alpha,\beta|\ \mathbf{y}\right) = \sum_i \left(y_i - \exp(\alpha+\beta x_i)\right)$$
$$\frac{\partial}{\partial\beta} l\left(\alpha,\beta|\ \mathbf{y}\right) = \sum_i x_i\ \left(y_i - \exp(\alpha+\beta x_i)\right)$$

mlest

$$\frac{\partial}{\partial\alpha} l\left(\widehat{\alpha},\widehat{\beta}\Big|\ \mathbf{y}\right) = 0$$
$$\frac{\partial}{\partial\beta} l\left(\widehat{\alpha},\widehat{\beta}\Big|\ \mathbf{y}\right) = 0$$

Second derivative matrix (**Hessian**) $\mathbf{H}(\alpha,\beta|\ \mathbf{y})$

$$\frac{\partial^2}{\partial\alpha^2} l\left(\alpha,\beta|\ \mathbf{y}\right) = -\sum_i \exp(\alpha+\beta x_i)$$
$$\frac{\partial^2}{\partial\beta^2} l\left(\alpha,\beta|\ \mathbf{y}\right) = -\sum_i x_i^2 \exp(\alpha+\beta x_i)$$
$$\frac{\partial^2}{\partial\alpha\partial\beta} l\left(\alpha,\beta|\ \mathbf{y}\right) = -\sum_i x_i \exp(\alpha+\beta x_i)$$

[1] suppress dependence on the independent variable, write $L(\alpha,\beta|\ \mathbf{y})$ instead of $L(\alpha,\beta|\ \mathbf{x},\mathbf{y})$

$\Longrightarrow$ Asymptotic normality $\sqrt{\mathbb{I}_n(\alpha,\beta)}\left(\begin{bmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{bmatrix}-\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\right)\to\mathcal{N}(0,\mathbf{I})$ with[2]

Fisher information

$$\mathbb{I}_n(\alpha,\beta)=-\mathbb{E}\mathbf{H}(\alpha,\beta|\ \mathbf{y})$$

Therefore, for large n

$$\begin{bmatrix}\mathrm{var}(\widehat{\alpha}) & \mathrm{cov}(\widehat{\alpha},\widehat{\beta})\\ \mathrm{cov}(\widehat{\alpha},\widehat{\beta}) & \mathrm{var}(\widehat{\beta})\end{bmatrix}\approx\mathbb{I}_n^{-1}(\alpha,\beta)$$

If the expectation needed for the Fisher information is hard to find, replace the expectation of the Hessian by the Hessian: define **observed** Fisher information: :

$$\widehat{\mathbb{I}}_n(\alpha,\beta)=-\begin{bmatrix}\sum_i^n\frac{\partial^2}{\partial\alpha^2}l(\alpha,\beta\mid y_i,x_i) & \sum_i^n\frac{\partial^2}{\partial\alpha\partial\beta}l(\alpha,\beta|\ y_i,x_i)\\ \sum_i^n\frac{\partial^2}{\partial\alpha\partial\beta}l(\alpha,\beta\mid y_i,x_i) & \sum_i^n\frac{\partial^2}{\partial\beta^2}l(\alpha,\beta\mid y_i,x_i)\end{bmatrix}$$

Since the values of α,β are unknown, replace for practical purposes$\widehat{\mathbb{I}}_n(\alpha,\beta)$ by $\widehat{\mathbb{I}}_n(\widehat{\alpha},\widehat{\beta})$

$$\Longrightarrow\ \begin{bmatrix}\mathrm{var}(\widehat{\alpha}) & \mathrm{cov}(\widehat{\alpha},\widehat{\beta})\\ \mathrm{cov}(\widehat{\alpha},\widehat{\beta}) & \mathrm{var}\ (\widehat{\beta})\end{bmatrix}\approx\begin{bmatrix}\sum_i\exp(\widehat{\alpha}+\widehat{\beta}x_i) & \sum_i x_i\exp(\widehat{\alpha}+\widehat{\beta}x_i)\\ \sum_i x_i\exp(\widehat{\alpha}+\widehat{\beta}x_i) & \sum_i x_i^2\exp(\widehat{\alpha}+\widehat{\beta}x_i)\end{bmatrix}^{-1}$$

2.3 ML-tests

$X_1,\ldots,X_n$ iid $f(x\mid\theta)$. Hypothesis

$$\mathrm{H}_0:\ \theta\in\mathbf{\Theta}_0$$

where $\mathbf{\Theta}_0\subset\mathbf{\Theta}$

2.3.1 Likelihood ratio test

compute $\widehat{\theta}=\arg\max_{\theta\in\mathbf{\Theta}}l(\theta|\ \mathbf{y})$. compute $\widehat{\theta}_0=\arg\max_{\theta\in\mathbf{\Theta}_0}l(\theta|\ \mathbf{y})$.

It can be shown that

$$2\sum\left(l(\widehat{\theta}|\ y_i)-l(\widehat{\theta}_0|\ y_i)\right)\to\chi^2(df_1-df_0)$$

where df =number of free parameters $(\dim\mathbf{\Theta},\dim\mathbf{\Theta}_0)$

[2] $\sqrt{\mathbb{I}_n(\alpha,\beta)}$ is a matrix such that $\left(\sqrt{\mathbb{I}_n(\alpha,\beta)}\right)'\sqrt{\mathbb{I}_n(\alpha,\beta)}=\mathbb{I}_n(\alpha,\beta)$

2.4 Netwon method and the MLE connection

sample $Y_1, \ldots, Y_n$ iid $f(y|\ \theta) \implies$ log likelihood function

$$l\left(\theta \mid y_1, \ldots, y_n\right) = l\left(\theta \mid \mathbf{y}\right) = \sum_i \log f(y_i|\ \theta) = \sum_i l(\theta|\ y_i)$$

score function ($p-$vector)

$$\frac{\partial}{\partial \theta} l\left(\theta \mid \mathbf{y}\right) = \left[\frac{\partial}{\partial \theta_j} l\left(\theta \mid \mathbf{y}\right)\right] = \sum_i^n \frac{\partial}{\partial \theta} l\left(\theta \mid y_i\right)$$

mle ($p-$vector)

$$\frac{\partial}{\partial \theta} l\left(\widehat{\theta} \,\middle|\, \mathbf{y}\right) = \left[\frac{\partial}{\partial \theta_j} l\left(\widehat{\theta} \,\middle|\, \mathbf{y}\right)\right] = \mathbf{0}$$

Fisher information ($p \times p-$matrix) of n observations ($\mathbb{I}_n$) resp of 1 observation ($\mathbb{I}_1$)

$$\begin{aligned} \mathbb{I}_n &= -\mathbb{E}\frac{\partial^2}{\partial \theta^2} l\left(\theta \mid \mathbf{y}\right) = -\left[\sum_i^n \mathbb{E}\frac{\partial^2}{\partial \theta_j \partial \theta_k} l\left(\theta \mid y_i, x_i\right)\right] \\ &= n\mathbb{I}_1 \quad \text{under iid assumption} \end{aligned}$$

use Newton method to solve

$$\frac{\partial}{\partial \theta} l\left(\widehat{\theta} \,\middle|\, \mathbf{y}\right) = \left[\frac{\partial}{\partial \theta_j} l\left(\widehat{\theta} \,\middle|\, \mathbf{y}\right)\right] = \mathbf{0}$$

need to compute second derivative matrix (**Hessian**)

$$\mathbf{H}\left(\theta\right) = \left[\frac{\partial l^2}{\partial \theta_i \partial \theta_j} l\left(\widehat{\theta} \,\middle|\, \mathbf{y}\right)\right]$$

$\implies$ **Newton method:**

$$\theta^{(t+1)} = \theta^{(t)} - \mathbf{H}^{-1}(\theta^{(t)}) \frac{\partial l(\theta^{(t)}|\ \mathbf{y})}{\partial \theta}$$

Define Fisher Information

$$\mathbb{I}_n(\theta \mid \mathbf{y}) = -\mathbb{E}\mathbf{H}(\theta) = -\left[\mathbb{E}\sum_i^n \frac{\partial^2 l(\theta \mid y_i)}{\partial \theta_i \partial \theta_j}\right]$$

the Hessian can be considered the negative **observed** Fisher information

$$\widehat{\mathbb{I}}_n(\theta \mid \mathbf{y}) = -\mathbf{H}(\theta)$$

$\Longrightarrow$ **Newton method**:

$$\theta^{(t+1)} = \theta^{(t)} + \widehat{\mathbb{I}}_n^{-1}(\theta^{(t)} \mid \mathbf{y}) \frac{\partial l(\theta^{(t)} \mid \mathbf{y})}{\partial \theta}$$

Using $\mathbb{I}(\theta \mid \mathbf{y})$ instead of $H(\theta)$ the Newton method is called the method of "**Fisher scoring**":

$$\theta^{(t+1)} = \theta^{(t)} + \mathbb{I}_n^{-1}(\theta^{(t)} \mid \mathbf{y}) \frac{\partial l(\theta^{(t)} \mid \mathbf{y})}{\partial \theta}$$

2.4.1 Example: Poisson regression

$$\begin{aligned} \mathbb{P}((Y = y \mid x) &= \frac{\mu^y}{y!} \exp(-\mu) \\ \mu &= \exp(\alpha + \beta x) \end{aligned}$$

log likelihood: sample $(x_1, y_1), \ldots, (x_n, y_n)$:

$$l(\alpha, \beta \mid \mathbf{y}, \mathbf{x}) = \sum_i ((\alpha + \beta x_i) y_i - \exp(\alpha + \beta x_i))$$

score function:

$$\begin{aligned} \frac{\partial}{\partial \alpha} l(\alpha, \beta \mid \mathbf{y}, \mathbf{x}) &= \sum_i (y_i - \exp(\alpha + \beta x_i)) \\ \frac{\partial}{\partial \beta} l(\alpha, \beta \mid \mathbf{y}, \mathbf{x}) &= \sum_i x_i \ (y_i - \exp(\alpha + \beta x_i)) \end{aligned}$$

mlest:

$$\begin{aligned} \frac{\partial l}{\partial \alpha}(\widehat{\alpha}, \widehat{\beta}) &= 0 \\ \frac{\partial l}{\partial \beta}(\widehat{\alpha}, \widehat{\beta}) &= 0 \end{aligned}$$

Second derivative matrix (**Hessian**)

$$H(\alpha,\beta) = \begin{bmatrix} \frac{\partial^2 l(\alpha,\beta \mid \mathbf{y},\mathbf{x})}{\partial \alpha^2} & \frac{\partial^2 l(\alpha,\beta \mid \mathbf{y},\mathbf{x})}{\partial \alpha \partial \beta} \\ \frac{\partial^2 l(\alpha,\beta \mid \mathbf{y},\mathbf{x})}{\partial \alpha \partial \beta} & \frac{\partial^2 l(\alpha,\beta \mid \mathbf{y},\mathbf{x})}{\partial \beta^2} \end{bmatrix} = \begin{bmatrix} -\sum_i \exp(\alpha+\beta x_i) & -\sum_i x_i \exp(\alpha+\beta x_i) \\ -\sum_i x_i \exp(\alpha+\beta x_i) & -\sum_i x_i^2 \exp(\alpha+\beta x_i) \end{bmatrix}$$

Newton method:

$$\begin{bmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{bmatrix} = \begin{bmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{bmatrix} - H(\alpha^{(t)},\beta^{(t)}) \begin{bmatrix} \frac{\partial}{\partial \alpha} l\left(\alpha^{(t)},\beta^{(t)}\right) \\ \frac{\partial}{\partial \beta} l\left(\alpha^{(t)},\beta^{(t)}\right) \end{bmatrix}$$

It turns out that *in this special case* the Hessian equals its expectation (the Hessian is non-random in this case), i.e., Information *(=−Hessian)*

$$\mathbb{I}_n(\alpha,\beta \mid \mathbf{y},\mathbf{x}) = -\mathbb{E}\mathbf{H}(\alpha,\beta) = -\begin{bmatrix} \mathbb{E}\sum_i^n \frac{\partial^2 l(\alpha,\beta \mid y_i,x_i)}{\partial^2 \alpha} & \mathbb{E}\sum_i^n \frac{\partial^2 l(\alpha,\beta \mid y_i,x_i)}{\partial \alpha \partial \beta} \\ \mathbb{E}\sum_i^n \frac{\partial^2 l(\alpha,\beta \mid y_i,x_i)}{\partial \alpha \partial \beta} & \mathbb{E}\sum_i^n \frac{\partial^2 l(\alpha,\beta \mid y_i,x_i)}{\partial \beta^2} \end{bmatrix}$$

Generally, the Hessian can be considered the negative observed information. Define

$$\widehat{\mathbb{I}}_n(\alpha,\beta \mid \mathbf{y},\mathbf{x}) = -\mathbf{H}(\alpha,\beta)$$

Using $\mathbb{I}(\alpha,\beta \mid \mathbf{y},\mathbf{x})$ instead of $\mathbf{H}(\alpha,\beta)$ the Newton method is called the method of

"**Fisher scoring**":

$$\begin{bmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{bmatrix} = \begin{bmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{bmatrix} + \mathbb{I}_n^{-1}(\alpha^{(t)},\beta^{(t)}) \begin{bmatrix} \frac{\partial}{\partial \alpha} l\left(\alpha^{(t)},\beta^{(t)}\right) \\ \frac{\partial}{\partial \beta} l\left(\alpha^{(t)},\beta^{(t)}\right) \end{bmatrix}$$

Newton method:

$$\begin{bmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{bmatrix} = \begin{bmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{bmatrix} + \widehat{\mathbb{I}}_n^{-1}(\alpha^{(t)},\beta^{(t)}) \begin{bmatrix} \frac{\partial}{\partial \alpha} l\left(\alpha^{(t)},\beta^{(t)}\right) \\ \frac{\partial}{\partial \beta} l\left(\alpha^{(t)},\beta^{(t)}\right) \end{bmatrix}$$

Newton method for Poisson regression: with $\alpha = \alpha^{(old)},\ \beta = \beta^{(old)}$

$$\begin{bmatrix} \alpha^{(new)} \\ \beta^{(new)} \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \sum_i e^{\alpha+\beta x_i} & \sum_i x_i e^{\alpha+\beta x_i} \\ \sum_i x_i e^{\alpha+\beta x_i} & \sum_i x_i^2 e^{\alpha+\beta x_i} \end{bmatrix}^{-1} \begin{bmatrix} \sum_i \left(y_i - e^{\alpha+\beta x_i}\right) \\ \sum_i x_i \left(y_i - e^{\alpha+\beta x_i}\right) \end{bmatrix}$$

Note: in the special case of Poisson regression $\widehat{\mathbb{I}}_n(\alpha,\beta) = \mathbb{I}_n(\alpha,\beta)$ (the expectation is taken of $\mathbf{y}$ conditional on $\mathbf{x}$) hence *the Newton method is identical to Fisher scoring in this case.*

More over the (*observed*) Fisher information:

$$\widehat{\mathbb{I}}(\widehat{\alpha},\widehat{\beta}) = -\begin{bmatrix} \sum_i^n \frac{\partial^2 l(\widehat{\alpha},\widehat{\beta} \mid y_i,x_i)}{\partial \alpha^2} & \sum_i^n \frac{\partial^2 l(\widehat{\alpha},\widehat{\beta} \mid y_i,x_i)}{\partial \alpha \partial \beta} \\ \sum_i^n \frac{\partial^2 l(\widehat{\alpha},\widehat{\beta} \mid y_i,x_i)}{\partial \alpha \partial \beta} & \sum_i^n \frac{\partial^2 l(\widehat{\alpha},\widehat{\beta} \mid y_i,x_i)}{\partial \beta^2} \end{bmatrix}$$

is an estimator of the the asymptotic covariance matrix

$$\Longrightarrow \begin{bmatrix} \operatorname{var}(\widehat{\alpha}) & \operatorname{cov}(\widehat{\alpha},\widehat{\beta}) \\ \operatorname{cov}(\widehat{\alpha},\widehat{\beta}) & \operatorname{var}(\widehat{\beta}) \end{bmatrix} \approx \begin{bmatrix} \sum_i \exp(\widehat{\alpha}+\widehat{\beta}x_i) & \sum_i x_i \exp(\widehat{\alpha}+\widehat{\beta}x_i) \\ \sum_i x_i \exp(\widehat{\alpha}+\widehat{\beta}x_i) & \sum_i x_i^2 \exp(\widehat{\alpha}+\widehat{\beta}x_i) \end{bmatrix}^{-1}$$

2.4.2 WLS connection

Newton method for Poisson regression

$$\begin{bmatrix} \alpha^{(new)} \\ \beta^{(new)} \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} - \begin{bmatrix} \sum_i e^{\alpha+\beta x_i} & \sum_i x_i e^{\alpha+\beta x_i} \\ \sum_i x_i e^{\alpha+\beta x_i} & \sum_i x_i^2 e^{\alpha+\beta x_i} \end{bmatrix}^{-1} \begin{bmatrix} \sum_i \left(y_i - e^{\alpha+\beta x_i}\right) \\ \sum_i x_i \left(y_i - e^{\alpha+\beta x_i}\right) \end{bmatrix}$$

or, with $\mathbb{E}y_i = e^{\alpha+\beta x_i} = \mu_i$

$$\begin{aligned} \begin{bmatrix} \alpha^{(new)} \\ \beta^{(new)} \end{bmatrix} &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \sum_i \mu_i & \sum_i x_i \mu_i \\ \sum_i x_i \mu_i & \sum_i x_i^2 \mu_i \end{bmatrix}^{-1} \sum_i \begin{bmatrix} 1 \\ x_i \end{bmatrix} (y_i - \mu_i) \\ &= \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \left(\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \end{bmatrix} \begin{bmatrix} 1 & \mu_1 \\ \vdots & \vdots \\ 1 & \mu_n \end{bmatrix} \right)^{-1} \begin{bmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{bmatrix} \end{aligned}$$

or with $\beta = \begin{bmatrix} \alpha & \beta \end{bmatrix}'$

$$\beta^{(new)} = \beta^{(old)} + \left(X' W^{(old)} X\right)^{-1} (\mathbf{y} - \mu^{(old)})$$

weighted least squares connection. Therefore, this representation is called the **method of iteratively reweighted least squares (IRLS)**.

2.4.3 Notes on Poisson regression

excess zeros is a common problem in applications of the Poisson model. An alternative to excess zeros is a mixture model

overdispersion is a common problem in applications of the Poisson regression model: the variance of the observations is visibly larger than the mean. An alternative overdispersion is the negative binomial model.

Powers & Xie gives a nice example of the Poisson regression as modeling the rate of premarital teenage pregnancies, depending on age and socio economic covariates. (time to marriage is a censoring variable)

- the most important case for application for us will be the analysis of contingency tables (i.e., "regression on dummy variables")