TMA4315 Generalized linear models H2018

Module 5: Generalized linear models - common core

Mette Langaas, Department of Mathematical Sciences, NTNU - with contributions from Ingeborg Hem

11.10.2017 [PL], 12.10.2017 [IL]

(Latest changes: 11.010, added links to handwritten materials and dispersion formula, 07.10.2018 first version)

Overview

Learning material

- Textbook: Fahrmeir et al (2013): Chapter 5.4, 5.8.2.
- Classnotes 27.09.2018

Additional notes (with theoretical focus):

- Exponential family from Module 1
- Proof of E and Var for exp fam
- Proof of two forms for F
- Orthogonal parameters
- IRWLS

Topics

- random component: exponential family
  - elements: θ, ϕ, w, $b(\theta)$
  - elements for normal, binomial, Poisson and gamma
  - properties: $\mathsf{E}(Y) = b'(\theta)$ and $\mathsf{Var}(Y) = b''(\theta)\frac{\phi}{w}$ (and proof)
- systematic component= linear predictor
  - requirements: full rank of design matrix
- link function and response function
  - link examples for normal, binomial, Poisson and gamma
  - requirements: one-to-one and twice differentiable
  - canonical link

- likelihood inference set-up: $\theta_i \leftrightarrow \mu_i \leftrightarrow \eta_i \leftrightarrow \beta$
- the loglikelihood
- the score function
- expected Fisher information matrix for the GLM and covariance for $\hat{\beta}$
  - what about covariance of $\hat{\beta}$ when ϕ needs to be estimated?
  - estimator for dispersion parameter
- Fisher scoring and iterated reweighted least squares (IRWLS)
- Pearson and deviance statistic
- AIC

– so, for the first time: no practical examples or data sets to be analysed!

Jump to interactive.

GLM — three ingredients

Random component - exponential family

In Module 1 we introduced distributions of the Y_i, that could be written in the form of a *univariate exponential family*

$$f(y_i \mid \theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} \cdot w_i + c(y_i, \phi, w_i) \right)$$

where we said that

- θ_i is called the canonical parameter and is a parameter of interest
- ϕ is called a nuisance parameter (and is not of interest to us=therefore a nuisance (plage))
- w_i is a weight function, in most cases $w_i = 1$ (NB: can not contain any unknown parameters)
- b and c are known functions.

Elements - Poisson

$$\begin{aligned}\theta &= \log(\mu) \\ b(\theta) &= e^{\theta} \\ \phi &= 1 \\ w &= 1 \\ \mathsf{E}(Y) &= e^{\theta} \\ \mathsf{Var}(Y) &= \phi/w\end{aligned}$$

You can get equivalent results for the normal, Bernoulli, and gamma. Here we will look at the general results

Elements - for normal, Bernoulli, Poisson and gamma

We have seen:

					E$(Y) =$		Var$(Y) =$
Distribution	θ	$b(\theta)$	ϕ	w	$b'(\theta)$	$b''(\theta)$	$b''(\theta)\phi/w$
normal	μ	$\frac{1}{2}\theta^2$	σ^2	1	$\mu = \theta$	1	σ^2
Bernoulli	$\ln\left(\frac{p}{1-p}\right)$	$\ln(1+\exp(\theta))$	1	1	$p = \frac{\exp(\theta)}{1+\exp(\theta)}$	$p(1-p)$	$p(1-p)$
Poisson	$\ln\mu$	$\exp(\theta)$	1	1	$\lambda = \exp(\theta)$	λ	λ
gamma	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	1	$\mu = -1/\theta$	μ^2	μ^2/ν

Systematic component - linear predictor

Nothing new - as always in this course: $\eta_i = \mathbf{x}_i^T \beta$, and we require that the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_n^T)$ has full rank (which is p).
Remark: in this course we always assume that $n >> p$.

Link function - and response function

Link function: $\eta_i = g(\mu_i)$
Response function: $\mu_i = h(\eta_i)$
Canonical link: $\eta_i = \theta_i$, so $g(\mu_i) = \theta_i$ When the canonical link is used some of the results for the GLM (to be studied in the next sections) are simplified.

Examples for normal, binomial, Poisson and gamma

random component	response function and link function
normal	$h(\eta_i) = \eta_i$ and $g(\mu_i) = \mu_i$, “identity link”.
binomial	$h(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ and $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) =$ logit(p_i). NB: $\mu_i = p_i$ in our set-up.
Poisson	$h(\eta_i) = \exp(\eta_i)$ and $g(\mu_i) = \ln(\mu_i)$, log-link.
gamma	$h(\eta_i) = -\frac{1}{\eta_i}$ and $g(\mu_i) = -\frac{1}{\mu_i}$, negative inverse, or $h(\eta_i) = \exp(\eta_i)$ and $g(\mu_i) = \ln(\mu_i)$, log-ink.

Requirements

There are a few formal requirements for the mathematics to work, in particular:

- one-to-one (inverse exists)
- twice differential (for score function and expected Fisher information matrix)

Properties of the exponential family

We have two general properties:

$$\mathsf{E}(Y_i) = b'(\theta_i)$$

and

$$\mathsf{Var}(Y_i) = b''(\theta_i)\frac{\phi}{w_i}$$

In class we study the handwritten proof together: Proof
$b''(\theta_i)$ is often called the variance function $v(\mu_i)$.

The Score (as a function of θ)

The score is $\frac{\partial l}{\partial \theta}$, i.e.

$$\frac{\partial l_i}{\partial \theta} = s_i(\theta) = \frac{\partial \left(\frac{y_i \theta_i - b(\theta_i)}{\phi} \cdot w_i + c(y_i, \phi, w_i) \right)}{\partial \theta}$$

$$= (y_i - b'(\theta)) \frac{w_i}{\phi}$$

The Expected Score

As a general result we have $E(s_i(\theta_i)) = 0$

Proof:

$E(s_i(\theta_i)) = \int \frac{dl(\theta)}{d\theta} f(y|\theta) dy$

and because $d\log(y)/dx = 1/ydy/dx$, we get

$$E(s_i(\theta_i)) = \int \frac{1}{f(y|\theta)} \frac{df(y|\theta)}{d\theta} f(y|\theta) dy = \int \frac{df(y|\theta)}{d\theta} dy$$

Now, if everything is well behaved, we can reverse the integration and differentiation:

$$E(s_i(\theta_i)) = \int \frac{d(y|\theta)}{d\theta} dy = \frac{d \int (y|\theta) dy}{d\theta} = \frac{d1}{d\theta} = 0$$

A Different Proof that E$(Y_i) = b'(\theta_i)$

This is straightforward, from $E(s_i(\theta_i)) = 0$

$$\begin{aligned} E(s) &= E\left((y_i - b'(\theta))\frac{w_i}{\phi}\right) \\ &= (E(y_i) - b'(\theta))\frac{w_i}{\phi} = 0 \\ &= E(y_i) - b'(\theta) = 0 \end{aligned}$$

So $E(y_i) = b'(\theta)$

Variance, $Var(Y_i) = b^{''}(\theta)\phi/w$

Strategy: calculate $\partial^2 f/\partial\theta^2$, then integrate over y

$\int \partial^2 f(y)/\partial\theta^2 dy = 0$ (see notes: we can swap integration & partial derivative)

Go to the notes

Observed Fisher Information

The observed Fisher information is

$$\begin{aligned} \frac{\partial^2 l_i}{\partial \theta^2} &= \frac{\partial s_i(\theta)}{\partial \theta} \\ &= \frac{\partial (y_i - b'(\theta)) \frac{w_i}{\phi}}{\partial \theta} \\ &= -b''(\theta) \frac{w_i}{\phi} \end{aligned}$$

Likelihood inference set-up

We want to estimate β, going from $f(Y|\theta)$:

$\theta_i \leftrightarrow \mu_i \leftrightarrow \eta_i \leftrightarrow \beta$

$$f(y_i|\theta_i) = exp\left(\frac{y_i\theta - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right)$$
$$\theta_i = b^{'-1}(\mu)(\text{from } \mu_i = b'(\theta_i)(= E(Y_i)))$$

$$\mu_i = g^{-1}(\eta_i)$$
$$\eta_i = x_i'\beta$$

$b^{'-1}(\mu)$ is horrible. With the canonical link, $\eta_i = \theta_i$, so $g(\mu_i) = \theta_i$.

See class notes or Fahrmeir et al (2015), Section 5.8.2 for the derivation of the loglikelihood, score and expected Fisher information matrix.

Loglikelihood

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} \frac{1}{\phi}(y_i\theta_i - b(\theta_i))w_i + \sum_{i=1}^{n} c(y_i, \phi, w_i)$$

The part of the loglikelihood involving both the data and the parameter of interest is for a *canonical link* equal to

$$\sum_{i=1}^{n} y_i\theta_i = \sum_{i=1}^{n} y_i\mathbf{x}_i^T\beta = \sum_{i=1}^{n} y_i \sum_{j=1}^{p} x_{ij}\beta_j = \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} y_i x_{ij}$$

Score function

$\theta_i \leftrightarrow \mu_i \leftrightarrow \eta_i \leftrightarrow \beta$

What is the score function as a function of β? We need a long chain rule...

$$s(\beta) = \frac{\partial l}{\partial \beta} = \frac{\partial l(\theta)}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta}$$

We already have $\partial l/\partial \theta = (y_i - b'(\theta))\frac{w_i}{\phi}$, so we need the rest

Score function

$$s(\beta) = \frac{\partial l}{\partial \beta} = \frac{\partial l(\theta)}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta}$$

$$\frac{\partial l}{\partial \theta_i} = (y_i - b'(\theta_i))\frac{w_i}{\phi}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = ...$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial h(\eta_i)}{\partial \eta_i} = h'(\eta_i)$$

$$\frac{\partial \eta_i}{\partial \beta} = \frac{\partial \mathbf{x}_i' \beta}{\partial \beta} = \mathbf{x}_i$$

We get $\frac{\partial\theta_i}{\partial\mu_i}$ by reversing numerator and denominator:

$$\frac{\partial\mu_i}{\partial\theta_i} = \frac{\partial b'(\theta_i)}{\partial\theta_i} = b''(\theta_i) = \frac{w_i \mathsf{Var}(y_i)}{\phi}$$

So

$$\frac{\partial\theta_i}{\partial\mu_i} = \frac{\phi}{w_i \mathsf{Var}(y_i)}$$

Putting it together

$$\begin{aligned}
\frac{\partial l}{\partial \theta_i} &= (y_i - b'(\theta_i))\frac{w_i}{\phi} \\
\frac{\partial \theta_i}{\partial \mu_i} &= \frac{\phi}{w_i \mathsf{Var}(y_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial h(\eta_i)}{\partial \eta_i} = h'(\eta_i) \\
\frac{\partial \eta_i}{\partial \beta} &= \frac{\partial \mathbf{x}_i'\beta}{\partial \beta} = \mathbf{x}_i
\end{aligned}$$

So

$$s(\beta) = (y_i - b'(\theta_i))\frac{w_i}{\phi}\frac{\phi}{w_i \mathsf{Var}(y_i)}h'(\eta_i)\mathbf{x}_i = \frac{(y_i - b'(\theta_i))}{\mathsf{Var}(y_i)}h'(\eta_i)\mathbf{x}_i$$

Total Score

$$s(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mathbf{x}_i h'(\eta_i)}{\mathsf{Var}(Y_i)} = \mathbf{X}^T \mathbf{D} \Sigma^{-1} (\mathbf{y} - \mu)$$

where $\Sigma = \text{diag}(\text{Var}(Y_i))$ and $\mathbf{D} = \text{diag}(h'(\eta_i))$ (derivative wrt η_i).

Remark: observe that $s(\beta) = 0$ only depends on the distribution of Y_i through μ_i and Var(Y_i).

Canonical link

This is neat, because $\frac{\partial \mu_i}{\partial \eta_i} = b''(\theta_i)$:

$$s(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)\mathbf{x}_i w_i}{\phi}$$

Expected Fisher information matrix for the GLM and covariance for $\hat{\beta}$

$$F_{[h,l]}(\beta) = \sum_{i=1}^{n} \frac{x_{ih}x_{il}(h'(\eta_i))^2}{\mathsf{Var}(Y_i)}$$

$$F(\beta) = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

where $\mathbf{W} = \mathsf{diag}(\frac{h'(\eta_i)^2}{\mathsf{Var}(Y_i)})$.

Canonical link:

$$\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_l} = -\frac{x_{ij}w_i}{\phi}(\frac{\partial\mu_i}{\partial\beta_l})$$

which do not contain any random variables, so the observed must be equal to the expected Fisher information matrix.

Fisher scoring and iterated reweighted least squares (IRWLS)

Details on the derivation: IRWLS

$$\beta^{(t+1)} = \beta^{(t)} + F(\beta^{(t)})^{-1} s(\beta^{(t)})$$

Insert formulas for expected Fisher information and score function.

$$\beta^{(t+1)} = (\mathbf{X}^T \mathbf{W}(\beta^{(t)})\mathbf{X})^{-1}\mathbf{X}^T \mathbf{W}(\beta^{(t)})\tilde{\mathbf{y}}_i^{(t)}$$

where $\mathbf{W}$ is as before $\mathbf{W} = \text{diag}(\frac{h'(\eta_i)^2}{\text{Var}(Y_i)})$ - but now the current version of $\beta^{(t)}$ is used. The diagonal elements are called the *working weights*. The $\tilde{\mathbf{y}}_i^{(t)}$ is called the *working response vector* and has element i given as

$$\tilde{\mathbf{y}}_i^{(t)} = \mathbf{x}_i^T \beta^{(t)} + \frac{y_i - h(\mathbf{x}_i^T \beta^{(t)})}{h'(\mathbf{x}_i^T \beta^{(t)})}.$$

Remark: Convergence? With full rank of $\mathbf{X}$ and positive diagonal elements of $\mathbf{W}$ we are certain that the inverse will exist, but there might be that the temporary version of $\mathbf{W}$ can cause problems.

See what is output from `glm`- observe working weights as `weights`..

```
fitgrouped = glm(cbind(y, n - y) ~ ldose, family = "binomia
# names(fitgrouped)
round(fitgrouped$weights, 2)
round(fitgrouped$residuals, 2)
```

```
##     1     2     3     4     5     6     7     8
##  3.25  8.23 14.32 13.38 10.26  5.16  2.65  1.23
##     1     2     3     4     5     6     7     8
##  0.78  0.38 -0.31 -0.44  0.19 -0.06  0.67  1.02
```

Estimator for dispersion parameter

Let data be grouped as much as possible. With G groups (covariate pattern) with n_i observations for each group (then $n = \sum^G n_i = n$):

$$\hat{\phi} = \frac{1}{G-p} \sum_{i=1}^{G} \frac{(y_i - \hat{\mu}_i)^2}{b''(\theta_i)/w_i}$$

The motivation behind this estimator is as follows:

$$\mathsf{Var}(Y_i) = \phi b''(\theta_i)/w_i \Leftrightarrow \phi = \mathsf{Var}(Y_i)/(b''(\theta_i)/w_i)$$

Distribution of the MLE

As before we have that maximum likelihood estimator $\hat{\beta}$ asymptotically follows the multivariate normal distribution with mean β and covariance matrix equal to the inverse of the expected Fisher information matrix. This is also true when we replace the unknown β with the estimated $\hat{\beta}$ for the expected Fisher information matrix.

$$\hat{\beta} \approx N_p(\beta, F^{-1}(\hat{\beta}))$$

and with

$$F(\hat{\beta}) = \mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$$

where $\hat{\mathbf{W}}$ denotes that $\hat{\beta}$ is used then calculating $\mathbf{W} = \text{diag}(\frac{h'(\eta_i)^2}{\text{Var}(Y_i)})$.

What about the distribution of $\hat{\beta}, \hat{\phi}$?

The concept of orthogonal parameters

Hypothesis testing

Same as before - for the Wald we insert the formula for the covariance matrix of $\hat{\beta}$, for the LRT we insert the loglikelihoods and for the score test we insert formulas for the score function and expected Fisher information matrix.

Model assessment and model choice

Pearson and deviance statistic

Group observations together in groups of maximal size (covariate patterns? interval versions thereof?). Group i has n_i observations, and there are G groups. Asymptotic distribution correct if all groups have big n_i. For the non-continuous individual data asymptotic results can not be trusted.

Deviance

$$D = -2[\sum_{i=1}^{g}(l_i(\hat{\mu}_i) - l_i(\bar{y}_i))]$$

with approximate χ^2-distribution with $G - p$ degrees of freedom.

Pearson:

$$X_P^2 = \sum_{i=1}^{G} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

with approximate $\phi \cdot \chi^2$-distribution with $G - p$ degrees of freedom.

Remember that the variance function $v(\hat{\mu}_i) = b''(\theta_i)$ (this is a function of μ_i because $\mu_i = b'(\theta_i)$).

AIC

Let p be the number of regression parameters in our model.

$$\text{AIC} = -2 \cdot l(\hat{\beta}) + 2p$$

If the dispersion parameter is estimated use $(p+1)$ in place of p.

Further reading

- A. Agresti (2015): “Foundations of Linear and Generalized Linear Models.” Wiley.