Chapter 3. Generalised Linear Models (GLM)

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Contents

§3.1 Introduction (1)

- Let $Y_1, Y_2, \cdots, Y_n$ be independent observations of random variable Y, with associated covariate vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$.
- In the case of general linear models, we have, for $i = 1, \cdots, n$

$$Y_i \stackrel{d}{=} N(\mu_i, \sigma^2) \quad \text{independently} \quad \text{and} \quad \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

- We seek to generalise this model to the exponential family: we assume

$$Y_i \stackrel{d}{=} \mathcal{EF}(\text{mean} = \mu_i) \quad \text{independently} \quad \text{and} \quad g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

  where $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ is the **linear predictor**.
- The function $g(\cdot)$ is called the **link function**: it provides the link between the linear predictor and the mean of the response Y.
- If $g(\mu_i)$ is set to equal the **canonical parameter** in the exponential family, $g(\cdot)$ is called the **natural link** or **canonical link**.

§3.1 Introduction (2)

- If the distribution of Y is in **canonical form**, then its pdf $f(y|\theta,\phi)$ satisfies

$$\ln f(y|\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$$

where θ is called the **canonical parameter** or **natural parameter** representing the location, while ϕ is called the **dispersion parameter** representing the scale. We may define various members of the exponential family by specifying functions a, b and c. Often $a(\phi) = \phi/w$ with w being a known **weight**.

- In follows that

$$\frac{\partial \ln f(y|\theta,\phi)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \quad \text{and} \quad \frac{\partial^2 \ln f(y|\theta,\phi)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}$$

- On the other hand

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \ln f(y|\theta,\phi)}{\partial\theta}\right] &= \int \frac{\partial \ln f(y|\theta,\phi)}{\partial\theta} \cdot f(y|\theta,\phi)dy \\
&= \frac{\partial}{\partial\theta}\int f(y|\theta,\phi)dy = \frac{\partial}{\partial\theta}1 = 0, \quad \text{and} \\
\mathbb{E}\left[\frac{\partial^2 \ln f(y|\theta,\phi)}{\partial\theta^2}\right] &= \int \frac{\partial}{\partial\theta}\left(\frac{f'_\theta(y|\theta,\phi)}{f(y|\theta,\phi)}\right) \cdot f(y|\theta,\phi)dy \\
&= \int f''_\theta(y|\theta,\phi)dy - \int\left(\frac{\partial \ln f(y|\theta,\phi)}{\partial\theta}\right)^2 f(y|\theta,\phi)dy \\
&= 0 - \mathbb{E}\left[\left(\frac{\partial \ln f(y|\theta,\phi)}{\partial\theta}\right)^2\right]
\end{aligned}$$

§3.1 Introduction (4)

From the previous two slides, it follows that

$$\mu = \mathbb{E}(Y) = b'(\theta) \quad \text{and} \quad \sigma^2 = \text{var}(Y) = a(\phi)b''(\theta)$$

- **Example.** If $Y \stackrel{d}{=} \text{Poi}(\lambda)$, then $\ln f(y|\lambda) = -\lambda + y\ln\lambda - \ln y!$, so that $\theta = \ln\lambda$.
  Thus, we have $\ln f(y|\theta) = -e^\theta + y\theta - \ln y!$; so that, $b(\theta) = e^\theta$ and $a(\phi) = 1$. Applying the above result gives
  $\mu = \mathbb{E}(Y) = b'(\theta) = e^\theta = \lambda$ and $\sigma^2 = \text{var}(Y) = a(\phi)b''(\theta) = e^\theta = \lambda$.
- **Example.** If $Y \stackrel{d}{=} \text{Bin}(m, p)$, then $\theta = \ln\frac{p}{1-p}$, and $b(\theta) = m\ln(1-p) = m\ln(1+e^\theta)$.
  Thus, we have $\mu = \mathbb{E}(Y) = b'(\theta) = \dfrac{me^\theta}{1+e^\theta} = mp$ and
  $\sigma^2 = \text{var}(Y) = a(\phi)b''(\theta) = \dfrac{me^\theta}{(1+e^\theta)^2} = mp(1-p)$.

§3.2 Estimation (1)

We consider in some detail a fairly standard case — the **Poisson regression model with log link**, which will act as a template for the GLM with link function equal the natural parameter:

Example. *Poisson regression model with log link*

- $Y_i \stackrel{d}{=} \text{Poi}(\lambda_i)$, $i = 1, 2, \cdots, n$; independent.
- The mean parameters λ's depend on the covariates $x_1, x_2, \cdots, x_q$.
- The natural parameter $\theta_i = \ln \lambda_i$. The natural link is $g(\lambda_i) = \ln \lambda_i$.
- This gives a log-linear model $\eta_i = \ln \lambda_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \sum_{\ell=1}^{q} x_{i\ell}\beta_\ell$,
- Its matrix form: $\ln \boldsymbol{\lambda} = \boldsymbol{\eta} = X\boldsymbol{\beta}$, where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_n)^\top$.

§3.2 Estimation (2)

- Joint log-likelihood function

$$\begin{aligned}\ell(\boldsymbol{\beta}) = \ln f &= -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} y_i \ln \lambda_i - \sum_{i=1}^{n} \ln y! \\ &= -\sum_{i=1}^{n} e^{\sum_{\ell=1}^{q} x_{i\ell}\beta_\ell} + \sum_{i=1}^{n} y_i \sum_{\ell=1}^{q} x_{i\ell}\beta_\ell + \text{const}\end{aligned}$$

- Score function

$$\mathbf{u}(\boldsymbol{\beta}) = \left(\frac{\partial \ln f}{\partial \beta_j}\right) = \left(-\sum_{i=1}^{n} x_{ij} e^{\sum_{\ell=1}^{q} x_{i\ell}\beta_\ell} + \sum_{i=1}^{n} x_{ij} y_i\right) = X^\top(\mathbf{y} - \boldsymbol{\lambda})$$

- Hessian function

$$H(\boldsymbol{\beta}) = \left(\frac{\partial^2 \ln f}{\partial \beta_j \partial \beta_k}\right) = \left(-\sum_{i=1}^{n} x_{ij} x_{ik} e^{\sum_{\ell=1}^{p} x_{i\ell}\beta_\ell}\right) = -X^\top \Lambda X$$

where $\Lambda = \Lambda(\boldsymbol{\beta}) = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

§3.1 Estimation (3)

Thus

- Observed information $J(\boldsymbol{\beta}) = -H(\boldsymbol{\beta}) = X^\top \Lambda X$.
- Fisher information $I(\boldsymbol{\beta}) = -\mathbb{E}[H(\boldsymbol{\beta})] = X^\top \Lambda X$.
- Variance of score function $\mathrm{var}(\mathbf{u}(\boldsymbol{\beta})) = I(\boldsymbol{\beta}) = X^\top \Lambda X \overset{denoted}{=} V(\boldsymbol{\beta})$.

Method of scoring (1)

MLE $\hat{\beta}$ is solved by *Newton-Raphson* or *Fisher scoring*

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{(k+1)} &= \hat{\boldsymbol{\beta}}^{(k)} + \left[\hat{V}^{(k)}\right]^{-1} \mathbf{u}(\hat{\boldsymbol{\beta}}^{(k)}), \quad \text{equivalently} \\
\hat{V}^{(k)}\hat{\boldsymbol{\beta}}^{(k+1)} &= \hat{V}^{(k)}\hat{\boldsymbol{\beta}}^{(k)} + \mathbf{u}(\hat{\boldsymbol{\beta}}^{(k)}) \\
X^\top \hat{\Lambda}^{(k)} X \hat{\boldsymbol{\beta}}^{(k+1)} &= X^\top \hat{\Lambda}^{(k)} X \hat{\boldsymbol{\beta}}^{(k)} + X^\top \hat{\Lambda}^{(k)} \left[(\hat{\Lambda}^{(k)})^{-1}(\mathbf{y} - \hat{\boldsymbol{\lambda}}^{(k)})\right] \\
X^\top \hat{\Lambda}^{(k)} X \hat{\boldsymbol{\beta}}^{(k+1)} &= X^\top \hat{\Lambda}^{(k)} \hat{\mathbf{z}}^{(k)}
\end{aligned}
$$

where $\hat{\mathbf{z}}^{(k)} = X\hat{\boldsymbol{\beta}}^{(k)} + (\hat{\Lambda}^{(k)})^{-1}(\mathbf{y} - \hat{\boldsymbol{\lambda}}^{(k)})$, $\hat{\boldsymbol{\lambda}}^{(k)} = \boldsymbol{\lambda}(\hat{\boldsymbol{\beta}}^{(k)})$ and $\hat{\Lambda}^{(k)} = \Lambda(\hat{\boldsymbol{\beta}}^{(k)})$.

- This essentially is weighted least squares equation of the form

$$X^\top \mathbf{W} X \hat{\boldsymbol{\beta}} = X^\top \mathbf{W}\mathbf{y}.$$

Method of scoring (2)

- Thus the iterative step in the method of scoring amounts to fitting a weighted regression of $\hat{z}_i^{(k)}$ on $\mathbf{x}_i$ with weights $\hat{\lambda}_i^{(k)}$.
- Note that values of $\hat{\lambda}_i^{(k)}$ and $\hat{z}_i^{(k)}$ need to be updated at each iteration since both depend on $\hat{\beta}^{(k)}$.
- This representation of the iterative step in the method of scoring as a weighted regression can always be done for a generalised linear model: the procedure is called the method of **iteratively re-weighted least squares**(IRWLS).
- The IRWLS method is implemented in the `R` function `glm()`:

```
glm(formula, family = poisson, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)
```

Method of scoring (3)

Example. Fit the model $Y_i \stackrel{d}{=} \text{Poi}(\lambda_i)$, where $\ln \lambda_i = \beta_0 + \beta_1 x_i$ to the following data:

x	-1	-1	0	0	0	0	1	1	1
y	2	3	6	7	8	9	10	12	15

```
> x <- c(-1, -1, 0, 0, 0, 0, 1, 1, 1)
> y <- c(2, 3, 6, 7, 8, 9, 10, 12, 15)
```

In R

```
> pr.1 <- glm(y ~ x, family = poisson)
> summary(pr.1)

Call:
glm(formula = y ~ x, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8472  -0.2601  -0.2137   0.5214   0.8788

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8893     0.1421  13.294  < 2e-16 ***
x             0.6698     0.1787   3.748 0.000178 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18.4206  on 8  degrees of freedom
Residual deviance:  2.9387  on 7  degrees of freedom
AIC: 41.052

Number of Fisher Scoring iterations: 4
```

GLM estimation with natural link (1)

The general exponential family case with natural link function is little different to the above derivation for the Poisson case.
In the general case with natural link and $a(\phi) = \phi/w$ we have

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \phi^{-1} \sum_{i=1}^{n} w_i [y_i \theta_i - b(\theta_i)] + \text{const} \quad \text{where } \theta_i = \eta_i = \textstyle\sum_{j=1}^{q} \beta_j x_{ij} \\
u(\boldsymbol{\beta}) &= \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \left(\phi^{-1} \sum_{i=1}^{n} w_i (y_i - \mu_i) x_{ij} \right) = \phi^{-1} X^\top W (\mathbf{y} - \boldsymbol{\mu}) \\
J(\boldsymbol{\beta}) &= -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \boldsymbol{\beta}^\top} = \left(\phi^{-1} \sum_{i=1}^{n} w_i b''(\theta_i) x_{ij} x_{ik} \right) = \phi^{-1} X^\top W \Sigma X = I(\boldsymbol{\beta})
\end{aligned}$$

where $W = \text{diag}\{w_1, \cdots, w_n\}$ and $\Sigma = \text{diag}\{b''(\theta_1), \cdots, b''(\theta_n)\}$.

GLM estimation with natural link(2)

Writing $V = X^\top W\Sigma X$, the **MLE** $\hat{\beta}$ is solved by the method of scoring

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{(k+1)} &= \hat{\boldsymbol{\beta}}^{(k)} + \phi\left[\hat{V}^{(k)}\right]^{-1}\mathbf{u}(\hat{\boldsymbol{\beta}}^{(k)}), \quad \text{equivalently} \\
\phi^{-1}\hat{V}^{(k)}\hat{\boldsymbol{\beta}}^{(k+1)} &= \phi^{-1}\hat{V}^{(k)}\hat{\boldsymbol{\beta}}^{(k)} + \mathbf{u}(\hat{\boldsymbol{\beta}}^{(k)}) \\
X^\top W\hat{\Sigma}^{(k)}X\hat{\boldsymbol{\beta}}^{(k+1)} &= X^\top W\hat{\Sigma}^{(k)}X\hat{\boldsymbol{\beta}}^{(k)} + X^\top W\hat{\Sigma}^{(k)}\left[(\hat{\Sigma}^{(k)})^{-1}(\mathbf{y}-\hat{\boldsymbol{\mu}}^{(k)})\right] \\
X^\top W\hat{\Sigma}^{(k)}X\hat{\boldsymbol{\beta}}^{(k+1)} &= X^\top W\hat{\Sigma}^{(k)}\hat{\mathbf{z}}^{(k)}
\end{aligned}
$$

where $\hat{\mathbf{z}}^{(k)} = X\hat{\boldsymbol{\beta}}^{(k)} + (\hat{\Sigma}^{(k)})^{-1}(\mathbf{y}-\hat{\boldsymbol{\mu}}^{(k)})$, $\hat{\boldsymbol{\mu}}^{(k)} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(k)})$ and $\hat{\Sigma}^{(k)} = \Sigma(\hat{\boldsymbol{\beta}}^{(k)})$.

GLM estimation with natural link (3)

From the last two lines, the method of scoring becomes the IRWLS

$$X^\top W\hat{\Sigma}^{(k)}X\hat{\boldsymbol{\beta}}^{(k+1)} \;=\; X^\top W\hat{\Sigma}^{(k)}\hat{\mathbf{z}}^{(k)}$$

where $\hat{\mathbf{z}}^{(k)} = X\hat{\boldsymbol{\beta}}^{(k)} + (\hat{\Sigma}^{(k)})^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(k)})$, $\hat{\boldsymbol{\mu}}^{(k)} = \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}^{(k)})$ and $\hat{\Sigma}^{(k)} = \Sigma(\hat{\boldsymbol{\beta}}^{(k)})$.

- The IRWLS procedure produces the MLE $\hat{\boldsymbol{\beta}}$ in convergence.
- The variance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated as

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\left(X^\top W\Sigma(\hat{\boldsymbol{\beta}})X\right)^{-1}.$$

§3.3 Inference

- To do any inference, we need to know something about the distributions of the statistics. The underpinning of GLM is the asymptotic likelihood theory (Note $q = \dim(\mathbf{x})$):

$$\mathbf{u} \overset{d}{\sim} N_q(0, I(\boldsymbol{\beta})) \quad \text{and} \quad \mathbf{u}^\top I^{-1}\mathbf{u} \overset{d}{\sim} \chi^2(q)$$

- For GLMs, V does not involve $\mathbf{y}$ and so $I(\boldsymbol{\beta}) = \phi^{-1}V$ and we have:

$$\mathbf{u} \overset{d}{\sim} N_q(0, \phi^{-1}V) \quad \text{and} \quad \phi\mathbf{u}^\top V^{-1}\mathbf{u} \overset{d}{\sim} \chi^2(q)$$

- Further, the ML estimators are such that

$$\hat{\boldsymbol{\beta}} \overset{d}{\sim} N_q(\boldsymbol{\beta}, \phi V^{-1}) \quad \text{and} \quad \phi^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top V(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\sim} \chi^2(q)$$

- In general V involves $\boldsymbol{\beta}$, and so we need to use $V(\hat{\boldsymbol{\beta}})$ to compute se's.

$$V = [v_{ij}(\boldsymbol{\beta})] \quad \text{and} \quad V^{-1} = \left[v^{ij}(\boldsymbol{\beta})\right]$$

$$\text{se}(\hat{\beta}_j) = \sqrt{\phi v^{jj}(\hat{\boldsymbol{\beta}})}; \qquad \text{approx 95\% CI:} \;\; \hat{\beta}_j \pm 1.96 \cdot \text{se}(\hat{\beta}_j)$$

Checking the adequacy of the model (1)

- The adequacy of a model M can be assessed by comparing the likelihood of the model M with the likelihood of the 'full (saturated) model', F — which is a GLM with the same distribution and same link but with n parameters.
- If model M is a good one, then $L(\boldsymbol{\beta}_M)$ will be close to $L(\boldsymbol{\beta}_F)$: it is able to explain most of the variation in the data.
- We define as test statistic

$$D = 2\phi[\ln L(\hat{\boldsymbol{\beta}}_F) - \ln L(\hat{\boldsymbol{\beta}}_M)]$$

This is called the **residual deviance** (or just the **deviance**) and is a sort-of analogue of the residual sum of squares in linear model (but now there is no σ^2 to estimate, so now if D is big it's because the model is of poor fit).

Checking the adequacy of the model (2)

- Testing is based on the result

$$\phi^{-1} D \overset{d}{\approx} \chi^2(n-q) \quad \text{if model } M \text{ is correct}$$

  and otherwise, D tends to be bigger (indicating that M is not a good fit).

- Thus

$$\text{model } M \text{ is acceptable if} \quad D < \chi^2_{0.95}(n-q)$$

- The R output

```
Residual Deviance: 2.938747 on 7 degrees of freedom
```

  now makes a bit more sense.

- It says that $D = 2.94$ and $n - q = 7$ (since $n = 9$ and $q = 2$) so that the model we fitted is quite acceptable: $\chi^2_{0.95}(7) = 14.07$.

Theory behind the model adequacy test

- The distributional result for D comes from the asymptotic likelihood result:

$$\ln L(\boldsymbol{\beta}) \approx \ln L(\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{u}(\hat{\boldsymbol{\beta}}) - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top J(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

from which it follows that (Note $\mathbf{u}(\hat{\boldsymbol{\beta}}) = 0$ at MLE $\hat{\boldsymbol{\beta}}$)

$$2[\ln L(\hat{\boldsymbol{\beta}}) - \ln L(\boldsymbol{\beta})] \approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top J(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\sim} \chi^2(q)$$

- A further standard check on the adequacy of the model is to look at the residuals, using the diagnostic tools introduced in Chapters 1 & 2.
- Denote $D = \sum_{i=1}^n d_i^2$ with d_i^2 being the contribution from the i-th observation. Then the **deviance residual** for observation i is defined to be

$$r_i = \text{sign}(y_i - \hat{y}_i)|d_i|, \quad i = 1, \cdots, n$$

Comparing nested models

- Tests for the comparison of nested models proceed in much the same way as for the general linear model theory — except that now there is no σ^2 to be estimated.
- So the test can be based directly on the change in SS analogue (i.e. changes in residual deviance).
- Suppose model M_0 is nested in model M_1.
- Assume M_1 (the model under H_1) is true. Then

$$\Delta D = D_0 - D_1 \overset{d}{\approx} \chi^2(q_1 - q_0) \quad \text{if } M_0 \text{ is also true;}$$

  and if M_0 is not true, then ΔD tends to be larger.
- Thus

$$\text{we reject } M_0 \text{ if } \Delta D > \chi^2_{0.95}(q_1 - q_0)$$

In R

```
> anova(pr.1, test = "Chi")
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     8    18.4206
x     1  15.4819         7     2.9387    0.0001
```

From R

The output tells us that $D_0 = 18.42$ and $D_1 = 2.94$ so that

$$\Delta = D_0 - D_1 = 15.48$$

which, if M_0 were true would be an observation on $\chi^2(1)$.

As $\chi^2_{0.95}(1) = 3.841$, we reject M_0; and conclude that $\beta_1 \neq 0$
— as we already would have from consideration of $\hat{\beta}_1$ and $\mathrm{se}(\hat{\beta}_1)$.

Remark: Comparing $M_0 \subset M_1$ with M_1 is equivalent to testing the linear hypothesis $H_0 : \boldsymbol{\beta}_{M_1 - M_0} = 0$. We have learned three tests for this comparison: LR test, Wald test and Score test. The test corresponding to $\Delta = D_0 - D_1$ is the LR test.