Exponential families and GLMs

Exponential families

Exponential families

Y comes from an exponential family if it has density/mass function of the form

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

- θ is the *canonical parameter* (captures location)
- ϕ is the *dispersion parameter* (captures scale)

Example: normal

$Y \sim N(\mu, \sigma^2)$

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}} \\ &= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \end{aligned}$$

where $\theta = \mu$, $\phi = \sigma^2$, and

$$\begin{aligned} b(\theta) &= \theta^2/2 \\ a(\phi) &= \phi \\ c(y, \phi) &= -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right) \end{aligned}$$

Example: Poisson

$Y \sim \text{pois}(\lambda)$

$$\begin{aligned} f(y) &= e^{-\lambda}\lambda^y/y! \text{ for } y = 0, 1, 2, \ldots \\ &= \exp\left[y \log \lambda - \lambda - \log y!\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \end{aligned}$$

where $\theta = \log \lambda$, $\phi = 1$, and

$$\begin{aligned} b(\theta) &= e^{\theta} \\ a(\phi) &= \phi \\ c(y, \phi) &= -\log y! \end{aligned}$$

Example: binomial

$Y \sim \text{bin}(m, p)$ for known m (not a parameter)

$$\begin{aligned} f(y) &= \binom{m}{y} p^y (1-p)^{m-y} \text{ for } y = 0, 1, \ldots, m \\ &= \exp\left[y \log \frac{p}{1-p} + m \log(1-p) + \log \binom{m}{y} \right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right] \end{aligned}$$

where $\theta = \log \frac{p}{1-p}$, $\phi = 1$, and

$$\begin{aligned} b(\theta) &= m \log(1 + e^\theta) \\ a(\phi) &= \phi \\ c(y, \phi) &= \log \binom{m}{y} \end{aligned}$$

Other examples of exponential families are the gamma and the inverse Gaussian.

Exponential family: mean and variance

Lemma If Y is from an exponential family then

$$\begin{aligned} \mathbb{E}\, Y &= b'(\theta) \\ \operatorname{Var} Y &= b''(\theta) a(\phi) \end{aligned}$$

[Proof] Exercise.

Exponential family: variance function

Let $\mu = \mathbb{E}\, Y$ and write

$$\operatorname{Var} Y = v(\mu)a(\phi)$$

(so $v = b'' \circ (b')^{-1}$). v is called the *variance function*

Examples:

normal	$v(\mu) = 1$
Poisson	$v(\mu) = \mu$
binomial	$v(\mu) = \mu(1 - \mu/m)$

Generalised Linear Model

Generalised Linear Model

Definition: Y is a GLM if it is from an exponential family, and

$$\mu := \mathbb{E}\, Y = g^{-1}(\mathbf{x}^T \boldsymbol{\beta})$$

where

- g is a monotonic differentiable function called the *link function*.
- $\mathbf{x}$ is a vector of independent (predictor) variables, and
- $\boldsymbol{\beta}$ is a vector of parameters

Remark: We model *location* using $\eta = \mathbf{x}^T \boldsymbol{\beta}$, and let the *scale* sort itself out. That is, we do not model the scale explicitly.

Canonical link

Recall Y is from an exponential family if

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

If $g(\mu) = g(\mathbb{E}\,Y) = \theta$ then g is called the *cannonical* link. Since $\mu = b'(\theta)$, it follows that the canonical link must be $(b')^{-1}$.

Examples: canonical links

normal $\theta = \mu,\ g(\mu) = \mu$

Poisson $\theta = \log \lambda = \log \mu,\ g(\mu) = \log \mu$

binomial $\theta = \log \frac{p}{1-p} = \log \frac{\mu}{m-\mu},\ g(\mu) = \log \frac{\mu}{m-\mu}$

Estimation of parameters in GLM

GLM fitting

We fit GLMs using maximum likelihood.

Suppose we have independent observations y_i from an exponential family, with canonical parameter θ_i and dispersion parameter ϕ, for $i = 1, \ldots, n$.

Furthermore suppose that y_i has mean

$$\mu_i = b'(\theta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

If g is the canonical link then $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$.

The log-likelihood is then

$$l(\boldsymbol{\beta}, \phi; \mathbf{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right).$$

Newton-Raphson method

If f has a continuous derivative f', then the problem of finding the maximum of f is equivalent to finding the maximum of $f(x_1^*), \ldots, f(x_k^*)$, where $x_1^*, \ldots, x_k^*$ are the roots of f'.

We apply the Newton-Raphson method for root-finding to f':

$$x^{n+1} = x^n - \frac{f'(x^n)}{f''(x^n)}.$$

Newton-Raphson method to find MLE

Suppose we wish to maximise a log likelihood $l(\boldsymbol{\theta})$ using Newton-Raphson method. Our update step is

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n - H(\boldsymbol{\theta}^n)^{-1} U(\boldsymbol{\theta}^n)$$

where

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \text{ and } H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

That is, $-H(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$, the observed information.

If we replace $\mathcal{J}$ by $\mathcal{I}$, the Fisher information, then the algorithm is called Fisher scoring.

The Fisher information is guaranteed to be positive definite (unlike the observed information).

Weighted Least Squares

It turns out that Fisher scoring applied to a GLM can be interpreted as a least squares problem (if you want to know details, read "McCullagh & Nelder on IWLS" posted on LMS).
Consider

$$\begin{aligned} g(Y_i) \quad \approx \quad Z_i \quad &:= \quad g(\mu_i) + (Y_i - \mu_i) g'(\mu_i) \\ &= \quad \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \text{ say} \end{aligned}$$

where

$$\operatorname{Var} \epsilon_i = (g'(\mu_i))^2 \operatorname{Var} Y_i.$$

If we knew $\operatorname{Var} \epsilon_i$ then the estimator of $\boldsymbol{\beta}$ would be the solution to the weighted least squares problem:

$$\min_{\boldsymbol{\beta}} (\mathbf{z} - X\boldsymbol{\beta})^T \Sigma^{-1} (\mathbf{z} - X\boldsymbol{\beta})$$

where Σ is diagonal with $\Sigma_{ii} = \operatorname{Var} \epsilon_i$.

Aside: Weighted LS from Ordinary LS

Suppose $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Sigma)$, and X and Σ are full rank. Multiplying by $\Sigma^{-1/2}$ we get $\Sigma^{-1/2}\mathbf{Y} = \Sigma^{-1/2}X\boldsymbol{\beta} + \boldsymbol{\varepsilon}'$ where $\boldsymbol{\varepsilon}' \sim N(\mathbf{0}, I)$.

From this we see that the estimator of $\boldsymbol{\beta}$ minimises the sum of squares

$$(\Sigma^{-1/2}\mathbf{y} - \Sigma^{-1/2}X\boldsymbol{\beta})^T(\Sigma^{-1/2}\mathbf{y} - \Sigma^{-1/2}X\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{y} - X\boldsymbol{\beta})$$

and the least squares estimator is

$$(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{y}.$$

Weighted LS for GLM

We have $g(\mathbf{Y}) \approx \mathbf{Z} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\epsilon_i = (Y_i - \mu_i)g'(\mu_i)$ and $\Sigma = \operatorname{Var} \boldsymbol{\varepsilon}$ is diagonal with entries

$$\Sigma_{ii} = (g'(\mu_i))^2 \operatorname{Var} Y_i = (g'(\mu_i))^2 v(\mu_i) a(\phi).$$

This suggests

$$\hat{\boldsymbol{\beta}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \mathbf{z}.$$

Problem: clearly z_i depends on $\boldsymbol{\beta}$, but $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$, so Σ_{ii} also depends on $\boldsymbol{\beta}$.

Solution: iterate!

Note that the $a(\phi)$ factor in the expression for $\hat{\boldsymbol{\beta}}$ cancels out.

Iterated Weighted Least Squares (IWLS)

IWLS algorithm (for estimating β)

1. Start with $\hat{\mu}^0 = \mathbf{y}$.
2. Given $\hat{\mu}^n$ calculate, for each i,
   $z_i^n = g(\hat{\mu}_i^n) + (y_i - \hat{\mu}_i^n)g'(\hat{\mu}_i^n)$ and
   $W_{ii}^n = 1/[g'(\hat{\mu}_i^n)^2 v(\hat{\mu}_i^n)]$.
3. Put $\hat{\beta}^{n+1} = (X^T W^n X)^{-1} X^T W^n \mathbf{z}^n$ and
   for each i, $\hat{\mu}_i^{n+1} = g^{-1}(\mathbf{x}_i^T \hat{\beta}^{n+1})$.
4. If $\hat{\beta}^{n+1}$ is sufficiently close to $\hat{\beta}^n$ then stop, otherwise return to (2).

For a GLM, IWLS is equivalent to Fisher scoring.
Example: see the section "IWLS" in Bliss.pdf.

Variance of $\hat{\boldsymbol{\beta}}$ from IWLS

Suppose that the IWLS algorithm converges to the estimate $\hat{\boldsymbol{\beta}}$, then

$$\hat{\boldsymbol{\beta}} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} \mathbf{z}$$

where, elementwise, we have

$$\begin{aligned} \hat{\mu}_i &= g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \\ z_i &= \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + (y_i - \hat{\mu}_i) g'(\hat{\mu}_i) \\ \hat{\Sigma}_{ii} &= (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i) a(\phi) \end{aligned}$$

Since $\operatorname{Var} \mathbf{z} = \hat{\Sigma}$ we have

$$\begin{aligned} \operatorname{Var} \hat{\boldsymbol{\beta}} &= [(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}] \hat{\Sigma} [(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}]^T \\ &= (X^T \hat{\Sigma}^{-1} X)^{-1} \end{aligned}$$

Variance of $\hat{\boldsymbol{\beta}}$ from IWLS

Note that the $a(\phi)$ term in $\hat{\Sigma}$ does not cancel here, as it did in the IWLS algorithm, so we need to estimate it.

Now $(Y_i - \mu_i)/\sqrt{v(\mu_i)}$ has mean 0 and variance $a(\phi)$, so it should come as no surprise that

$$X^2 := \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \approx a(\phi)\chi^2_{n-p}$$

where p is the number of parameters used to estimate $\boldsymbol{\mu}$. Thus $X^2/(n-p)$ will be an estimator for $a(\phi)$.

X^2 is called Pearson's χ^2 statistic, and it can be shown that $X^2/(n-p)$ is a consistent estimator for $a(\phi)$.

Deviance

Deviance

From here on we will w.l.o.g. adopt the convention that $a(\phi) = \phi$.

Definition: the scaled deviance for model A is

$$\frac{D^A}{\phi} = -2 \log \frac{\mathcal{L}(\hat{\beta}^A)}{\mathcal{L}(\mathrm{S})}$$

where $\hat{\beta}^A$ is the MLE of β^A, the true parameter value for model A, and $\mathcal{L}(\mathrm{S})$ is the maximum likelihood for the saturated model

The deviance is just D^A.

Example: normal

The saturated normal model uses y_i to estimate μ_i.
The deviance can be written as $D = \sum_i d_i$ where

$$d_i = (y_i - \hat{\mu}_i)^2$$

where $\hat{\mu}_i$ is the fitted mean using the MLE.

Example: Poisson

The saturated Poisson model uses y_i to estimate $\mu_i = \lambda_i$.
The deviance can be written as $D = \sum_i d_i$ where

$$d_i = -2\left(y_i \log \frac{\hat{\mu}_i}{y_i} - (\hat{\mu}_i - y_i)\right)$$

where $\hat{\mu}_i$ is the fitted mean using the MLE.

Example: binomial

The saturated binomial model uses y_i to estimate μ_i, that is y_i/m_i to estimate p_i.
The deviance can be written as $D = \sum_i d_i$ where

$$d_i = -2\left(y_i \log \frac{\hat{\mu}_i}{y_i} + (m_i - y_i) \log \frac{m_i - \hat{\mu}_i}{m_i - y_i}\right)$$

where $\hat{\mu}_i$ is the fitted mean using the MLE.

Deviance

If the model is adequate then the scaled deviance will often (but not always) be $\approx \chi^2_{n-p}$, where the full model has n parameters (equal to the number of observations) and the fitted model has p parameters.

For nested models, if the smaller model is correct then the difference between two scaled deviances is the log likelihood ratio and will be $\approx \chi^2_s$ for large n, where s is the difference in the number of parameters.

The scaled deviance can be used to test model adequacy.
The difference between two scaled deviances can be used to test the significance of the extra parameters (a log likelihood ratio test).

Example : See "LR test using deviance" in Bliss.pdf

Deviance

For the binomial and Poisson models $\phi = 1$ and the scaled deviance is just the deviance.
For these models the sclaed deviance will be approximately χ^2 when the individual responses are somewhat normal. As a rule of thumb we need the Poisson rate or the binomial mean (when the success rate is small) to be at least 5. (When the success rate is close to 1 we want the number trials minus the mean to be at least 5.)

For the normal, gamma or inverse gaussian models, we have to estimate ϕ, using $X^2/(n-p)$.
For these models we can't use the scaled deviance to test model adequacy.

Akaike Information Criterion (AIC)

The AIC is defined as

$$\text{AIC} = 2p - 2\log \mathcal{L}(\hat{\boldsymbol{\beta}})$$

where p is the number of parameters in the model. Given a choice, we prefer that model with the smaller AIC.

If model B has s more parameters than model A (not necessarily nested within B), then

$$\begin{aligned} \text{AIC}^B - \text{AIC}^A &= 2s - 2\log \mathcal{L}(\hat{\boldsymbol{\beta}}^B) + 2\log \mathcal{L}(\hat{\boldsymbol{\beta}}^A) \\ &= 2s + \frac{D^B}{\phi} - \frac{D^A}{\phi}. \end{aligned}$$

Like the log likelihood ratio, the AIC needs an estimate of ϕ.

Diagnostics

Residuals

Response residuals: $y_i - \hat{\mu}_i$
Unless $v(\mu)$ is constant, as in the Gaussian case, the response residuals are not homoskedastic and hence not very useful.

Pearson residuals:

$$r_P(i) = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$$

Pearson residuals are (approximately) homoskedastic, and $\sum_i r_P(i)^2 = X^2$.

Deviance residuals:

$$r_D(i) = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where the deviance is $D = \sum_i d_i = \sum_i r_D(i)^2$.

Residuals

As for linear models, patterns in the residuals indicate structure in the data that has not been captured by the model.

We can plot the residuals against predictor variables, the responses, or the fitted means. Often a plot against $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ works well.

With count data residual plots exhibit banding due to the discrete nature of the responses, and this can make it hard to see other patterns.
In this case we can use a smoothed fit of the residuals to help spot trends/patterns.

Leverage

The leverage measures the potential influence of a point on the fitted model. We borrow the definition of leverage from the theory of linear models, and use the hat matrix from the IWLS fitting.

In the IWLS scheme we have $\Sigma_{ii} = (g'(\mu_i))^2 v(\mu_i)\phi$ (assuming $a(\phi) = \phi$) and

$$\mathbf{Z}' = \Sigma^{-1/2}\mathbf{Z} = \Sigma^{-1/2}X\beta + \varepsilon'$$

where $\operatorname{Var}\varepsilon' = I$. The hat matrix for $\hat{\mathbf{Z}}'$ is the matrix H' such that $H'\mathbf{Z}' = \hat{\mathbf{Z}}'$. From the theory of linear models

$$H' = \Sigma^{-1/2}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1/2}$$

Thus $\hat{\mathbf{Z}} = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\mathbf{Z}$, and the hat matrix for $\hat{\mathbf{Z}}$ is

$$H = X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}.$$

Leverage

$$H = X(X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}.$$

The leverage for the i-th observation is H_{ii}, the i-th diagonal element of H.

Note that H does not depend on ϕ (it cancels out).

A large leverage does not necessarily mean a point *has* influenced the fit.

Studentised residuals

The variance of $\hat{\mathbf{Z}}$ is

$$\begin{aligned} & H\mathrm{Var}\,\mathbf{Z}H^T \\ &= [X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}]\Sigma[\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T] \\ &= X(X^T\Sigma^{-1}X)^{-1}X^T \quad = \quad H\Sigma \end{aligned}$$

Whence $\mathrm{Var}\,(\mathbf{Z} - \hat{\mathbf{Z}}) = (I - H)\Sigma$.

Studentised residuals

From the above we have $\operatorname{Var} \frac{Z_i - \hat{Z}_i}{\sqrt{(1-H_{ii})g'(\mu_i)^2 v(\mu_i)\phi}} = 1$. Also

$$\begin{aligned} Z_i - \hat{Z}_i &= g(\mu_i) + (Y_i - \mu_i)g'(\mu_i) - [g(\mu_i) + (\hat{Y}_i - \mu_i)g'(\mu_i)] \\ &= (Y_i - \hat{Y}_i)g'(\mu_i) \end{aligned}$$

so, noting that $\hat{Y}_i = \hat{\mu}_i$,

$$\operatorname{Var} \frac{Y_i - \hat{\mu}_i}{\sqrt{(1 - H_{ii})v(\hat{\mu}_i)\hat{\phi}}} \approx 1$$

The LHS is just $r_P(i)/\sqrt{(1 - H_{ii})\hat{\phi}} =: r_{SP}(i)$, which we call the i-th **studentised Pearson residual**.

Studentised residuals

Studentised Pearson residual:

$$r_{SP}(i) = \frac{r_P(i)}{\sqrt{(1 - H_{ii})\hat{\phi}}}$$

By analogy we define **Studentised deviance residual:**

$$r_{SD}(i) = \frac{r_D(i)}{\sqrt{(1 - H_{ii})\hat{\phi}}}$$

Jack-knife residuals

A direct measure of the influence of a point is the **jack-knife residual**, which is the change in $\hat{\mu}_i$ when you remove y_i from the set of observations, then scaled to standardise the variance.

The jack-knife residual can be approximated by

$$\text{sign}(y_i - \hat{\mu}_i)\sqrt{(1 - H_{ii})r_{SD}^2(i) + H_{ii}r_{SP}^2(i)}.$$

Cook's distance

Another measure of the influence of the i-th observation is **Cook's distance**:

$$\frac{(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}})^T X^T W X (\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\phi}}$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is the estimate of $\boldsymbol{\beta}$ obtained when y_i is omitted. (Note that $W/\hat{\phi} = \hat{\Sigma}^{-1}$)

The jack-knife residual and Cook's distance are both useful for detecting potential outliers.

Plotting residuals

When looking at residuals it is helpful to consider them ordered by absolute size.

If we plot the ordered absolute values against the percentage points of a half-normal dsitribution then it is easier to see if the largest values are in keeping with the others.

That is, plot the i-th ordered absolute residual against $\Phi^{-1}((n+i)/(2n+1))$, for $i = 1, \ldots, n$. If all is well we expect to see a smooth plot, while a jump or kink in the tail indicates a potential problem. Note that for a glm the residuals will not in general be normal, so don't expect a straight line.

Checking linearity

A non-linear link g (anything except the identity) makes it harder to check the assumption that $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$.

The easiest thing to do is to plot $g(y_i)$ against $\{x_{ij}\}_{i=1}^n$ for each j and look for linear relationships. A more sophisticated approach is to plot $z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$ against $\{x_{ij}\}_{i=1}^n$, where we use $\hat{\mu}_i$ for μ_i. From the IWLS scheme we know that these plots should be linear.

If there are non-linearities present then we can consider transforming $\{x_{ij}\}_{i=1}^n$ or adding extra variables.
Transforming the responses y_i is often not a good idea for a glm, as this can break assumptions made about the distribution of Y_i.

Example

See Gala.pdf.