3. Exponential families and GLM’s

Do you spend time with your family? Good. Because a man that doesn’t spend time with his family can never be a real man.

Don Corleone (from “The Godfather” by Mario Puzo)
Exponential families

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

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

$\theta$ is the *canonical parameter* (captures location)

$\phi$ is the *dispersion parameter* (captures scale)
Example: normal

\[ Y \sim N(\mu, \sigma^2) \]

\[
f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(y-\mu)^2}{\sigma^2}}
\]

\[
= \exp \left[ \frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2} \left( \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right]
\]

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

where \( \theta = \mu, \phi = \sigma^2 \), and

\[
b(\theta) = \theta^2/2
\]

\[
a(\phi) = \phi
\]

\[
c(y, \phi) = -\frac{1}{2} \left( \frac{y^2}{\phi} + \log(2\pi\phi) \right)
\]

3. Exponential families and GLM's
Example: Poisson

\[ Y \sim \text{pois}(\lambda) \]

\[
f(y) = e^{-\lambda} \frac{\lambda^y}{y!} \quad \text{for } y = 0, 1, 2, \ldots
\]

\[
= \exp \left[ y \log \lambda - \lambda - \log y! \right]
\]

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

where \( \theta = \log \lambda \), \( \phi = 1 \), and

\[
b(\theta) = e^\theta
\]

\[
a(\phi) = \phi
\]

\[
c(y, \phi) = -\log y!
\]
Example: binomial

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

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

where \( \theta = \log \frac{p}{1 - p}, \phi = 1, \text{ and} \)

\[
\begin{align*}
  b(\theta) &= m \log(1 + e^\theta) \\
  a(\phi) &= \phi \\
  c(y, \phi) &= \log \binom{m}{y}
\end{align*}
\]
Example: scaled binomial

\[ X \sim \text{bin}(m, p) \text{ and } Y = X / m \]

\[
f(y) = \binom{m}{my} p^{my} (1 - p)^{m(1-y)} \quad \text{for } y = 0, 1/m, \ldots, 1
\]

\[
= \exp \left[ \frac{y \log \frac{p}{1-p} + \log(1 - p)}{1/m} + \log \binom{m}{my} \right]
\]

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

where \( \theta = \log \frac{p}{1-p} \), \( \phi = 1/m \), and

\[
b(\theta) = \log(1 + e^\theta)
\]

\[
a(\phi) = \phi
\]

\[
c(y, \phi) = \log \binom{m}{my}
\]
Other examples of exponential families are the gamma and the inverse Gaussian.
Lemma: If $Y$ is from an exponential family then

$$\mathbb{E} Y = b'(\theta)$$
$$\text{Var } Y = b''(\theta) a(\phi)$$

The proof is part of Assignment 2.
Exponential family: variance function

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

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

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

**Examples:**

- **normal** $v(\mu) = 1$
- **Poisson** $v(\mu) = \mu$
- **binomial** $v(\mu) = \mu(1 - \mu/m)$
- **scaled binomial** $v(p) = p(1 - p)$
Definition: $Y$ is a GLM if it is from an exponential family, and

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

where

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

Remark: We model \textit{location} using $\eta = x^T \beta$, and let the \textit{scale} sort itself out. That is, we do not model the scale explicitly.
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 \( \theta = \log \frac{p}{1-p} = \log \frac{\mu}{m-\mu}, \ g(\mu) = \log \frac{\mu}{m-\mu} \)

scaled binomial \( \theta = \log \frac{p}{1-p} = \log \frac{\mu}{1-\mu}, \ g(\mu) = \log \frac{\mu}{1-\mu} \)
GLM fitting

We fit GLMs using maximum likelihood.

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

Furthermore suppose that $y_i$ has mean

$$\mu_i = b'(\theta_i) = g^{-1}(x_i^T \beta)$$

If $g$ is the canonical link then $\theta_i = x_i^T \beta$.

The log-likelihood is then

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

Suppose we wish to maximise a log likelihood $l(\theta)$ using Newton’s method. Our update step is

$$\theta(n + 1) = \theta(n) - H(\theta(n))^{-1} \nabla l(\theta(n))$$

where $H(\theta) = (H_{ij}(\theta))$ for

$$H_{ij}(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j}$$

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

If we replace $J$ by $I$, the Fisher information, then the algorithm is called *Fisher scoring*.

The Fisher information is often easier/quicker to calculate, and is guaranteed to be positive definite (unlike the observed information).
It turns out that Fisher scoring applied to a GLM can be interpreted as a least squares problem. Consider

\[
g(Y_i) \approx Z_i := g(\mu_i) + (Y_i - \mu_i)g'(\mu_i) = x_i^T \beta + \epsilon_i\text{ say}
\]

where

\[
\text{Var } \epsilon_i = (g'(\mu_i))^2 \text{Var } Y_i.
\]

If we knew \text{Var } \epsilon_i then the Best Linear Unbiased Estimator (BLUE) of \( \beta \) would be the solution to the weighted least squares problem:

\[
\min_{\beta} (z - X\beta)^T \Sigma^{-1} (z - X\beta)
\]

where \( \Sigma \) is diagonal with \( \Sigma_{ii} = \text{Var } \epsilon_i \).
Aside: Weighted LS from Ordinary LS

Suppose $Y = X\beta + \varepsilon$ where $\varepsilon \sim N(0, \Sigma)$, and $X$ and $\Sigma$ are full rank.

Multiplying by $\Sigma^{-1/2}$ we get $\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \varepsilon'$ where $\varepsilon' \sim N(0, I)$.

From this we see that the BLUE of $\beta$ minimises the sum of squares

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

and the least squares estimator is

$$(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y.$$
We have $g(Y) \approx Z = X\beta + \varepsilon$ where $\varepsilon_i = (Y_i - \mu_i)g'(\mu_i)$ and $\Sigma = \text{Var } \varepsilon$ is diagonal with entries $\Sigma_{ii} = (g'(\mu_i))^2 \text{Var } Y_i = (g'(\mu_i))^2 v(\mu_i) a(\phi)$.

This suggests

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

**Problem:** clearly $z_i$ depends on $\beta$, but $\mu_i = g^{-1}(x_i^T \beta)$, so $\Sigma_{ii}$ also depends on $\beta$.

**Solution:** iterate!

Note that the $a(\phi)$ factor in the expression for $\hat{\beta}$ cancels out.
Iterated Weighted Least Squares (IWLS)

**IWLS algorithm (for estimating $\beta$)**

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

An experiment measuring death rates for insects, with 30 insects at each of five treatment levels.

```r
> library(faraway)
> data(bliss)
> bliss
  dead  alive  conc
1   2      28     0
2   8      22     1
3  15      15     2
4  23       7     3
5  27       3     4
```

We model this with a binomial regression model, and fit using IWLS...
```
# IWLS
y <- bliss$dead
m <- bliss$dead + bliss$alive

mu <- y
eta <- logit(mu/m)
z <- eta + (y - mu)*m/mu/(m - mu)
w <- mu*(m - mu)/m
lmod <- lm(z ~ conc, weights=w, bliss)
coef(lmod)
for (i in 1:5) {
    eta <- lmod$fit
    mu <- m*ilogit(eta)
z <- eta + (y - mu)*m/mu/(m - mu)
w <- mu*(m - mu)/m
lmod <- lm(z ~ conc, weights=w, bliss)
cat(i, coef(lmod), "\n")
}
```
<table>
<thead>
<tr>
<th>Exponential families</th>
<th>GLMs</th>
<th>Weighted LS</th>
<th>Inference</th>
<th>Diagnostics</th>
</tr>
</thead>
</table>

3. Exponential families and GLM's
Theorem

For a GLM, IWLS is equivalent to Fisher scoring
<table>
<thead>
<tr>
<th>Exponential families</th>
<th>GLMs</th>
<th>Weighted LS</th>
<th>Inference</th>
<th>Diagnostics</th>
</tr>
</thead>
</table>

3. Exponential families and GLM's
Variance of $\hat{\beta}$

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

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

where, elementwise, we have

$$\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$$
$$z_i = x_i^T \hat{\beta} + (y_i - \hat{\mu}_i) g'(\hat{\mu}_i)$$
$$\hat{\Sigma}_{ii} = (g'(\hat{\mu}_i))^2 v(\hat{\mu}_i) a(\phi)$$

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

$$\text{Var} \hat{\beta} = [(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}] \hat{\Sigma} [(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1}]^T$$
$$= (X^T \hat{\Sigma}^{-1} X)^{-1}$$
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 $\mu$. Thus $X^2/(n-p)$ will be an estimator for $a(\phi)$.

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

3. Exponential families and GLM’s
From here on we will w.l.o.g. adopt the convention that \( a(\phi) = \phi \).

**Definition:** the *scaled deviance* for model A is

\[
\frac{D^A}{\phi} = -2 \log \frac{\mathcal{L}(\hat{\beta}^A)}{\mathcal{L}(\text{full})}
\]

where \( \hat{\beta}^A \) is the MLE of \( \beta^A \), the true parameter value for model A, and \( \mathcal{L}(\text{full}) \) is the maximum likelihood for the “full” model.

The *deviance* is just \( D^A \).
Example: normal

The full normal model uses $y_i$ to estimate $\mu_i$. 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 full 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 full binomial model uses $y_i$ to estimate $\mu_i$, 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.
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).
For the binomial and Poisson models $\phi = 1$ and the scaled deviance is just the deviance.
For these models the scaled deviance will be approximately $\chi^2$ 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 $\phi$, using $X^2/(n - p)$.
For these models we can’t use the scaled deviance to test model adequacy.
For a linear model A nested within linear model B, under the null hypothesis that model A is correct we have

\[
\frac{(D^A - D^B)/s}{X^2/(n - p)} \sim F_{s, n-p}
\]

where we have \( n \) observations, \( A \) has \( p - s \) parameters and \( B \) has \( p \) parameters.

In R \( X^2 \) (Pearson’s chi-squared) is calculated using the fitted model A.

For other GLM’s this distributional result only holds approximately, but it can still be used for comparing models. In particular it can be used to compare gamma models (though we can also use the AIC).
Akaike Information Criterion (AIC)

For model selection, the AIC generally produces better models than comparing (scaled) deviances, and has the benefit that you don’t have to choose a significance level. The AIC is defined as

\[ \text{AIC} = 2p - 2 \log \mathcal{L}(\hat{\beta}^\star) \]

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

\[ \text{AIC}^B - \text{AIC}^A = 2s - 2 \log \mathcal{L}(\hat{\beta}^{\star B}) + 2 \log \mathcal{L}(\hat{\beta}^{\star A}) \]

\[ = 2s + \frac{D^B}{\phi} - \frac{D^A}{\phi}. \]

Like the log likelihood ratio, the AIC needs an estimate of \( \phi \).
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 \( D = \sum_i d_i = \sum_i r_D(i)^2 \).

For GLMs deviance residuals are often the most useful.
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 $\eta_i = x_i^T \beta$ 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

\[
Z' = \Sigma^{-1/2} Z = \Sigma^{-1/2} X \beta + \epsilon'
\]

where \( \text{Var} \epsilon' = I \). The hat matrix for \( \hat{Z}' \) is the matrix \( H' \) such that \( H'Z' = \hat{Z}' \). From the theory of linear models

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

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

\[
H = X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1}.
\]
The leverage for the $i$-th observation is $H_{ii}$, the $i$-th diagonal element of $H$.

Note that $H$ does not depend on $\phi$ (it cancels out).

The variance of $\hat{Z}$ is

\[
\text{HVar } \hat{Z} H^T = \left[ X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \right] \Sigma \left[ \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T \right] = X(X^T\Sigma^{-1}X)^{-1}X^T = H\Sigma
\]

Whence $\text{Var } (Z - \hat{Z}) = (I - H)\Sigma$. 
Studentised residuals

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

\[
Z_i - \hat{Z}_i = g(\mu_i) + (Y_i - \mu_i)g'(\mu_i) - [g(\mu_i) + (\hat{Y}_i - \mu_i)g'(\mu_i)]
= (Y_i - \hat{Y}_i)g'(\mu_i)
\]

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

\[
\text{Var} \frac{Y_i - \hat{\mu}_i}{\sqrt{(1-H_{ii})v(\hat{\mu}_i)\phi}} \approx 1
\]

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

By analogy we define the \( i \)-th \textbf{studentised deviance residual} to be

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

A large leverage does not necessarily mean a point \textit{has} influenced the fit.

A direct measure of the influence of a point is the \textbf{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

\[
 \text{sign}(y_i - \hat{\mu}_i)\sqrt{(1 - H_{ii})r^2_{SD}(i) + H_{ii}r^2_{SP}(i)}.
\]
Another measure of the influence of the \( i \)-th observation is **Cook’s distance**:

\[
(\hat{\beta}^{(i)} - \hat{\beta})^T X^T \Sigma^{-1} X (\hat{\beta}^{(i)} - \hat{\beta})
\]

\[
\frac{p \hat{\phi}}{p}\]

where \( \hat{\beta}^{(i)} \) is the estimate of \( \beta \) obtained when \( y_i \) is omitted.

Recall that for linear models the Cook’s distance can be expressed in terms of the leverage and the studentised (Pearson) residual, and is large when both of these are large.

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 distribution 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
\[
\Phi^{-1}\left(\frac{n + i}{2n + 1}\right), \text{ for } i = 1, \ldots, n.
\]
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.

**Examples:** bliss.r, gala.r
A non-linear link $g$ (anything except the identity) makes it harder to check the assumption that $g(\mu_i) = x_i^T \beta$.

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 $z_i = g(\hat{\mu}_i) + (y_i - \mu_i)g'(\mu_i)$ against $\{x_{ij}\}_{i=1}^n$, where we use $\hat{\mu}_i$ for $\mu_i$. From the IWLS scheme we know that—provided the predictor variables are independent—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 $Y_i$. 

3. Exponential families and GLM's