1 Densities for generalized linear models
   1.1 Mean and variance .................................................. 2

2 The components of a generalized linear models
   2.1 Linear predictor and link function .............................. 2
   2.2 Variance function .................................................... 3
   2.3 Canonical link .......................................................... 3
   2.4 Example: Bernoulli distribution and canonical link ........ 3

3 Estimation: Iteratively reweighted least squares ................. 4
   3.1 Saturated model, deviance and estimation of \( \phi \) ......... 5
   3.2 Binomial distribution ................................................. 5

4 Ressources .................................................................... 7

5 EXERCISES: Generalized linear models .......................... 7

1 Densities for generalized linear models

Consider densities of the form

\[
f_Y(y; \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi) \right) \tag{1}
\]

where \( \phi \) is a fixed parameter (not part of the model).
1.1 Mean and variance

For random variables with densities of the above form we have:

\[ \begin{align*}
(\ast) \mathbb{E}(y) &= b'(\theta) \\
(\ast\ast) \text{Var}(y) &= b''(\theta)a(\phi)
\end{align*} \tag{2} \]

Notice that \( \frac{\partial}{\partial \theta} f_Y(y; \theta, \phi) = y - b'(\theta) a(\phi) f_Y(y; \theta, \phi) \).

Assuming that the order of integration and differentiation can be interchanged we get:

\[ \begin{align*}
0 &= \frac{d}{d\theta} \int f_Y dy = \int \frac{\partial}{\partial \theta} f_Y dy = \int \frac{y - b'(\theta)}{a(\phi)} f_Y dy \\
&= \frac{1}{a(\phi)} \left[ \int y f_Y dy - b'(\theta) \int f_Y dy \right] = \frac{1}{a(\phi)} [\mu - b'(\theta)]
\end{align*} \tag{3} \]

from which (2) (\ast) follows.

By differentiating twice we get

\[ \begin{align*}
0 &= \frac{d^2}{d\theta^2} \int f_Y dy = \int \frac{\partial}{\partial \theta} \left[ \frac{y - b'(\theta)}{a(\phi)} f_Y \right] dy \\
&= \int \left[ -\frac{b''(\theta)}{a(\phi)} + \left( \frac{y - b'(\theta)}{a(\phi)} \right)^2 \right] f_Y dy
\end{align*} \]

From this (2) (\ast\ast) follows since

\[ \begin{align*}
0 &= \int \left[ \left( \frac{y - \mu}{a(\phi)} \right)^2 - \frac{b''(\theta)}{a(\phi)} \right] f_Y dy \\
&= \frac{1}{a(\phi)^2} \int (y - \mu)^2 f_Y dy - \frac{b''(\theta)}{a(\phi)} \int f_Y dy \\
&= \frac{1}{a(\phi)^2} \text{Var}(Y) - \frac{b''(\theta)}{a(\phi)}
\end{align*} \tag{4} \]

2. The components of a generalized linear models

Consider the situation where \( Y_1, \ldots, Y_n \) are independent random variables and \( Y_i \) has density \( f_Y(y; \theta_i, \phi) \) of the form (1). In the following it is assumed that \( a(\phi) = \phi/w \) where \( w \) is a known weight.

2.1 Linear predictor and link function

To each \( Y \) there is associated a vector \( x \) of explanatory variables and a fixed number \( \omega \). The linear predictor is

\[ \eta = \eta(\beta) = \omega + x^\top \beta. \]
I generalized linear models (GLIMs) the linear predictor is related to the mean $\mu = \mathbb{E}(Y)$ by the link function $g()$ \( \eta = g(\mu) \) as follows

$$\mu = g^{-1}(\eta) = g^{-1}(\omega + x^T \beta)$$

The density (1) is parameterized by the parameter $\theta$ and we therefore need to establish the connections between $\theta$ and $\mu$. From (2) we have that $E(Y) = \mu = b'(\theta)$. Letting the inverse function of $b'$ denoted by $H$. Then

$$\theta = (b')^{-1}(\mu) = H(\mu) = H(g^{-1}(\omega + x^T \beta))$$

which establishes the link between the parameter $\theta$ and $x^T \beta$.

### 2.2 Variance function

From (2) we also have that $\text{Var}(Y) = b''(\theta)a(\phi)$. Since $\theta = H(\mu)$ we have

$$\text{Var}(Y) = b''(H(\mu))a(\phi).$$

The function

$$V(\mu) = b''(H(\mu))$$

is called the variance function as it expresses how the variance depends on the mean.

### 2.3 Canonical link

The connection between $\theta$ and the linear predictor $\eta$ is given by $\theta = (b')^{-1}(g^{-1}(\eta))$. So if $b' = g^{-1}$ then $\theta = \eta$. In this case $g$ is said to be the canonical link function. Using the canonical link function ensures that the log likelihood is concave such it has a unique maximum.

### 2.4 Example: Bernoulli distribution and canonical link

Consider the Bernoulli distribution

$$Pr(Y = y) = p^y(1-p)^{1-y}, \quad y \in \{0,1\}$$

Rewrite as

$$Pr(Y = y) = \exp \left( y \log p + (n - y) \log(1 - p) \right)$$

$$= \exp \left( y \log \frac{p}{1-p} + \log(1 - p) \right)$$

$$= \exp \left( y \theta - \log(1 + \exp(\theta)) \right)$$

where we have introduced the logit as $\theta = g(p) = \log \frac{p}{1-p}$. Notice that $p = \frac{e^\theta}{1+e^\theta}$.

Hence $b(\theta) = \log(1 + e^\theta)$, $w = 1$ and $\phi = 1$. Now $b'(\theta) = \frac{e^\theta}{1+e^\theta} = p$ and $b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = p(1-p)$.  

3
Consider the situation where \( Y_1, \ldots, Y_n \) are independent where the distribution of \( Y_i \) has density of the form (1) with weight \( w_i \). We have

- \( \mathbb{E}(Y_i) = \mu_i \)
- \( \text{Var}(Y_i) = V(\mu_i)\phi/w_i \)
- \( g(\mu_i) = \eta_i = \omega_i + x_i^\top \beta \)

Then the log-likelihood becomes

\[
    l(\beta) = \frac{1}{\phi} \sum_i w_i (y_i \theta_i - b(\theta_i)) - \sum_i c(y_i, \phi)
\]

where \( \theta_i = H(\mu_i) = H(g^{-1}(\eta_i)) = H(g^{-1}(\omega_i + x_i^\top \beta)) \).

Imagine a Taylor expansion of \( g(Y_i) \) around \( \mu_i \):

\[
    g(Y_i) \approx g(\mu_i) + g'(\mu_i)(Y_i - \mu_i) = Z_i
\]

The key to the algorithm is to work with adjusted dependent variables \( z_i \). To actually calculate these we need an initial guess of \( \mu_i \). More about this later.

We have \( \mathbb{E}(Z_i) = g(\mu_i) = \eta_i \) and \( \text{Var}(Z_i) = (g'(\mu_i))^2 V(\mu_i)\phi/w_i \).

In vector form we get

\[
    \mathbb{E}(Z) = \eta = \omega + X \beta, \quad \text{Var}(Z) = \phi \text{diag} \left( (g'(\mu_i))^2 V(\mu_i)/w_i \right) = \phi W(\mu)
\]

The least squares estimate of \( \beta \) is

\[
    \hat{\beta} = (X^\top W^{-1} X)^{-1} X^\top W^{-1} (z - \omega)
\]

Notice: Both \( W \) and \( z \) depend on \( \beta \) so some iterative scheme is needed.

Initialization: Since \( g(\mu_i) = \omega_i + x_i^\top \beta \) we must have \( g(y_i) \approx \omega_i + x_i^\top \beta \). So we may start by applying the link function to data, and obtaining an initial estimate of \( \beta \) as

\[
    \beta^0 = (X^\top X)^{-1} X^\top (g(y) - \omega)
\]

Iteration: Given \( \hat{\beta}^m \) from the \( m \)th iteration we calculate

\[
    z^{m+1} = X \hat{\beta}^m + \text{diag} \left( g'(\mu_i^m) \right) (y - \mu^m) = X \beta^m + r
\]
Update $W$ accordingly as $W = W(\mu^m)$ and calculate
\[
\beta^{m+1} = (X^T W^{-1} X)^{-1} X^T W^{-1}(z^{m+1} - \omega)
\]

Notice that
\[
\beta^{m+1} = \beta^m + (X^T W^{-1} X)^{-1} X^T W^{-1}(r - \omega)
\]
This means that $\alpha = \beta^{m+1} - \beta^m$ is determined as the MLE in the model $r - \omega \sim N(X\alpha, W)$.

3.1 Saturated model, deviance and estimation of $\phi$

Recall that the log–likelihood $l$ has the form $l(\beta) = l(\mu, y) = l_1(\mu, y)/\phi$.
The saturated model is obtained by making no restrictions on $\mu$ so in this case $\hat{\mu} = y$ and $l(y, y)$ is the log–likelihood for the saturated model.
The goodness of fit of a generalized linear model is investigated using the deviance:
\[
D(\hat{\mu}, y) = -2(l(\hat{\mu}, y) - l(y, y)) = -2(l_1(\hat{\mu}, y) - l_1(y, y))/\phi = D_1(\hat{\mu}, y)/\phi
\]
Under certain assumptions on the covariates $X$, it can be shown that $D$ will be asymptotically $\chi^2_{n-p}$ distributed where $p$ is the rank of $X$.

- When $\phi$ is known, this result provides a goodness–of–fit test of the model.
- When $\phi$ is unknown, this results provides a natural estimate of $\phi$ as
\[
\tilde{\phi} = \frac{D_1(\hat{\mu}, y)}{n-p}
\]
(but in this case we do not have a goodness–of–fit test).

3.2 Binomial distribution

Suppose $Y_i \sim Bin(n_i, \mu_i)$ where $g(\mu_i) = \eta_i = x_i^T \beta$ for $i = 1, \ldots, N$.
There are different ways of putting the binomial distribution into the GLIM framework. First notice that $\phi = 1$ and $w_i = 1$ so these terms can be ignored. For later use, let $X = [x_1 : \cdots : x_N]^T$ and $v_i = \{g'(\mu_i)\}^2 \mu_i(1 - \mu_i)$; We may regard $Y_i$ as a sum of $n_i$ independent Bernoulli variables:
\[
Y_i = \sum_{j=1}^{n_i} u_{ij}, \quad u_{ij} \sim Bern(p_i)
\]
We may therefore fit the binomial distribution this way, but this is potentially very inefficient in terms of storage and computing time.

An alternative approach is described below.

In the Bernoulli setup:

- The row vector of covariates \( x_i^\top \) must then be repeated \( n_i \) times: Let \( \tilde{X}_i = 1_{n_i} x_i^\top \), \( \tilde{X} = [\tilde{X}_1 : \cdots : \tilde{X}_N]^\top \). With \( M = \sum_i n_i \), \( \tilde{X} \) is an \( M \times p \) matrix.
- Let also \( \tilde{V}_i = v_i I_{n_i} \); \( \tilde{V} = diag(\tilde{V}_1, \ldots, \tilde{V}_N) \).
- Finally, let \( z_{ij} = g(\mu_i) + g'(\mu_i)(u_{ij} - \mu_i); \tilde{z}_i = (z_{i1}, \ldots, z_{in_i})^\top \) and \( \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_N)^\top \).

We omit the subscripts \( n_i \) indicating dimensions of 1 and \( I \). We then have

\[
\tilde{X}^\top \tilde{V}^{-1} = \begin{bmatrix} 1_{n_1} x_1^\top : \cdots : 1_{n_N} x_N^\top \end{bmatrix}
\]

\[
\tilde{X}^\top \tilde{V}^{-1} \tilde{z} = \frac{1}{v_1} x_1^\top \tilde{z}_1 + \cdots + \frac{1}{v_N} x_N^\top \tilde{z}_N
\]

\[
\tilde{X}^\top \tilde{V}^{-1} \tilde{X} = \begin{bmatrix} n_1 x_1^\top + \cdots + n_M x_N^\top \end{bmatrix}
\]

Recall that \( z_{ij} = g(\mu_i) + g'(\mu_i)(u_{ij} - \mu_i) \). Let

\[
\tilde{z}_i = \frac{1}{n_i} \sum_j z_{ij} = \frac{1}{n_i} 1_{n_i}^\top \tilde{z}_i = g(\mu_i) + g'(\mu_i) \left( \frac{y_i}{n_i} - \mu_i \right)
\]

and let \( \tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_N)^\top \). Then

\[
\mathbb{E}(\tilde{z}_i) = g(\mu_i) \quad \forall \text{Var}(\tilde{z}_i) = \frac{1}{n_i} \{g'(\mu_i)\}^2 \mu_i(1 - \mu_i) = \frac{v_i}{n_i}
\]

Let \( V = diag(\frac{v_1}{n_1}, \ldots, \frac{v_N}{n_N}) \). Then

\[
X^\top V^{-1} = \begin{bmatrix} n_1 x_1^\top : \cdots : n_N x_N^\top \end{bmatrix}
\]

\[
X^\top V^{-1} \tilde{z} = \frac{n_1}{v_1} x_1^\top \tilde{z}_1 + \cdots + \frac{n_N}{v_N} x_N \tilde{z}_N
\]

\[
X^\top V^{-1} X = \begin{bmatrix} n_1 x_1^\top + \cdots + n_N x_N^\top \end{bmatrix}
\]

Notice that \( 1^\top \tilde{z}_i = \sum_j z_{ij} = n_i \tilde{z}_i \). This means that

\[
\tilde{X}^\top \tilde{V}^{-1} \tilde{z} = X^\top V^{-1} \tilde{z}, \quad \tilde{X}^\top \tilde{V}^{-1} \tilde{X} = X^\top V^{-1} X
\]

So iterating using \((\tilde{X}, \tilde{V})\) or \((X, V)\) produces the same result.
4 Ressources

- The standard reference on generalized linear models is McCullagh and Nelder *Generalized Linear Models*


5 EXERCISES: Generalized linear models

The function `model.matrix()` can be used for generating a model matrix $X$ from a given formula and a dataset. So in the following we shall assume that we have a vector of responses $y$, a model matrix $X$ and possibly also a weight vector $w$.

1. Implement your own function `myglm()` which takes $(y, X, w)$ as input. In doing so, you may find it helpful to use that the various families are implemented in R. See e.g. `?binomial`. You may seek inspiration from the function `glm.fit`.

2. The result from your function could be of a certain class, and you could create interesting methods. For example, it would be nice to obtain the approximate covariance matrix of $\hat{\beta}$, which is

   $$(X^TW^{-1}X)^{-1}$$