1 Statistics 249: Categorical Data/GLM

1.1 A reminder of the Linear model:

\[ Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \varepsilon \]

Having defined mean \( \mu \equiv X\beta \) and \( \hat{Y} \equiv X\beta \)

1.2 Generalized GLM

- \( Y \) composed of iid draws from an exponential family.

- \( \mu_{n \times 1} = \mathbb{E}[Y_{n \times 1}] \)

- \( \eta_{n \times 1} = g(\mu_{n \times 1}) = X\beta \)

1.3 A few examples when \( Y \) is discrete

1.3.1 Example: Modern Applied Statistics in Splus, Venalles & Ripley p 218

Results of experiment on a budworm. Giving doses of a chemical to worms, calculating how many live and die (20 worms in each batch):

<table>
<thead>
<tr>
<th>Sex</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

This is the number that died. We could choose \( Y_i \) to be the number that died in a batch of 20, (distributed Binomial(20, \( p_i \))), 12 cells, 12 tests. \( Y_i \) could also be considered the proportion of worms which survived.

A covariate matrix (using log_2 dose covariate):

\[ X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 1 & 4 \\ 1 & 1 & 5 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 2 \\ 1 & 0 & 3 \\ 1 & 0 & 4 \\ 1 & 0 & 5 \end{bmatrix} \]

Sometimes you’ll need to specify the order with a third column: indicator female.
1.4 Exponential Families

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

### 1.4.1 For Normal Distribution

- \( \theta = \mu \)
- \( b(\theta) = \frac{\theta^2}{2} \)
- \( \phi = \sigma^2 \)
- \( c(y, \phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2} \log_e (2\pi\sigma^2) \)

We have shown that:

- \( E[Y] = \mu = b'(\theta) \)
- \( \text{Var}(Y) = b''(\theta)a(\phi) \)

(Reminder, we showed this last time by requiring integral of the density is 1, taking a derivative by \( \frac{\partial}{\partial \theta} \) results in zero, a second derivative results in double zero).

In general \( b''(\theta) \) is called the variance function. And is reexpressed as a function of the mean \( \mu \) so \( b'(\theta) = \mu \) and \( b''(\theta) = V(\mu) \)

### 1.5 Binomial Distribution

For random variables that are discrete and expressed in counts. \( B(n, p) \).

- \( \mu = np \)
- \( \text{Var}(x) = np(1-p) \) so \( V(\mu) = \frac{\mu(\mu-n)}{n} \)
1.6 Poisson distribution $P(\mu)$

Now to consider the amount of worms survived $Y_i = \frac{X_i}{N_i}$ if the numbers are not the same we want to maintain the power in our.

Splus has an option “weights” where you give it proportions and add weights or give numbers and decrease weights.

Looking at the distribution of the proportions. In Binomial

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$  \hspace{1cm} (1)$$

So

$$P\left(Y = \frac{x}{n} = y\right) = \binom{n}{ny} p^{ny} (1 - p)^{n-ny}$$  \hspace{1cm} (2)$$

Exponential family form:

$$P(Y = y) = \exp \left\{ ny \log_e p + n(1 - y) \log_e (1 - p) + \log_e \binom{n}{ny} \right\}$$  \hspace{1cm} (3)$$

So do we now have

1. $\theta = \log_e \frac{p}{1-p}$ so $p = \frac{e^\theta}{1+e^\theta} = b'(\theta)$
2. $b(\theta) = -\log_e (1 - p)$ or $\log_e (1 + e^\theta)$
3. $\phi = n$
4. $a(\phi) = \frac{1}{n} = \frac{1}{\phi}$
5. $c(y, \phi) = \log_e \binom{n}{ny}$

Filling in for this function:

1.6 Poisson distribution $P(\mu)$

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}$$  \hspace{1cm} (4)$$

$$f(y; \mu) = \exp \left\{ -\mu + y \log_e \mu - \log_e y! \right\}$$

1. $\theta = \log_e \mu$
2. $b(\theta) = \mu = \exp(\theta)$
3. $\phi = 1$
4. $\phi = 1$
5. $c(y, \phi) = -\log_e (y!)$

The link we are missing is the canonical link: $g(\mu) = \theta$
1.6.1 Canonical link in glm is \( g(\mu) = \eta = \theta \)

For example, the canonical link for Binomial is \( \frac{x}{1-x} \).

1.7 We’re going to skip the Inverse Gaussian (in the book)

1.8 Gamma \( \alpha, \beta \)

\[
f(y, \alpha, \beta) = \frac{1}{\Gamma(\alpha) \beta^\alpha} y^{\alpha-1} e^{-y/\beta} \text{ for } y > 0 \text{ and } \alpha \beta > 0
\]

Use \( \alpha \beta = \mu \) and \( V(\mu) = \alpha \beta^2 = \frac{\mu^2}{\alpha} \).

\[
\frac{1}{\Gamma(\alpha) \left( \frac{\alpha}{\beta} \right)^\alpha} y^{\alpha-1} e^{-\alpha y/\mu}
\]

\[
f(y, \mu, \alpha) = \exp \left\{ -\frac{\alpha}{\mu} y + (\alpha - 1) \log_e y - \log_e \Gamma(\alpha) - \alpha (\log_e \mu - \log_e \alpha) \right\}
\]

\[
= \exp \left[ -\frac{\alpha}{\mu} y - \alpha \log_e \mu + (\alpha - 1) \log_e y - \log_e \Gamma(\alpha) + \alpha \log_e \alpha \right]
\]

- \( \theta = -\frac{1}{\mu} \)
- \( b(\theta) = \log_e(\mu) = -\log_e(-\theta) \)
- \( \phi = \)
- \( a(\phi) = \frac{1}{\alpha} \)
- \( c(y, \phi) = (\alpha - 1) \log_e y - \log_e \Gamma(\alpha) + \alpha \log_e \alpha \)

1.9 GLM Link and the example

\[
g(\mu_i) = \beta_0 + \text{dose}_i \beta_2 + \beta_3 X_{F_i} + \beta_4 X_{M_i} + \beta_6 (X_X \times \text{dose})
\]