LECTURE 13

Generalized Linear Models – I

Modelling Non-Normal Data
The action of transforming the data can be tried sometimes when the data are skewed. The hope is that:

1. the systematic effects are additive,
2. the random effects are independent of the systematic effects,
3. the resulting errors are normal with constant variance.
The last feature may not always be possible, e.g., for binary data, when the response is a *dichotomous* variable, i.e., having only two possible outcomes, say survival or death, working or broken etc.
Types of possible data by error distribution:

- Normal
- Binary - (0,1) data
- Binomial - eg proportions, x out n
- Poisson - eg counts (and contingency tables)
- Gamma - eg failure times
- Weibull and Negative Binomial
- Inverse Gaussian
Non-Normal Data:
- (Binary)

Minor faults occur irregularly in an industrial process and, as an aid to their diagnosis, an experiment was performed. Batches of raw material were selected and each batch was divided into two equal sections: for each batch, one of the sections was processed by the standard method and the other by a modified process. Before processing, a purity index was measured for the whole batch of material. For the product from each section of material it was recorded whether the minor faults did or did not occur. Results for 22 batches are given in the table.
<table>
<thead>
<tr>
<th>Purity index</th>
<th>Standard process</th>
<th>Modified process</th>
<th>Purity index</th>
<th>Standard process</th>
<th>Modified process</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.2</td>
<td>NF</td>
<td>NF</td>
<td>6.5</td>
<td>NF</td>
<td>F</td>
</tr>
<tr>
<td>6.3</td>
<td>F</td>
<td>NF</td>
<td>4.9</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>8.5</td>
<td>F</td>
<td>NF</td>
<td>5.3</td>
<td>F</td>
<td>NF</td>
</tr>
<tr>
<td>7.1</td>
<td>NF</td>
<td>F</td>
<td>7.1</td>
<td>NF</td>
<td>F</td>
</tr>
<tr>
<td>8.2</td>
<td>F</td>
<td>NF</td>
<td>8.4</td>
<td>F</td>
<td>NF</td>
</tr>
<tr>
<td>4.6</td>
<td>F</td>
<td>NF</td>
<td>8.5</td>
<td>NF</td>
<td>F</td>
</tr>
<tr>
<td>8.5</td>
<td>NF</td>
<td>NF</td>
<td>6.6</td>
<td>F</td>
<td>NF</td>
</tr>
<tr>
<td>6.9</td>
<td>F</td>
<td>F</td>
<td>9.1</td>
<td>NF</td>
<td>NF</td>
</tr>
<tr>
<td>8.0</td>
<td>NF</td>
<td>NF</td>
<td>7.1</td>
<td>F</td>
<td>NF</td>
</tr>
<tr>
<td>8.0</td>
<td>F</td>
<td>NF</td>
<td>7.5</td>
<td>NF</td>
<td>F</td>
</tr>
<tr>
<td>9.1</td>
<td>NF</td>
<td>NF</td>
<td>8.3</td>
<td>NF</td>
<td>NF</td>
</tr>
</tbody>
</table>

F, Faults occur  
NF, No Faults occur
Thus the response is a binary *categorical* variable, (F,NF).
(Binomial)

The table summarises a two-factor industrial investigation in which the number of ingots not ready for rolling out of those tested is shown for combinations of heating time and soaking time.

<table>
<thead>
<tr>
<th>Soaking time</th>
<th>Heating time</th>
<th>7</th>
<th>14</th>
<th>27</th>
<th>51</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.10</td>
<td>0.31</td>
<td>1.56</td>
<td>3.13</td>
<td>4.110</td>
<td></td>
</tr>
<tr>
<td>1.7</td>
<td>0.17</td>
<td>0.43</td>
<td>4.44</td>
<td>0.1</td>
<td>4.105</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>0.7</td>
<td>2.33</td>
<td>0.21</td>
<td>0.1</td>
<td>2.62</td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>0.12</td>
<td>0.31</td>
<td>1.22</td>
<td>0.0</td>
<td>1.65</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>0.9</td>
<td>0.19</td>
<td>1.16</td>
<td>0.1</td>
<td>1.45</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.55</td>
<td>2.157</td>
<td>7.159</td>
<td>3.16</td>
<td>12.387</td>
<td></td>
</tr>
</tbody>
</table>

Entry = (number not ready, number tested)
So the response variable is $x$ ready, out of $n$ tested.
• (Poisson)

The table gives the number of faults in rolls of textile fabric. The distribution of the number of faults is of interest, especially in its relation to that expected if faults occur at random at a fixed rate per metre.
<table>
<thead>
<tr>
<th>Roll No.</th>
<th>Roll length (metres)</th>
<th>No. of faults</th>
<th>Roll No.</th>
<th>Roll length (metres)</th>
<th>No. of faults</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>551</td>
<td>6</td>
<td>17</td>
<td>543</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>651</td>
<td>4</td>
<td>18</td>
<td>842</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>832</td>
<td>17</td>
<td>19</td>
<td>905</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>375</td>
<td>9</td>
<td>20</td>
<td>542</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>715</td>
<td>14</td>
<td>21</td>
<td>522</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>868</td>
<td>8</td>
<td>22</td>
<td>122</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>271</td>
<td>5</td>
<td>23</td>
<td>657</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>630</td>
<td>7</td>
<td>24</td>
<td>170</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>491</td>
<td>7</td>
<td>25</td>
<td>738</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>372</td>
<td>7</td>
<td>26</td>
<td>371</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>645</td>
<td>6</td>
<td>27</td>
<td>735</td>
<td>17</td>
</tr>
<tr>
<td>12</td>
<td>441</td>
<td>8</td>
<td>28</td>
<td>749</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>895</td>
<td>28</td>
<td>29</td>
<td>495</td>
<td>7</td>
</tr>
<tr>
<td>14</td>
<td>458</td>
<td>4</td>
<td>30</td>
<td>716</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>642</td>
<td>10</td>
<td>31</td>
<td>952</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>492</td>
<td>4</td>
<td>32</td>
<td>417</td>
<td>2</td>
</tr>
</tbody>
</table>
So the response is a count, the number of faults, with no upper limit, in theory.
(Gamma)

The data in the table give the clotting times (secs) of blood for normal plasma diluted to nine different percentage concentrations with prothrombin–free plasma; clotting was induced by two lots of thromboplastin.
<table>
<thead>
<tr>
<th>Concentration</th>
<th>Clotting time</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lot 1</td>
<td>Lot 2</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>118</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>58</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>42</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>35</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>27</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>25</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>21</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>19</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>18</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

So the response is continuous and non zero.
Definition

A linear model is defined by

\[ Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]
A GLM is defined by

\[ Y = \mu + \epsilon, \quad \epsilon \text{ not necessarily normal} \]
\[ E(Y) = \mu \quad \text{systematic component} \]
\[ \eta = X\beta \quad \text{linear predictor} \]
\[ \mu = g^{-1}(\eta) \quad \text{the inverse function} \]
\[ y = g^{-1}(\eta) + \epsilon(\mu) = g^{-1}(X\beta) + \epsilon \]

So the variance can depend on the mean, in general.
Thus the function fitted to the data is always a function of a linear model, hence the term *generalized* linear model.

The function $g$ is called the *link* function, so named because

$$g(\mu) = \eta = X\beta$$
Deviance

The generalized form of discrepancy (between data and fitted) function is called the *Deviance*. The forms of the deviance are:
<table>
<thead>
<tr>
<th>Error</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>$\sum (y - \hat{\mu})^2$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$2 \sum [y \ln(y/\hat{\mu}) - (y - \hat{\mu})]$</td>
</tr>
<tr>
<td>Binomial</td>
<td>$2 \sum \left[ y \ln(y/\hat{\mu}) + (m - y) \ln \left( \frac{m-y}{m-\hat{\mu}} \right) \right]$</td>
</tr>
</tbody>
</table>

If $y \approx \hat{\mu}$ the deviance is small, else it is large.
Binary or Bernoulli Data

The response $Y$ is dichotomous, ie, $(0,1)$.

Thus we have \textit{two} possible values only.

The (probability) model is

$$P(Y = 1) = \pi, \quad P(Y = 0) = 1 - \pi$$

and the deterministic component then models $\pi$.

This is the most severe departure from Normality!

Now

$$Y = \mu + \epsilon$$
with, in general,

\[ P(Y = y) = \pi^y (1 - \pi)^{1-y}, \; y = 0, 1 \]

giving

\[ \epsilon = 0 - \mu \text{ or } 1 - \mu \]

and so \( \epsilon \) is \textit{discrete}.

Now

\[ E(Y) = 0P(Y = 0) + 1P(Y = 1) = \pi = \mu \]

and

\[ V(Y) = \pi(1 - \pi) \]
Link Function

A common link function used in binary data is the logit;

$$\ln\left(\frac{\mu}{1 - \mu}\right) = \eta = X\beta$$

or in terms of the mean (proportion)

$$\mu = \frac{e^\eta}{1 + e^\eta} = \pi$$
Example

A hypothetical situation shows the connection between the linear model of the $\eta$ scale and the scale of the mean $\mu$ or scale of proportions $\pi$. Now the linear model (predictor) is given by

$$\eta_i = 10 + 5x_i, \ -6 < x_i < 4$$
while

$$\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \pi_i$$

is constrained to lie between 0 and 1 as shown in the graph
> x <- 1:10
> x <- x-6
> lp <- 10 + 5 * x

  linear predictor on the logit scale
> plot(lp~x,type="b")

  the scale of proportions
> p <- exp(lp)/(1+exp(lp))
> plot(p~x,type="b")
Example of Binary Data

The data were obtained in a controlled study of the effect of rate and volume of air inspired by humans on the occurrence (1) or non-occurrence (0) of a transient vasoconstriction response in the skin of the fingers, a proxy measure. We wish to use rate and volume to model the vasoconstriction response. The following data plot shows the potential relation between the response and the predictors:
1 = vaso–constriction, 0 = no vaso–constriction

---

**Diagram:**

- **X-axis:** vol
- **Y-axis:** rate

- Points marked with 1 represent vaso-constriction.
- Points marked with 0 represent no vaso-constriction.

---

**Note:**

The diagram illustrates the relationship between vol and rate, with values indicating whether vaso-constriction occurs (1) or not (0) at various vol levels.
As expected, both predictors contribute to explaining the occurrence of the vasoconstriction, since the reduction in deviance on 2 df (= 38-36) is 24.268 (= 54.04-29.772), which is well above what might be expected for random variation.
> fdatdf <- read.table("finney1.txt",header=T)
> str(fdatdf)
'data.frame': 39 obs. of 3 variables:
$ vol : num 3.7 3.5 1.25 0.75 0.8 0.7 0.6 1.1 0.9 0.9 ...
$ rate: num 0.825 1.09 2.5 1.5 3.2 3.5 0.75 1.7 0.75 0.45 ...
$ resp: int 1 1 1 1 1 1 1 0 0 0 0 ...
> attach(fdatdf)
> f1.glm <- glm(resp ~ rate + vol,family=binomial)
> summary(f1.glm)

Call:
glm(formula = resp ~ rate + vol, family = binomial)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.50657  -0.73465   0.03997   0.48854   2.32935
Coefficients:

|             | Estimate | Std. Error | z value | Pr(>|z|)  |
|-------------|----------|------------|---------|-----------|
| (Intercept) | -9.5296  | 3.2276     | -2.953  | 0.00315 **|
| rate        | 2.6491   | 0.9129     | 2.902   | 0.00371 **|
| vol         | 3.8821   | 1.4262     | 2.722   | 0.00649 **|

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 54.040 on 38 degrees of freedom
Residual deviance: 29.772 on 36 degrees of freedom
AIC: 35.772

Number of Fisher Scoring iterations: 5

> tresp <- as.character(resp)
> plot(rate~vol, pch=tresp, main="1 = vaso-constriction,"
The utility of the predictors is verified by a plot of fitted values.
similar to the original data plot:
a = (p above 0.5), b = (p below 0.5)
A proper plot of the fitted surface against the two predictors, shows that we are fitting a 'cliff' to the data, as expected since the data are dichotomous.
The binary nature of the residuals is shown in the plot of the Pearson residuals vs the fitted values (proportions).
The curvature in the residuals is due to their divisor being \( \sqrt{\mu(1 - \mu)} \).
Binary data take the values 0, 1

\[ Pr(Y_i = 0) = 1 - p_i; \quad Pr(Y_i = 1) = p_i \]

<table>
<thead>
<tr>
<th>Subject</th>
<th>((x_1, x_2))</th>
<th>(Y)</th>
<th>Listed by Subject Number</th>
<th>((x_1, x_2))</th>
<th>(m_i)</th>
<th>(Y)</th>
<th>Listed by Covariate Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,1</td>
<td>0</td>
<td></td>
<td>1,1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1,2</td>
<td>1</td>
<td></td>
<td>1,2</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1,2</td>
<td>0</td>
<td></td>
<td>2,1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2,1</td>
<td>0</td>
<td></td>
<td>2,2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2,2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1,2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1,1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In R, binary data are entered as a vector of 0s and 1s, while binomial data are entered as a two column matrix holding successes and failures.
Example of Binomial Data

The data presented relate to a dosage mortality problem, where increasing dosage levels were applied and deaths out of a total exposed to the drug were recorded.

<table>
<thead>
<tr>
<th>dose</th>
<th>killed</th>
<th>out of</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.57</td>
<td>37</td>
<td>132</td>
</tr>
<tr>
<td>1.17</td>
<td>40</td>
<td>51</td>
</tr>
<tr>
<td>1.49</td>
<td>114</td>
<td>127</td>
</tr>
<tr>
<td>1.66</td>
<td>115</td>
<td>117</td>
</tr>
<tr>
<td>1.79</td>
<td>125</td>
<td>125</td>
</tr>
</tbody>
</table>

Is mortality related to dosage?
Two models are fitted, (i) the constant probability model, where mortality is unrelated to dose, and (ii) the glm where mortality is related to a linear model on the logit scale.

A plot of model (i), on the scale of proportions killed:
A plot of model (ii), on the scale of proportions killed:
So we expect model (ii) to be preferred.
The Analysis of Deviance Table (via \texttt{anova.glm}) shows the change in deviance between model (i) on 4 df and model (ii) on 3 df is 257.5 on 1 df, indicating a clear choice of model (ii). Thus mortality does appear to be related to dosage.
> fdosedf <- read.table("fdose.txt",header=T)
> str(fdosedf)
'data.frame': 5 obs. of 3 variables:
$ dose : num 0.57 1.17 1.49 1.66 1.79
$ killed: int 37 40 114 115 125
$ outof : int 132 51 127 117 125
> attach(fdosedf)
> propn <- killed/outof
> plot(propn~dose,pch="d")
> fdosedf$Y <- cbind(killed, outof-killed)
> attach(fdosedf)
> Y

     killed
[1,]    37   95
[2,]    40    11
[3,]   114   13
> glm(Y ~ 1, family=binomial, data=fdosedf)

Call: glm(formula = Y ~ 1, family = binomial, data = fdosedf)

Coefficients:
(Intercept)
1.270

Degrees of Freedom: 4 Total (i.e. Null); 4 Residual
Null Deviance: 266.7
Residual Deviance: 266.7  AIC: 284.7

> g1 <- glm(Y ~ 1, family=binomial, data=fdoseddf)
> g1$fitted

[4,]  115  2
[5,]  125  0
> points(dose,g1$fitted,pch="f",type="l")
> glm(Y ~ dose, family=binomial, data=fdosedf)

Call:  glm(formula = Y ~ dose, family = binomial, data = fdosedf)

Coefficients:
  (Intercept)    dose
    -3.286        4.010

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:     266.7
Residual Deviance:  9.125   AIC:  29.16
> g2 <- glm(Y ~ dose, family=binomial, data=fdosedef)
> g2$fitted
> plot(propn~dose,pch="d")
> points(dose,g2$fitted,pch="f",type="l")
> anova.glm(g2)

Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance Resid.</th>
<th>Df Resid.</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>4</td>
<td>266.670</td>
<td></td>
</tr>
<tr>
<td>dose</td>
<td>1</td>
<td>257.545</td>
<td>3</td>
</tr>
</tbody>
</table>

> summary.glm(g2)
Call:
```r
glm(formula = Y ~ dose, family = binomial,
     data = fdosedf)
```

Deviance Residuals:
[1]  0.2927 -0.3344 -1.6532  1.0707  2.2468

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -3.2856 | 0.3384 | -9.71 | <2e-16 *** |
| dose     | 4.0100  | 0.3169 | 12.65 | <2e-16 *** |

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 266.6699  on 4  degrees of freedom
Residual deviance: 9.1254  on 3  degrees of freedom
AIC: 29.163

Number of Fisher Scoring iterations: 4
To verify the last fitted value for the second model,

$$\eta_5 = -3.286 + 4.01(1.79) = 3.8919 = \ln \left( \frac{\pi_5}{1 - \pi_5} \right)$$

So

$$\pi_5 = 1/(1 + e^{-\eta_5}) = 0.98$$

giving

$$\mu_5 = n_5 \pi_5 = 125(0.98) = 122.5$$
Binary/Binomial Data

This example shows the connection between binary and binomial data.

The data in the table give the number of seedlings that nodulated under given treatment regimes.

<table>
<thead>
<tr>
<th>Inoculation level</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Plot 1</td>
<td>Plot 2</td>
</tr>
<tr>
<td>Seedlings/plot (n)</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>No. nodulating (r)</td>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>
Two models are fitted (i) the constant probability model, and (ii) the model where the nodulation is allowed to vary across treatment regimes.

This is performed using both the binomial and binary formulation of the problem.
The binomial form is undertaken at the *plot* level, while the binary form fits a model at the *seedling* level. Note the difference in the df of the null model for the two formulations,

\[ 3 = 4 - 1 \text{ vs } 119 = 20 + 30 + 40 + 30 - 1 ! \]
Binomial form:
Model (i)
hollow circle=data, solid circle=fit
and Model (ii)
hollow circle=data, solid square=fit
Model (ii) should be chosen.

Model (i) gives a deviance 17.57 on 3 df, while model (ii) gives 3.914 on 2 df. Thus the changes in deviance is 13.65 on 1 df.

Binary form:

Model (i) gives a deviance of 155.4 on 119 df, while model (ii) gives a deviance of 141.7 on 118 df, giving a change in deviance of 13.7 on 1 df

Thus the two forms are equivalent.
> cahdf <- read.table("candh.txt",header=T)
> str(cahdf)
‘data.frame’: 4 obs. of 3 variables:
  $ tmt : int 1 1 2 2
  $ nod : int 7 16 34 21
  $ seeds: int 20 30 40 30
> attach(cahdf)
> treat <- tmt
> cahdf$tmt <- factor(cahdf$tmt)
> cahdf$Y <- cbind(nod,seeds-nod)
> attach(cahdf)
> str(cahdf)
‘data.frame’: 4 obs. of 4 variables:
  $ tmt : Factor w/ 2 levels "1","2": 1 1 2 2
  $ nod : int 7 16 34 21
  $ seeds: int 20 30 40 30
  $ Y : int [1:4, 1:2] 7 16 34 21 13 14 6 9
  ..- attr(*, "dimnames")=List of 2
\[ \text{pr} \leftarrow \text{nod/seeds} \]
\[ \text{plot(pr} \sim \text{treat)} \]
\[ \text{glm(Y} \sim 1, \text{data=cahdf, family=binomial)} \]

Call: glm(formula = Y ~ 1, family = binomial, data = cahdf)

Coefficients:
(Intercept)
  0.619

Degrees of Freedom: 3 Total (i.e. Null); 3 Residual
Null Deviance: 17.57
Residual Deviance: 17.57  AIC: 34.01
\[ \text{g1} \leftarrow \text{glm(Y} \sim 1, \text{data=cahdf, family=binomial)} \]
\[ \text{fp} \leftarrow \text{g1$\text{fitted}} \]
> points(treat,fp,type="p", pch=16)
> glm(Y~tmt,data=cahdf,family=binomial)

Call:  glm(formula = Y ~ tmt, family = binomial,
data = cahdf)

Coefficients:
(Intercept)     tmt2
  -0.1603       1.4596

Degrees of Freedom:  3 Total (i.e. Null);  2 Residual
Null Deviance:    17.57
Residual Deviance:  3.914  AIC:  22.35
> g2 <- glm(Y~tmt,data=cahdf,family=binomial)
> fp2 <- g2$fitted
> plot(pr~treat)
> points(treat,fp2,type="p", pch=15)
> anova.glm(g2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>3</td>
<td>17.5653</td>
<td></td>
</tr>
<tr>
<td>tmt</td>
<td>1</td>
<td>13.6516</td>
<td>2</td>
</tr>
</tbody>
</table>

```r
> cahdf$NODS <- c(rep(1,7),rep(0,13),rep(1,16),rep(0,14),
+ rep(1,34),rep(0,6),rep(1,21),rep(0,9))
> cahdf$LEVELS <- c(rep("lo",50),rep("hi",70))
> cahdf$LEVELS <- factor(cahdf$LEVELS)
> attach(cahdf)
> str(cahdf)
```
'data.frame': 4 obs. of 6 variables:
  $ tmt    : Factor w/ 2 levels "1","2": 1 1 2 2
  $ nod    : int 7 16 34 21
  $ seeds  : int 20 30 40 30
  $ Y      : int [1:4, 1:2] 7 16 34 21 13 14 6 9
     ..- attr(*, "dimnames")=List of 2
        ..$ : NULL
        ..$ : chr "nod"
  $ NODS   : num 1 1 1 1 1 1 1 0 0 0 ...
  $ LEVELS: Factor w/ 2 levels "hi","lo": 2 2 2 2 2 2 2 2 2 2 ...

> glm(NODS~1,data=cahdf,family=binomial)

Call: glm(formula = NODS ~ 1, family = binomial,  
           data = cahdf)

Coefficients:
(Intercept)    
       0.619
Degrees of Freedom: 119 Total (i.e. Null); 119 Residual
Null Deviance: 155.4
Residual Deviance: 155.4  AIC: 157.4

> glm(NODS~LEVELS,data=cahdf,family=binomial)

Call: glm(formula = NODS ~ LEVELS, family = binomial, data = cahdf)

Coefficients:
(Intercept)  LEVELSlo
  1.299       -1.460

Degrees of Freedom: 119 Total (i.e. Null); 118 Residual
Null Deviance: 155.4
Residual Deviance: 141.7  AIC: 145.7
Poisson Data

The data are discrete (counts), and the mean is equal to the variance. The model is

\[ Y = \mu + \epsilon \]

where

\[ Y \sim P(\mu). \]

So

\[ E(Y) = \mu \]

and a common link function used is log giving

\[ \log(\mu) = X\beta = \text{linear model} \]
Example 1

Consider the problem given earlier about faults in rolls of fabric.

The data are discrete, from the original description. Also the stem and leaf plot shows the discreteness.

The decimal point is 1 digit(s) to the right of the |  

0 | 1234444  
0 | 5666777888999999  
1 | 0044  
1 | 77  
1 | 77  
2 | 3  
2 | 8
A histogram of the data:
A plot of the data shows the variance increasing with the mean:
Faults vs length of roll

length

faults

0 5 10 15 20 25

200 400 600 800

length

648
The first model uses the length of the roll to predict the number of faults:

```r
> gp1 <- glm(faults ~ length, family = poisson)
> summary.glm(gp1)
```

Call:
glm(formula = faults ~ length, family = poisson)

Deviance Residuals:
```
          Min           1Q       Median           3Q          Max
-2.74127   -1.13312   -0.03904    0.66179    3.07446
```

Coefficients:
```
                Estimate Std. Error  z value Pr(>|z|)    
(Intercept) 0.9717506   0.2124693   4.574  4.79e-06 ***
length      0.0019297   0.0003063   6.300  2.97e-10 ***
```

649
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 103.714  on 31  degrees of freedom
Residual deviance:  61.758  on 30  degrees of freedom
AIC: 189.06
The change in deviance is $103.714 - 61.758 = 41.956$ in 1 df ($31 - 30$).

The over dispersion is $61.758/30 = 2.06$, and since $41/2$ is approx 20, we can still claim that length of the roll is a useful predictor of the number of faults in the roll, as expected ($\chi^2_{1,5\%} = 3.84$).
The histogram of residuals and plot of residuals vs length for this model:
An alternative formulation is to correct the number of faults for the length of the roll using an offset. Now is the glm is

$$\log(\text{counts}) = \text{linear model}$$

use

$$\log(\text{counts/length}) = \text{linear model}$$

to give

$$\log(\text{counts}) = \log(\text{length}) + \text{linear model}$$

The first term on the RHS is called the offset. It can be thought of as a correction to the response variable.
> ll <- log(length)
> gp2 <- glm(faults ~ 1 + offset(ll), family=poisson)
> summary(gp2)
Call:
glm(formula = faults ~ 1 + offset(ll), family = poisson)

Deviance Residuals:
            Min       1Q   Median       3Q      Max
-2.8167 -1.1265 -0.2367  0.6274  3.4376

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.19290  0.05934   -70.66   <2e-16 ***

(Dispersion parameter for poisson family taken to be 1)
Null deviance: 64.537  on 31  degrees of freedom
Residual deviance: 64.537  on 31  degrees of freedom
AIC: 189.84
The deviance (64.5) for this model is similar to the deviance (61.7) for the first model.
The histogram of residuals and plot of residuals vs length for this model:
Histogram of Residuals: offset model

Frequency

gp2$residuals

-1.0 -0.5 0.0 0.5 1.0 1.5

0 5 10 15
Residuals of offset model

length

gp2$residuals

200 400 600 800

-0.5 0.0 0.5 1.0 1.5

0.0

-0.5

660
Notice the similarity of these results to those of the first model.

The benefit of the offset is that we have saved 1 df, by not using length as a predictor.
> fpdf <- read.table("fault.txt",header=T)
> str(fpdf)
‘data.frame’: 32 obs. of 3 variables:
$ roll : int 1 17 2 18 3 19 4 20 5 21 ...
$ length: int 551 543 651 842 832 905 375 542 715 522 ...
$ faults: int 6 8 4 9 17 23 9 9 14 6 ...
> attach(fpdf)
> hist(faults)
> stem(faults)

The decimal point is 1 digit(s) to the right of the |
> plot(length, faults)
> gp1 <- glm(faults ~ length, family = poisson)
> plot(length, gp1$residuals)
> hist(gp1$residuals)
> summary.glm(gp1)

Call:
glm(formula = faults ~ length, family = poisson)

Deviance Residuals:

Min 1Q Median 3Q Max
-2.74127 -1.13312 -0.03904 0.66179 3.07446

Coefficients:

                           Estimate Std. Error   z value  Pr(>|z|)
(Intercept)               0.9717506   0.2124693  4.574 4.79e-06 ***
length                    0.0019297   0.0003063  6.300 2.97e-10 ***
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 103.714 on 31 degrees of freedom
Residual deviance: 61.758 on 30 degrees of freedom
AIC: 189.06

Number of Fisher Scoring iterations: 4

> ll <- log(length)
> gp2 <- glm(faults ~ 1 + offset(ll), family=poisson)
> summary(gp2)

Call:
glm(formula = faults ~ 1 + offset(ll), family = poisson)

Deviance Residuals:
          Min        1Q  Median        3Q       Max

-2.8167  -1.1265  -0.2367   0.6274   3.4376

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -4.19290   | 0.05934 | -70.66   | <2e-16 *** |

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 64.537 on 31 degrees of freedom
Residual deviance: 64.537 on 31 degrees of freedom
AIC: 189.84

Number of Fisher Scoring iterations: 4

> plot(length,gp2$residuals)
> hist(gp2$residuals)
Example 2

Light traps catch a small insect (midge), but the sampling intervals are uneven, due to the worker having to sleep!
<table>
<thead>
<tr>
<th>Hours after sundown</th>
<th>interval length</th>
<th>Counts in April</th>
<th>Counts in May</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>140</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>93</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

The unequal sampling needs to be taken into account. This can be done via an *offset* as shown in the previous example.
Basically, we correct the counts back to the same unit of measure, ie, 1 hour of sampling.
\[ E(y) = \mu \]
\[ \log(\mu) = X\beta \]
\[ \log\left(\frac{\text{counts}}{\text{hours}}\right) = X\beta \]
\[ \log(\text{counts}) = X\beta + \log(\text{hours}) \]
hours_c(1:6,12)
April_c(140,93,26,10,5,0,14)
May_c(24,9,3,4,5,0,0,1)
interval_c(rep(1,6),6)
Linterval_log(interval)
midge.mod_glm(April ~ hours + hours^2 + offset(Linterval),
            family=poisson)
print(anova(midge.mod,test="Chisq"))

Apr.preds_predict.glm(midge.mod,type="response",se.fit=T)
Apr.fit_Apr.preds$fit
Apr.ll_Apr.fit - 2*Apr.preds$se.fit
Apr.ul_Apr.fit + 2*Apr.preds$se.fit #$

plot(hours,April,pch=1,xlab="hours since sundown",
     ylab="midge counts",las=1)
lines(hours,Apr.fit,lty=1)
lines(hours,Apr.ll,lty=2)
lines(hours,Apr.ul,lty=2)
hours since sundown

midge counts
```r
> hours <- c(1:6,12)
> midges <- c(140,93,26,10,5,0,14)
> interval <- c(rep(1,6),6)
> midge.mod <- glm(midges ~ hours + offset(log(interval)),
                   family=poisson)
> midge.fit <- predict(midge.mod,type="response")
> print(cbind(hours,interval,midges,midge.fit))

   hours interval midges midge.fit
1      1       1    140     124.668010
2      2       1    93      72.432681
3      3       1    26      42.083718
4      4       1    10      24.450831
5      5       1     5     14.206044
6      6       1     0      8.253776
7     12       6    14      1.904939

> anova(midge.mod,test="Chisq")
Analysis of Deviance Table
```
Model: poisson, link: log

Response: midges

Terms added sequentially (first to last)

| Df | Deviance Resid. | Df Resid. | Dev | P(>|Chi|) |
|----|-----------------|-----------|-----|---------|
| NULL | 6 | 651.46 |
| hours | 1 | 570.01 | 5 | 81.45 | 5.597e-126 |

> summary(midge.mod)$coefficients

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| (Intercept) | 5.3686510 | 0.10217573 | 52.54331 | 0.0000000e+00 |
| hours | -0.5429967 | 0.03773791 | -14.38863 | 6.099575e-47 |

> midge.mod$deviance

[1] 81.44965

> midge.mod$null.deviance

[1] 651.4558
> midge.mod$df.null

[1] 6
To verify the last fitted value:

$\ln\left(\frac{\text{counts}}{6}\right) = 5.368651 - 0.5429967(12) = -1.147309$

$\frac{\text{counts}}{6} = 0.317490 \rightarrow \text{counts} = 1.9049$

The other fitted values use $\text{counts}/1$. 
Generalised Additive Models

The previous model used a quadratic model in time, but this may be too restrictive in practice. The GAM allows for general models such as splines etc to be used in place of the standard linear model.

```r
library(mgcv)
midge.gam <- gam(April ~ s(hours) + offset(log(interval)),
    family=poisson)

print(midge.gam)
plot(midge.gam)
```
Note that the scale of measure on the y axis is log counts . . . , so the apparent blow out in the confidence band around 8 hours is not so great, since it is from $e^{-4} = 0.01$ to $e^{-1} = 0.36$ on the scale of counts. By comparison the width of the confidence band at the zero end of the time scale is about 6 on the scale of counts.
> hours <- c(1:6,12)
> midges <- c(140,93,26,10,5,0,14)
> interval <- c(rep(1,6),6)
> midge.gam <- gam(midges~s(hours) + offset(log(interval)),
                   family=poisson)

> print(midge.gam)

Family: poisson
Link function: log

Formula:
midges ~ s(hours) + offset(log(interval))

Estimated degrees of freedom:
  4.106186   total =  5.106186

UBRE score:   0.7843005
> summary(midge.gam)

Family: poisson
Link function: log
Formula:
midges ~ s(hours) + offset(log(interval))

Parametric coefficients:

| Estimate | std. err. | t ratio | Pr(>|t|) |
|----------|-----------|---------|----------|
| (Intercept) | 2.4857    | 0.1519  | 16.37    | 0.0037127 |

Approximate significance of smooth terms:

<table>
<thead>
<tr>
<th>edf</th>
<th>chi.sq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s(hours)</td>
<td>4.106</td>
<td>350.69</td>
</tr>
</tbody>
</table>

Adjusted r-sq. = 0.995  
UBRE score = 0.7843

Scale estimate = 1  
n = 7

> plot(midge.gam)
> midge.gam$fitted.values
[1] 141.684523 89.080468 28.734126 9.855015
  3.549326 1.228838 13.867703
> midge.gam$residuals
[1] -0.01188923 0.04400019 -0.09515161
  0.01471181 0.40871898 -0.99994720
[7]  0.00953992
> midge.gam$deviance
[1]  3.44543
> midge.gam$null.deviance
[1] 391.1631
As expected, the GAM (4 df?) fits the data far better than the quadratic, with a deviance of 3.4 vs 81.4, and the fitted values clearly track the data better than the first model.
<table>
<thead>
<tr>
<th>midges</th>
<th>GLM</th>
<th>GAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>124</td>
<td>141</td>
</tr>
<tr>
<td>93</td>
<td>72</td>
<td>89</td>
</tr>
<tr>
<td>26</td>
<td>42</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>14</td>
</tr>
</tbody>
</table>
Gamma variables

Gamma data are continuous, non zero with constant coefficient of variation (CV).

\[ CV = \frac{\sigma}{\mu} \times \frac{100}{1} = 100 \frac{sd}{mean} \]
Example 3

Elbow flux data:
The variance increases with the mean, and so since the data are continuous and non zero, a gamma model is suggested.
Example 4

Accelerated test data:

The response is breakdown strength and the predictors are time in weeks and temperature in degrees C.
dialectr <- read.table("dialectr.txt",header=T)
gmod <- glm(Strength ~ Temperature*Time,family=Gamma(link=log),
             data=dialectr)
The analysis of deviance table agrees with the data plots; i.e., a linear response with interaction between time and temperature.
The data do not quite have a constant coefficient of variation ...?
> dialectr <- read.table("dialectr.txt", header=T)
> library(grid);library(lattice)
> dialectr$fTime <- factor(dialectr$Time)
> dialectr$fTemperature <- factor(dialectr$Temperature)
> attach(dialectr)

> tapply(Strength,fTime,mean)

1 2 4 8 16 32 48 64

> gm <- tapply(Strength,fTime,mean)
> tapply(Strength,fTime,sd)

1 2 4 8 16 32 48 64
1.695275 1.087811 2.239559 3.623994 4.183375 4.258594 5.146904 5.140211

> gs <- tapply(Strength,fTime,sd)
> cv <- (gs/gm) * 100
> cv

1 2 4 8 16 32 48 64
11.76764 8.44902 17.26889 31.51299 35.73625 45.94572 58.86381 60.93456

> gmet <- tapply(Strength,fTemperature,mean)
> gsd <- tapply(Strength, list(fTime, fTemperature), sd)
> (gsd/gme) * 100

<table>
<thead>
<tr>
<th></th>
<th>180</th>
<th>225</th>
<th>250</th>
<th>275</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.705</td>
<td>4.232</td>
<td>13.950</td>
<td>5.901</td>
<td>6.994</td>
</tr>
<tr>
<td>9.309</td>
<td>3.718</td>
<td>3.402</td>
<td>5.657</td>
<td>8.694</td>
</tr>
<tr>
<td>14.899</td>
<td>8.983</td>
<td>5.940</td>
<td>8.694</td>
<td>8.694</td>
</tr>
<tr>
<td>3.114</td>
<td>12.192</td>
<td>4.031</td>
<td>6.804</td>
<td>18.842</td>
</tr>
<tr>
<td>8.311</td>
<td>9.900</td>
<td>2.105</td>
<td>8.510</td>
<td>18.842</td>
</tr>
<tr>
<td>13.435</td>
<td>7.059</td>
<td>3.888</td>
<td>5.825</td>
<td>18.842</td>
</tr>
<tr>
<td>11.226</td>
<td>10.526</td>
<td>11.618</td>
<td>18.842</td>
<td>18.842</td>
</tr>
<tr>
<td>14.077</td>
<td>6.005</td>
<td>5.542</td>
<td>16.829</td>
<td>18.842</td>
</tr>
</tbody>
</table>

>
Lecture 14 !
<table>
<thead>
<tr>
<th>Injury</th>
<th>right</th>
<th>both</th>
<th>left</th>
<th>mean counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>no (counts)</td>
<td>313</td>
<td>15</td>
<td>21</td>
<td>349</td>
</tr>
<tr>
<td>yes (counts)</td>
<td>535</td>
<td>43</td>
<td>81</td>
<td>659</td>
</tr>
<tr>
<td>mean counts</td>
<td>848</td>
<td>58</td>
<td>102</td>
<td>1008</td>
</tr>
</tbody>
</table>
injury.model<-glm(counts ~ handed*injury,family=poisson(link=log))

injury.aov<-anova(injury.model,test="Chisq")

The analysis of deviance is:-

Analysis of Deviance Table

Poisson model

Response: counts

Terms added sequentially (first to last)

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance Resid. Df</th>
<th>Resid. Dev</th>
<th>Pr(Chi)</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>5</td>
<td>1234</td>
<td></td>
<td></td>
</tr>
<tr>
<td>handed</td>
<td>2</td>
<td>1123</td>
<td>3</td>
<td>111</td>
</tr>
<tr>
<td>injury</td>
<td>1</td>
<td>97</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>handed:injury</td>
<td>2</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
1 no relief
2 moderate relief
3 complete relief

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>P</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Treatment</td>
<td>Group = 1 Response</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------</td>
</tr>
<tr>
<td>A</td>
<td>3 5 8</td>
</tr>
<tr>
<td>P</td>
<td>11 5 0</td>
</tr>
</tbody>
</table>
library(MASS)
ordmod1 <- polr(Response ~ Group*Treatment, data=clintrial)
print(summary(ordmod1))

<table>
<thead>
<tr>
<th>category</th>
<th>$B_1$</th>
<th>$B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Call:
polr(formula = Response ~ Group * Treatment, data = clintrial)

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group2</td>
<td>-0.559</td>
<td>0.706</td>
<td>-0.791</td>
</tr>
<tr>
<td>TreatmentP</td>
<td>-2.406</td>
<td>0.739</td>
<td>-3.254</td>
</tr>
<tr>
<td>Group2:TreatmentP</td>
<td>0.539</td>
<td>1.055</td>
<td>0.511</td>
</tr>
</tbody>
</table>

Intercepts:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-1.496</td>
<td>0.538</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.011</td>
<td>0.483</td>
</tr>
</tbody>
</table>

Residual Deviance: 108.40
AIC: 118.40