12 Modelling Binomial Response Data

12.1 Examples of Binary Response Data

Binary response data arise when an observation on an individual can be placed into one of exactly two categories. Some examples:

- A manufacturing process is either satisfactory or unsatisfactory.
- A baby is delivered at either full term or prematurely.
- A customer applying for a loan is either a good risk or a bad risk.

Generally, we can think of an outcome as resulting in either success or failure.

12.2 Modelling Binary Response Probabilities

We consider a general framework for modelling 'success' probabilities.

Suppose that we have $G$ groups, with $N_i$ individuals in the $i$-th group. The value of the explanatory variables for those individuals in group $i$ is $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$.

Let $\pi_i$ be the probability of success for an individual in the $i$-th group, and $Y_i$ be the number of successes in that group also (thus $N_i - Y_i$ is the number of failures).

Then assuming mutual independence between the individuals, it follows that

$$Y_i \sim \text{Bin}(N_i, \pi_i) \quad i = 1, \ldots, G.$$  \hspace{1cm} (1)

Our goal is to estimate the $\{\pi_i\}$ and related quantities on the basis of the $\{x_i\}$ and the $\{y_i\}$. To this end, next we motivate the use of appropriate link functions.

12.3 Transformations

Since

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \in (-\infty, \infty)$$

and $\pi_i \in (0, 1)$, then clearly $\pi_i$ cannot be regressed on the explanatory variables directly.

In order to overcome this obstacle, we consider applying a transformation to $\pi_i$, yielding a
new quantity on $(-\infty, \infty)$, and then relating it to $\eta_i$.

Here are some possible transformations:

**Logistic**

$$\xi_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right)$$

i.e. $\xi_i = \logit(\pi_i)$

**Probit**

$$\xi_i = \Phi^{-1}(\pi_i)$$

where $\Phi(\cdot)$ is the cumulative distribution function (c.d.f.) for the *Standard Normal* distribution, i.e.

$$\pi_i = \int_{-\infty}^{\xi_i} \phi(u) du$$

with

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

**Complementary log-log**

$$\xi_i = \log[- \log(1 - \pi_i)]$$

### 12.4 Logistic Regression

Suitability of the transformations (2)-(4) depends on the context and situation. However, there are a few reasons for considering the logistic transformation when it is appropriate:

- $\log \left( \frac{\pi_i}{1 - \pi_i} \right)$ can be interpreted as the logarithm of the *odds in favour of success*;
- it is easy to compute;
- it is the canonical link for the Binomial error structure.

This choice of transformation yields the framework which is known as *logistic regression*; more formally

(a) $$Y_i \sim \text{Bin}(N_i, \pi_i), \ i = 1, \ldots, G$$

are mutually independent.

(b) The explanatory variables provide a set of linear predictors

$$\eta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \ i = 1, \ldots, G.$$
(c): (a) and (b) are connected via

\[ g(\pi_i) = \log \left( \frac{\pi_i}{1 - \pi_i} \right) \]

with \( g(\pi_i) = \eta_i \).

The general expression for the deviance is:

\[ D = 2 \sum_{i=1}^{G} \left[ g_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right] \quad (5) \]

Let us derive the form of the deviance under the conditions specified above. In previous calculations\(^1\) it was found that

\[ \phi = 1, \quad \alpha_i(\phi) = \phi, \quad \theta_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right) \]

\[ b(\theta_i) = N_i \log(1 + e^{\theta_i}) = -N_i \log \left( \frac{1}{1 + e^{\theta_i}} \right) = -N_i \log(1 - \pi_i) \]

and hence

\[ \mu_i = b'(\theta_i) = \frac{N_i e^{\theta_i}}{1 + e^{\theta_i}} = N_i \pi_i \]

with \( \pi_i = \mu_i / N_i \).

The maximum likelihood estimates for \( \theta_i \) and \( b(\theta_i) \) are given by

\[ \hat{\theta}_i = \log \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) = \log \left( \frac{N_i \hat{\pi}_i}{N_i - N_i \hat{\pi}_i} \right) = \log \left( \frac{\hat{\mu}_i}{N_i - \hat{\mu}_i} \right) \]

\[ b(\hat{\theta}_i) = -N_i \log(1 - \hat{\pi}_i) = -N_i \log \left( 1 - \frac{\hat{\mu}_i}{N_i} \right) = -N_i \log \left( \frac{N_i - \hat{\mu}_i}{N_i} \right) \]

respectively.

The corresponding values for the saturated model are

\[ \tilde{\theta}_i = \log \left( \frac{\tilde{\pi}_i}{1 - \tilde{\pi}_i} \right) = \log \left( \frac{y_i/N_i}{1 - y_i/N_i} \right) = \log \left( \frac{y_i}{N_i - y_i} \right) \]

\[ b(\tilde{\theta}_i) = -N_i \log(1 - \tilde{\pi}_i) = -N_i \log \left( 1 - \frac{y_i}{N_i} \right) = -N_i \log \left( \frac{N_i - y_i}{N_i} \right) \]

\(^1\)Examples 5 Question 2.
noting that $\tilde{\mu}_i = y_i$, and so $\tilde{\pi}_i = y_i/N_i$.

Plugging these expressions back into (5) and re-arranging terms yields

$$D = 2 \sum_{i=1}^{G} \left[ y_i \log \left( \frac{y_i}{\tilde{\mu}_i} \right) + (N_i - y_i) \log \left( \frac{N_i - y_i}{N_i - \tilde{\mu}_i} \right) \right]$$

(6)

which has a $\chi^2_{G-p}$ distribution approximately under the hypothesis that the model fits the data sufficiently well ($G > p$); however, we should be cautious about the use of $D$, as the aforementioned distribution may be a poor approximation to the true distribution of $D$.

### 12.5 Ungrouped Binary Data

To a limited extent, we can deal with data that are ungrouped: in effect, each individual forms its own group, so that $G = n$. Thus one can estimate parameters in the manner suggested in previous sections, with $N_i = 1, i=1, \ldots, G (= n)$.

However, it appears that, under logistic regression, $D$ is not useful for assessing goodness-of-fit; this is because $D$ can be reduced to an expression that involves fitted values, but not the observed values of the dependent variable, $\{y_i\}$ (thus the comparison between fitted and observed values does not appear to exist). The case for this is argued below.

Setting $N_i = 1$ for all $i$, and $G = n$ in (6) yields the expression

$$D = 2 \sum_{i=1}^{n} \left[ y_i \log y_i - y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - y_i) - (1 - y_i) \log(1 - \hat{\pi}_i) \right].$$

Defining $0^0 = 1$, then $y_i \log y_i$ and $(1 - y_i) \log(1 - y_i)$ are both equal to 0 at the two possible values of $y_i$ (which are 0 and 1).

Hence, our expression for $D$ reduces to

$$D = -2 \sum_{i=1}^{n} \left[ y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i) \right]$$

$$= -2 \sum_{i=1}^{n} \left[ y_i \log \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \log(1 - \hat{\pi}_i) \right]$$

Next we show that $y_i \log \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right)$ can be replaced by a term not involving $y_i$.

Since $Y_i \sim \text{Bin}(1, \pi_i) \sim \text{Bernoulli}(\pi_i)$, then the corresponding log-likelihood is given by

$$l(\beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left\{ y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right\}.$$
It can be shown that
\[
\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left\{ \frac{y_i}{\pi_i} - \frac{1}{1 - \pi_i} \right\} \pi_i (1 - \pi_i) x_{ij}
\]
\[
= \sum_{i=1}^{n} (y_i - \pi_i) x_{ij}
\]

\(j = 1, \ldots, p\).

Since the \(\left\{ \frac{\partial l}{\partial \beta_j} \right\}\) are equal to 0 at \(\beta = \hat{\beta}\), then
\[
\sum_{j=1}^{p} \beta_j \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \pi_i) \sum_{j=1}^{p} \beta_j x_{ij}
\]
\[
= \sum_{i=1}^{n} (y_i - \pi_i) \log \left( \frac{\pi_i}{1 - \pi_i} \right) = 0
\]
at \(\beta = \hat{\beta}\) also.

Hence
\[
\sum_{i=1}^{n} y_i \log \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) = \sum_{i=1}^{n} \hat{\pi}_i \log \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right)
\]
and thus \(D\) becomes
\[
D = -2 \sum_{i=1}^{n} \left\{ \hat{\pi}_i \log \left( \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \log (1 - \hat{\pi}_i) \right\}.
\] (7)

12.6 Examples
12.6.1 Dose-Response Model

**Bliss (1935)**

Recall the data showing the numbers of insects that have died after several hours exposure to a certain pesticide.

<table>
<thead>
<tr>
<th>Dose (x_i) log_{10} CS₂ mg⁻¹</th>
<th>Number of insects, (n_i)</th>
<th>Number killed, (y_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6907</td>
<td>69</td>
<td>6</td>
</tr>
<tr>
<td>1.7242</td>
<td>60</td>
<td>13</td>
</tr>
<tr>
<td>1.7552</td>
<td>62</td>
<td>18</td>
</tr>
<tr>
<td>1.7842</td>
<td>56</td>
<td>28</td>
</tr>
<tr>
<td>1.8113</td>
<td>63</td>
<td>52</td>
</tr>
<tr>
<td>1.8369</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>1.8610</td>
<td>62</td>
<td>61</td>
</tr>
<tr>
<td>1.8839</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>
Independence between observations, Binomial error structure, logit link, and linear predictor given by

\[ \eta_i = \beta_1 + \beta_2 x_i \]

was assumed and the model fitted in S-PLUS; an abbreviated version of the output is presented below.

```r
> pesticide.glm <- glm(prop ~ dose, family = binomial, weights = number, data = pesticide)
> summary(pesticide.glm)

Call: glm(formula = prop ~ dose, family = binomial, data = pesticide, weights = number)

Deviance Residuals:
Min 1Q Median 3Q Max
-1.594124 -0.3943968 0.8329153 1.259223 1.593985

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-60.71745</td>
<td>5.179902</td>
<td>-11.72174</td>
</tr>
<tr>
<td>dose</td>
<td>34.27032</td>
<td>2.911680</td>
<td>11.76995</td>
</tr>
</tbody>
</table>

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 284.2024 on 7 degrees of freedom
Residual Deviance: 11.23223 on 6 degrees of freedom
Number of Fisher Scoring Iterations: 5

This has a (scaled) deviance of 11.23223 on 6 degrees of freedom: this corresponds to an (approximate) \( p \)-value of 0.08145884; this is just about acceptable. Let us try to improve upon the fit by selecting other link functions.

To use the *probit* link, we invoke one of the optional arguments in the \texttt{glm} command:

```r
> pesticide.glm <- glm(prop ~ dose, family = binomial(link = probit), weights = number, data = pesticide)
> summary(pesticide.glm)

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-39.57238</td>
<td>3.239362</td>
<td>-12.21610</td>
</tr>
<tr>
<td>dose</td>
<td>22.04121</td>
<td>1.798838</td>
<td>12.25363</td>
</tr>
</tbody>
</table>

(Dispersion Parameter for Binomial family taken to be 1)
```

Finally, let us try the *complementary log-log* link:

```r
> pesticide.glm <- glm(prop ~ dose, family = binomial(link = cloglog), weights = number, data = pesticide)
> summary(pesticide.glm)

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-39.57238</td>
<td>3.239362</td>
<td>-12.21610</td>
</tr>
<tr>
<td>dose</td>
<td>22.04121</td>
<td>1.798838</td>
<td>12.25363</td>
</tr>
</tbody>
</table>
```

(Dispersion Parameter for Binomial family taken to be 1)
Null Deviance: 284.2024 on 7 degrees of freedom
Residual Deviance: 3.446439 on 6 degrees of freedom

This time the deviance is 3.446439 on 6 degrees of freedom: a serious improvement! So if there were nothing about the situation or experiment that would preclude the use of such a link, then this could be the model to go for.

The estimated values of the parameters are
\[ \hat{\beta}_1 = -39.57238, \quad \hat{\beta}_2 = 22.04121 \]

From these, related quantities like fitted values, and deviances can be evaluated. Indeed, from (4), we have
\[ \pi_i = 1 - \exp[-\exp(\xi_i)] \]

which implies that
\[ \hat{\pi}_i = 1 - \exp[-\exp(\eta_i)] = 1 - \exp[-\exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)] \]

The output below shows both an explicit calculation of the fitted values for the \{\pi_i\} as per the formula above, and the values obtained using `fitted`

```r
d <- coef(pesticide.glm)[1] + coef(pesticide.glm)[2] * dose
my.fitted <- 1 - exp(-exp(d))
```

```r
my.fitted
[1] 0.09473606 0.18801078 0.33797074 0.54231111 0.75835588 0.91767360 0.98569872 0.99912044
```

```r
fitted(pesticide.glm)
1 2 3 4 5 6 7 8
0.09473606 0.1880108 0.3379707 0.5423111 0.7583559 0.9176736 0.9856987 0.9991204
```

12.6.2 Challenger Mission

Dalal et. al(1989)

Data on the 23 space shuttle flights that occurred before the Challenger mission in 1986 are given in the following table. For each of the 23 missions, data on the temperature, in °F, at the time of flight (Temp.), and whether at least one primary O-ring suffered thermal distress (TD) were recorded.

<table>
<thead>
<tr>
<th>Temp</th>
<th>TD</th>
<th>Temp</th>
<th>TD</th>
<th>Temp</th>
<th>TD</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>0</td>
<td>57</td>
<td>1</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>70</td>
<td>1</td>
<td>63</td>
<td>1</td>
<td>81</td>
<td>0</td>
</tr>
<tr>
<td>69</td>
<td>0</td>
<td>70</td>
<td>1</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>0</td>
<td>78</td>
<td>0</td>
<td>79</td>
<td>0</td>
</tr>
<tr>
<td>67</td>
<td>0</td>
<td>67</td>
<td>0</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>72</td>
<td>0</td>
<td>53</td>
<td>1</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>73</td>
<td>0</td>
<td>67</td>
<td>0</td>
<td>58</td>
<td>1</td>
</tr>
<tr>
<td>70</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
TD can be thought of as being the observed probability of thermal distress. We fit a logistic regression model between TD and Temp.

```r
> td <- c(0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1)
> temp <- c(66, 70, 69, 68, 67, 72, 73, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70,
+ 81, 76, 79, 75, 76, 58)
> shuttle <- data.frame(td, temp)
> shuttle.glm <- glm(td ~ temp, family = binomial, data = shuttle)
> summary(shuttle.glm)
```

Call: `glm(formula = td ~ temp, family = binomial, data = shuttle)`

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Residual</td>
<td>-1.06</td>
<td>-0.76</td>
<td>-0.38</td>
<td>0.45</td>
<td>2.22</td>
</tr>
<tr>
<td>Max Residual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>15.04</td>
<td>7.34</td>
<td>2.05</td>
</tr>
<tr>
<td>temp</td>
<td>-0.23</td>
<td>0.11</td>
<td>-2.16</td>
</tr>
</tbody>
</table>

(Dispersion Parameter for Binomial family taken to be 1)

Null Deviance: 28.26715 on 22 degrees of freedom

Residual Deviance: 20.31519 on 21 degrees of freedom

Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>temp</td>
<td>-0.997</td>
<td>1.00</td>
</tr>
</tbody>
</table>

> anova(shuttle.glm)

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Terms added sequentially (first to last)</th>
<th>Df</th>
<th>Deviance</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>22</td>
<td>28.27</td>
<td>21</td>
<td>20.31</td>
</tr>
<tr>
<td>temp</td>
<td>1</td>
<td>7.95</td>
<td>20</td>
<td>20.31</td>
</tr>
</tbody>
</table>

Notice that the `weights` argument has not been specified since we are dealing with ungrouped Binary response data.

The residual deviance is 20.31519 on 21 degrees of freedom. However, this is not relevant as this is an example of ungrouped Binomial response data. Examination of the various residuals (as discussed in Chapter 10) can provide an alternative route for assessing the goodness-of-fit of the model.

Assuming, for the sake of argument, that the model does provide a good fit, then we see
that
\[
\hat{\pi} = \frac{\exp(\hat{\beta}_1 + \hat{\beta}_2 x)}{1 + \exp(\hat{\beta}_1 + \hat{\beta}_2 x)}
\]  \hspace{1cm} (8)

where \(\hat{\beta}_1 = 15.0422911\) and \(\hat{\beta}_2 = -0.2321537\), where \(x\) is the temperature at the time of flight, and \(\hat{\pi}\) is the (corresponding) estimated probability of thermal distress of at least one primary O-ring.

**Question:** What is the predicted probability of thermal distress at 31\(^{\circ}\)F (supposedly the temperature at the time of the Challenger flight)?

Setting \(x = 31\) in (8) yields
\[
\hat{\pi} = \frac{\exp(15.0422911 - 0.2321537 \times 31)}{1 + \exp(15.0422911 - 0.2321537 \times 31)} = 0.9996087
\]
i.e. it is predicted that thermal distress is almost certain at that temperature.

**Question:** At which temperature is it estimated that there is a 50% chance that thermal distress occurs?

Substituting for \(\hat{\pi} = 0.5\) into (8), we find that the corresponding \(x\) satisfies
\[
\hat{\beta}_1 + \hat{\beta}_2 x = 0
\]
which implies that
\[
x = -\frac{\hat{\beta}_1}{\hat{\beta}_2} = -\frac{15.0422911}{-0.2321537} = 64.79454
\]
i.e. about 65\(^{\circ}\)F.

**Exercise 12.1**
Try working through Example 6.1 in Krzanowski.