Binomial Link Functions

Lori Murray, Phil Munz
Binomial Link Functions

• Logit Link function: \[ \eta(p) = \ln\left(\frac{p}{1-p}\right) \]

• Probit Link function: \[ \eta(p) = \Phi^{-1}(p) \]

• Complementary Log Log function: \[ \eta(p) = \ln(-\ln(1-p)) \]
Motivating Example

• A researcher is examining beetle mortality after 5 hours of exposure to carbon disulphide, at various levels of concentration of the gas.

• Beetles were exposed to gaseous carbon disulphide at various concentrations (in mg/L) for five hours (Bliss, 1935) and the number of beetles killed were noted. The data are in the following table:
Example (continued)

```r
> beetle <- read.table("BeetleData.txt", header=TRUE)
> head(beetle)
  Dose Num.Beetles Num.Killed
1  1.6907      59          6
2  1.7242      60         13
3  1.7552      62         18
4  1.7842      56         28
5  1.8113      63         52
6  1.8369      59         53

> probitmodel <- glm(cbind(Num.Killed, Num.Beetles - Num.Killed) ~ Dose, data = beetle, family = binomial(link = probit)) > summary(probitmodel)
> logmodel <- glm(cbind(Num.Killed, Num.Beetles - Num.Killed) ~ Dose, data = beetle, family = binomial(link = cloglog)) > summary(logmodel)
```
Don’t forget to plot the data!
LOGIT MODEL:

Call:
 glm(formula = cbind(Num.Killed, Num.Beetles - Num.Killed) ~ Dose,
     family = binomial, data = beetle)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.5941 -0.3944   0.8329   1.2592   1.5940

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
     Dose        34.270      2.912   11.77   <2e-16 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
PROBIT MODEL:

Call:
glm(formula = cbind(Num.Killed, Num.Beetles - Num.Killed) ~ Dose, 
    family = binomial(link = probit), data = beetle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value  Pr(>|z|) 
(Intercept) -34.935     2.648  -13.19   <2e-16 ***
     Dose     19.728     1.487   13.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.20  on 7  degrees of freedom
Residual deviance: 10.12  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4
COMPLEMENTARY LOG-LOG MODEL:

Call:
glm(formula = cbind(Num.Killed, Num.Beetles - Num.Killed) ~ Dose,
    family = binomial(link = cloglog), data = beetle)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-0.80329  -0.55135   0.03089   0.38315   1.28883

Coefficients:
             Estimate  Std. Error     z value     Pr(> |z|)
(Intercept)  -39.5722    3.2400   -12.2100   < 2.2e-16 ***
Dose          22.0410    1.7990    12.2500   < 2.2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
    Residual deviance:   3.4464  on 6  degrees of freedom
    AIC: 33.644

Number of Fisher Scoring iterations: 4
Example (continued)
Binomial Link Functions

- Differences in choice of link affect model and deviance.

- Why have 3 link functions and what about them cause these differences.

- “All models are wrong, but some are useful” – George Box
Differences in Link Functions

![Graph showing differences in link functions](image-url)
Differences in Link Functions

• Numerically, consider the specific value of each function corresponding to various levels of p:

<table>
<thead>
<tr>
<th>p</th>
<th>Logit</th>
<th>Probit</th>
<th>C Log Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>-5.2933</td>
<td>-2.5758</td>
<td>-5.2958</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>-0.3665</td>
</tr>
<tr>
<td>0.99</td>
<td>4.5951</td>
<td>2.3263</td>
<td>1.5271</td>
</tr>
</tbody>
</table>
Deviances

\[ D = 2 \sum_{i=1}^{n} \left[ y_i \ln \left( \frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \ln \left( \frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right]; \hat{y}_i = n_i \hat{p}_i, \]

- **Logit:**
  \[ \hat{p}_i = \frac{e^{x_i^T \hat{\beta}}}{1 + e^{x_i^T \hat{\beta}}} \]

- **Probit:**
  \[ \hat{p}_i = \Phi(x_i^T \hat{\beta}) \]

- **C Log Log:**
  \[ \hat{p}_i = 1 - \exp \{- \exp [x_i^T \hat{\beta}] \} \]
Differences in Link Functions

```r
probLowerlogit <- vector(length=1000)
probLowercloglog <- vector(length=1000)
logitDeviance <- vector(length=1000)
probitDeviance <- vector(length=1000)
cloglogDeviance <- vector(length=1000)
probLowerlogitclog <- vector(length=1000)
for(i in 1:1000){

  x <- rnorm(1000)
  y <- rbinom(n=1000, size=1, prob=pnorm(x))

  logitModel <- glm(y~x, family=binomial(link="logit"))
  probitModel <- glm(y~x, family=binomial(link="probit"))
  cloglogModel <- glm(y~x, family=binomial(link="cloglog"))

  logitDeviance[i] <- deviance(logitModel)
  probitDeviance[i] <- deviance(probitModel)
  cloglogDeviance[i] <- deviance(cloglogModel)

  probLowerlogit[i] <- probitDeviance[i] < logitDeviance[i]
  probLowercloglog[i] <- probitDeviance[i] < cloglogDeviance[i]
  probLowerlogitclog[i] <- logitDeviance[i] < cloglogDeviance[i]
}
```
Differences in Link Functions

\[
\begin{align*}
&> \text{sum(probLowerlogit)/1000} \\
&\quad [1] 0.695 \\
&> \text{sum(probLowercloglog)/1000} \\
&\quad [1] 0.906 \\
&> \text{sum(probLowerlogitclog)/1000} \\
&\quad [1] 0.877 \\
\end{align*}
\]

Differences (last iteration):
\[
\begin{align*}
&> \text{deviance(logitModel)} - \text{deviance(probitModel)} \\
&\quad [1] 0.6076806 \\
&> \text{deviance(cloglogModel)} - \text{deviance(probitModel)} \\
&\quad [1] -1.152768 \\
\end{align*}
\]

Consider the last iteration of the script:

<table>
<thead>
<tr>
<th>Dev Probit</th>
<th>Dev Logit</th>
<th>Dev. cloglog</th>
</tr>
</thead>
<tbody>
<tr>
<td>1025.759</td>
<td>1026.366</td>
<td>1024.606</td>
</tr>
</tbody>
</table>
Origins of the Binominal Link Functions

1. Complementary log log link (1922)
2. Probit link (1933)
3. Logit link (1944)
Complementary log-log link (1922)

- R. A. Fisher, English Statistician
- Dilution assay §12.3
- Describes an experiment where a series of dilutions were made of a soil or water sample to determine the presence or absence of some microbial contaminant.
- Used a cll transformation and applied maximum likelihood estimation.
Complementary log-log link (1922)

• Assume that dilutions are made in powers of 2, then after $x$ dilutions the number of infective organisms, $p_x$, per unit volume is

$$p_x = p_0/2^x \quad x = 0,1,\ldots$$

• where $p_0$ is the density of infective organisms in the original solution (we wish to estimate).

• The expected number of organisms on any plate is $p_x v$, and the actual number of organisms follows a Poisson distribution with this parameter.
Complementary log-log link (1922)

• The probability that a plate is infected is

\[ \pi_x = 1 - \exp\{ -p_x v \} \]

• At dilution \( x \) we have,

\[
\log(-\log(1 - \pi_x)) = \log v + \log p_x \\
= \log v + \log p_o - x \log 2
\]

• If at dilution \( x \) we have \( r \) infected plates out of \( m \), the observed proportion of infected plates is \( y = r/m \), and \( E(Y| x) = \pi_x \)

• A complementary log-log transformation is

\[
\log(-\log(1 - \pi_x)) = \alpha + \beta x
\]
Probit link (1933/1934)

- John Gaddum was an English pharmacologist who wrote a comprehensive report on the statistical interpretation of bio-assay.

- Bliss was largely self taught, worked with Fisher, and eventually settled at Yale.
  - Published 2 brief notes in *Science* where he introduced the word ‘probit’ (probability unit).
Bliss uses an example of the effectiveness of a pesticide to combat an insect pest.

- Describes how a dosage-mortality curve has an asymmetrical S-shaped curve.
Probit link (1933/1934)

• Observation that in many physiological processes equal increments in response are produced when dose is increased by a constant proportion of the given dosage, rather than by a constant amount.

• Bliss proposed the same rule might hold for toxicological processes, in which case dosage would have to be plotted in logarithmic terms to show a uniform increase in mortality.

• Proposed to transform the percentage killed to a probit and then plot against the logarithm of the dose to achieve a straight line.
Probit link (1933/1934)

- Transformation by use of logarithms and probits.
Logit link (1944)

• Joseph Berkson was a medical doctor and chief statistician of the Mayo Clinic.

• Research was on statistical methodology of bio-assay.

• Proposed the use of the logistic instead of the normal probability function, coining the term ‘logit’ by analogy to the ‘probit’ of Bliss.
Logit link (1944)

• Berkson gives several reasons for using the logit
  – The logistic function is very close to the integrated normal curve.
  – Since it applies to a wide range of physiochemical phenomena, it may have a better theoretical basis than the integrated normal curve.
  – It is easier to handle statistically.

• Initially the logit was regarded as inferior and disreputable, since it cannot be related to an underlying normal distribution of tolerance levels.
Logit link (1944)

- By the 1960s, Berkson’s logit had gained acceptance.

- The power of the logistic’s analytical properties were starting to surface.

- By the 1970s, the logit takes the lead because it was now widely used among many disciplines.

Table 1. Number of articles in statistical journals containing the word 'probit' or 'logit'.

<table>
<thead>
<tr>
<th></th>
<th>probit</th>
<th>logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1935–39</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>1940–44</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1945–49</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>1950–54</td>
<td>50</td>
<td>15</td>
</tr>
<tr>
<td>1955–59</td>
<td>53</td>
<td>23</td>
</tr>
<tr>
<td>1960–64</td>
<td>41</td>
<td>27</td>
</tr>
<tr>
<td>1965–69</td>
<td>43</td>
<td>41</td>
</tr>
<tr>
<td>1970–74</td>
<td>48</td>
<td>61</td>
</tr>
<tr>
<td>1975–79</td>
<td>45</td>
<td>72</td>
</tr>
<tr>
<td>1980–84</td>
<td>93</td>
<td>147</td>
</tr>
<tr>
<td>1985–89</td>
<td>98</td>
<td>215</td>
</tr>
<tr>
<td>1990–94</td>
<td>127</td>
<td>311</td>
</tr>
</tbody>
</table>
Logit is Considered the Default Link

• Advantages of Logit link function:
  – Leads to simpler mathematics due to complexity of the standard normal CDF
  – It is easier to interpret (Log odds)
Final Remarks

• If the logit link is considered the default link, why do we still use probit and Complementary log log?
  – Theoretical Considerations
  – Influences by disciplinary tradition
    • Economists favour probit models
    • Toxicologists favour logit models
  – Underlying characteristics of the data
    • Complementary log log works best with extremely skewed distributions
References


References

