Introduction to Categorical Data Analysis

Adrian Dunne
Department of Statistics and Actuarial Science,
Roinn na Staitisticí agus na hAchtúreolaíochta
University College Dublin
An Coláiste Ollscoile Baile Átha Cliath
Data Types

• Quantitative
  – Continuous Plasma drug conc., BP, Muscle Tension, Time
  – Discrete Number of blood cells, Number of heart attacks

• Categorical
  – Nominal Religion, Nationality, Gender
  – Ordinal Social class, Treatment outcome
  – Binary Gender, Dead/Alive
Example: Categorical Data

Population of patients require a particular analgesic

Drug administered to sample of patients

Each patient is assessed for pain relief – none, some, complete

Estimate population proportions in each category – none, some, complete.
Binary Data

• Describe the two categories as “Success” (S) and “Failure” (F).

• Code

\[ Z = 0 \quad \text{for } F \]
\[ Z = 1 \quad \text{for } S \]
Binary Data

• Proportion of S in population.

• Randomly select a member of the population – probability of S.

• Proportion = Probability
Binary Data

- \( Z \) has the Bernoulli Distribution

\[
\begin{align*}
\Pr(Z = 1) &= \pi \quad \text{Prob of } S \\
\Pr(Z = 0) &= 1 - \pi \quad \text{Prob of } F \\
\Pr(Z = r) &= \pi^r (1 - \pi)^{1-r} \quad r = 0, 1
\end{align*}
\]
Estimation: Method of Maximum Likelihood

- Likelihood

\[ L(\pi) = \prod_{i=1}^{n} \pi^{z_i} (1 - \pi)^{(1-z_i)} \]

- \( \hat{\pi} \) is the value of \( \pi \) that maximises the likelihood
Simple Example

10 observations: 3S’s 7F’s
Simple model with no structure

\[ Z \sim \text{Bernoulli}(\pi) \]

\[ L(\pi) = \prod_{i=1}^{n} \pi^{z_i} (1 - \pi)^{(1-z_i)} \]
Example: Likelihood

![Likelihood Curve](image-url)
Data Modelling

• Previous example had no structure in the data.

• Consider the case where the subjects were administered different doses of drug and the response depends on dose.

• Another example would be where response changes with time following drug administration (PK-PD).
Data Modelling

• Take account of the structure by recording the values of covariates (x’s) for each member of the sample e.g. dose, time.

• Then construct a model which describes how the parameters depend on the covariates.
Modelling Binary Data

• We model $\pi_i$ e.g.

$$\pi_i = f(x_i, \theta)$$

• However,

$$0 \leq \pi_i \leq 1$$

• There is no guarantee that

$$0 \leq f(x_i, \hat{\theta}) \leq 1$$
Modelling Binary Data

Transform $\pi_i$ from $(0,1)$ to $(-\infty, +\infty)$ and model the transformed value to ensure that model predicted probabilities lie in $(0,1)$. 
Transformations

• Logit
  \[ \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) \]

• Probit
  \[ \text{probit}(\pi_i) = \Phi^{-1}(\pi_i) \]

• Log-log
  \[ \log(-\log(\pi_i)) \]

• Complementary log-log
  \[ \log(-\log(1 - \pi_i)) \]
Probit transformation

\[ \pi \]

Normal distribution

Probit
Transformations

- Log-log
- Comp. Log-log
- Probit
- Logit
Logistic Regression Model

Now consider a model with structure

$$\text{logit}(\pi_i) = f(x_i, \theta)$$

Example

$$\text{logit}(\pi_i) = \theta_1 + \theta_2 x_{1i} + \theta_3 x_{2i} + \ldots$$
## Example: Bioassay

Beetle deaths following dosing with an insecticide

<table>
<thead>
<tr>
<th>Dose</th>
<th># Exposed</th>
<th># Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0028</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>0.0056</td>
<td>40</td>
<td>19</td>
</tr>
<tr>
<td>0.0112</td>
<td>40</td>
<td>31</td>
</tr>
<tr>
<td>0.0225</td>
<td>40</td>
<td>34</td>
</tr>
<tr>
<td>0.0450</td>
<td>40</td>
<td>39</td>
</tr>
</tbody>
</table>
The graph shows the relationship between dose and the logit of proportion dead. As the dose increases, the logit of proportion dead also increases, indicating a positive correlation. The data points suggest that lower doses result in lower proportions of death, while higher doses lead to higher proportions of death.
Logit Model

Linear logistic model with log(dose)

\[ z_i \sim \text{Bernoulli}(\pi_i) \]

\[ \log \text{it}(\pi_i) = \theta_1 + \theta_2 \log(\text{dose}_i) \]

\[ L(\theta) = \prod_{i=1}^{n} \pi_i^{z_i} (1 - \pi_i)^{(1-z_i)} \]
Observed & Predicted Values

![Graphs showing observed and predicted values](image-url)
Mixed Effects Modelling

• Modelling correlation between responses/variation between groups.
  – Groups of related items
  – Repeated measures/Longitudinal data
Example: Binary PK-PD response

- 10 subjects all received dose of 100 units.

- Bolus iv administration.

- Binary response (dry mouth) recorded for each subject at times 0.5, 1, 2, 3, 5, 7, 9, 12, 15, 18, 24, 30 hours.
PK-PD Model

Bolus (D) at \( t=0 \)

Central Compartment

\( C_p(t) \)

\( \text{Effect Compartment} \)

\( C_e(t) \)

\( k_{1e} \)

\( k_{ce} \)

\[ \text{Eff} (t) = f (C_e(t)) \]
PK-PD Model

• Linear PD model

\[ Eff(t) = \theta_1 + \theta_2 C_e(t) \]

\[ \text{logit}(\pi(t)) = \theta_1 + \theta_2 C_e(t) \]
Example: Binary Population PK-PD model

• Variation between subjects

• Longitudinal (repeated measures) data – observations on same subject are correlated

• Model intrasubject correlation and intersubject variation using random effects
Example: Binary Population PK-PD model

\[
\text{logit } (\pi_i(t_j)) = \theta_1 + \theta_2 C_e(t_j) + \eta_i
\]

\[
\eta_i \sim N(0, \Omega)
\]

\[
L(\Theta, \Omega) = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} \prod_{j=1}^{m_i} \pi_i(t_j)^{z_{ij}} (1 - \pi_i(t_j))^{(1-z_{ij})} f(\eta_i, \phi) d\eta_i
\]
Observed & Predicted Values
Latent Variable

• Consider again the insecticide bioassay

• Assume that each insect has an (unobserved) tolerance $t_i$ which varies randomly across the population of insects

\[
\begin{align*}
    t_i \leq d_i & \implies z_i = 1 \\
    t_i > d_i & \implies z_i = 0
\end{align*}
\]
Tolerance distribution

\[ \pi_i = \Pr(t_i \leq d_i) \]

\[ z_i = 1 \quad \text{for} \quad \pi_i \]
\[ z_i = 0 \]
Latent Variables

- $d_i$ is known as the cut-point
- Here the latent variable is tolerance
\[ \pi_i = \Pr(t_i \leq d_i) \]

\[ \pi_i = \int_{-\infty}^{d_i} f(t_i, \beta) dt_i \]
\[ f(t_i, \beta) = \frac{\exp((t_i - \beta_0) / \beta_1)}{\beta_1 \left(1 + \exp((t_i - \beta_0) / \beta_1)\right)^2} \]

\[ \text{logit}(\pi_i) = \theta_1 + \theta_2 d_i \]
Log-logistic distribution

\[
f(t_i, \beta) = \frac{\exp((\log(t_i) - \beta_0) / \beta_1)}{t_i \beta_1 (1 + \exp((\log(t_i) - \beta_0) / \beta_1))^2}
\]

\[
\text{logit}(\pi_i) = \theta_1 + \theta_2 \log(d_i)
\]
\[ f(t_i, \beta) = \frac{\exp(-0.5(t_i - \beta_0)^2)}{\sqrt{2\pi} \beta_1} \]

\[ \text{probit}(\pi_i) = \theta_1 + \theta_2 d_i \]
Lognormal distribution

Latent Variable

\[ f(t_i, \beta) = \frac{\exp(-0.5((\log(t_i) - \beta_0)/\beta_1)^2)}{t_i \sqrt{2\pi \beta_1}} \]

\[ \text{probit}(\pi_i) = \theta_1 + \theta_2 \log(d_i) \]

Copyright Adrian Dunne 2005
Gumbel distribution

$$f(t_i, \beta) = \frac{\exp((t_i - \beta_0) / \beta_1) \exp(-\exp((t_i - \beta_0) / \beta_1))}{\beta_1}$$

$$\log(-\log(1 - \pi_i)) = \theta_1 + \theta_2 d_i$$
Log-Gumbel distribution

\[ f(t_i, \beta) = \frac{\exp\left(\frac{\log(t_i) - \beta_0}{\beta_1}\right) \exp\left(-\exp\left(\frac{\log(t_i) - \beta_0}{\beta_1}\right)\right)}{t_i \beta_1} \]

\[ \log(-\log(1 - \pi_i)) = \theta_1 + \theta_2 \log(d_i) \]
Ordinal Data

- Ordered categories e.g. severity of symptoms, none, mild, moderate, severe.

- Ordered categories \( Z = 1, 2, \ldots, K \)

- Probabilities \( \pi_1, \pi_2, \ldots, \pi_K \)
Latent Variable

\[ \alpha_1, \alpha_2 \]

\[ Z = 1 \quad Z = 2 \quad Z = 3 \]
Cumulative Logits

- Cumulative probabilities

\[ F_k = \pi_1 + \pi_2 + \ldots + \pi_k \]

- Cumulative Logits

\[ L_k = \logit(F_k) = \log \left( \frac{F_k}{1 - F_k} \right) \quad k = 1, 2, \ldots, K - 1 \]

- A model for \( L_k \) is a logit model for a binary response.

- We need \( K-1 \) logit models
Cumulative Logits

Based on $K-1$ dichotomizations.

- (1) and (2 to $K$)
- (1 and 2) and (3 to $K$)
- (1 to 3) and (4 to $K$)
- etc.
Proportional Odds Model

• Covariate $x$ influences all cumulative logits equally

$$\text{logit}(F_k) = \alpha_k - f(x, \theta)$$

• Such a model is equivalent to $x$ influencing the location (but not the spread) of the distribution of the latent variable.
Proportional Hazards Model

• Covariate \( x \) influences all cumulative complementary log-logs equally

\[
\log(-\log(1 - F_k)) = \alpha_k - f(x, \theta)
\]

• Such a model is equivalent to \( x \) influencing the location (but not the spread) of the distribution of the latent variable.