Session 2

Binary logistic model

MRC | Hubs for Trials Methodology Research

Edinburgh Hub

MRC | Hubs for Trials Methodology Research

North West Hub

Session 2

Binary Logistic Model

2.1 Dichotomisation of ordinal data to a binary response
2.2 Binary Methods
2.3 Logistic analysis using SAS Proc Logistic
2.4 Logistic analysis using SAS Proc Genmod
2.5 Why use logistic analyses?
2.6 Further example using SAS Proc Logistic

2.1 Dichotomisation of ordinal data to a binary response

Binary data are a special case of ***Ordinal data*** when there are just **two response categories**

e.g.

No Pain	Pain
No Bleeding	Bleeding
No Ulcer	Ulcer

However, even if we have multiple response categories, e.g.

no pain	mild pain	moderate pain	severe pain

these categories can be reduced to a **binary response:-**

no, mild, moderate pain	severe pain

no and mild pain	moderate and severe pain

no pain	some pain

Start by analysing binary data – as all **further methods** are developed from the **binary response**

2.2 Binary methods

Example 2: Outcome following a head injury

Glasgow Outcome Scale Count (%)	Treatment		Total
	Control	Treated	
1: Good recovery	42 (25)	71 (40)	113 (33)
2: Moderate disability	27 (16)	30 (17)	57 (17)
3: Severe disability	33 (20)	27 (15)	60 (18)
4: Vegetative state/Dead	63 (38)	48 (27)	111 (33)
Total	165 (100)	176 (100)	341 (100)

Objective: to relate

Outcome: Favourable = categories 1 and 2
Unfavourable = categories 3 and 4

to

Treatment: 0 = Control
1 = Treated

Baseline age

Standard notation for a 2×2 table

	Control	Treated	Total
Success	s_C	s_T	s
Failure	f_C	f_T	f
Total	n_C	n_T	n

Using Example 2

	Control	Treated	Total
Favourable	69	101	170
Unfavourable	96	75	171
Total	165	176	341

Estimation of difference

(1) Simple proportions

p_i = P(Success; Treatment Group i), i = C,T

Control	**Treated**
$\hat{p}_C = \frac{s_C}{n_C}$	$\hat{p}_T = \frac{s_T}{n_T}$
$\frac{69}{165} = 0.42$	$\frac{101}{176} = 0.57$

(2) Odds ratio: the ***Odds*** of a success for a patient in group T **relative** to the ***Odds*** of a success for a patient in group C

$$\psi = \frac{p_T(1-p_C)}{p_C(1-p_T)}$$

> 1 Group T better
= 1 No difference
< 1 Group T worse

Odds ratio of a favourable outcome in the treated relative to the control group

$$\hat{\psi} = \frac{101 \times 96}{69 \times 75} = 1.874$$

(3) Log odds ratio

denoted by θ : $\theta = \log \psi$

> 0 Group T better
= 0 No difference
< 0 Group T worse

estimated by $\hat{\theta} = \log 1.874 = 0.628$

$$\mathrm{se}\left(\hat{\theta}\right) = \left(\frac{1}{s_C} + \frac{1}{s_T} + \frac{1}{f_C} + \frac{1}{f_T}\right)^{\frac{1}{2}}$$

$$= \left(\frac{1}{69} + \frac{1}{101} + \frac{1}{96} + \frac{1}{75}\right)^{\frac{1}{2}} = 0.2194$$

95% confidence interval for θ

$$\hat{\theta} \pm 1.96\,\mathrm{se}\left(\hat{\theta}\right)$$

0.628 ± 1.96 (0.2194)

(0.198, 1.058)

Hence, 95% CI for ψ

$$\exp\left[\hat{\theta} \pm 1.96\,\mathrm{se}\left(\hat{\theta}\right)\right]$$

(1.22, 2.88)

Hypothesis testing

H_0: $\theta = 0$ vs H_1: $\theta \neq 0$

i.e. $\psi = 1$, $p_C = p_T$ i.e. $\psi \neq 1$, $p_C \neq p_T$

Pearson's chi-square test

$$X^2 = \sum \frac{(O - E)^2}{E} = 8.256 \qquad \text{c.f. } \chi_1^2$$

Significant result p = 0.004

Observed (Expected)	Control	Treated	Total
Favourable	69 (82.26)	101 (87.74)	170
Unfavourable	96 (82.74)	75 (88.26)	171
Total	165	176	341

Approach via efficient score and Fisher's information statistics for log odds ratio θ

Efficient score: Z : measure of group T advantage over group C

$$Z = \frac{s_T f_C - s_C f_T}{n}$$

Fisher's information: V' : amount of information in the data about the group effect

$$V' = \frac{n_C \, n_T \, s \, f}{n^3}$$

$$Z = \frac{s_T f_C - s_C f_T}{n} = \frac{101 \times 96 - 69 \times 75}{341} = 13.258$$

$$V' = \frac{n_C\, n_T\, s\, f}{n^3} = \frac{165 \times 176 \times 170 \times 171}{341^3} = 21.290$$

Test statistic for H_0: $\theta = 0$ $\quad \dfrac{Z^2}{V'}$

$$\frac{Z^2}{V'} = 8.256$$ **(equal to Pearson's chi-square statistic)**

Under H_0: $\quad \dfrac{Z^2}{V'} \sim \chi_1^2$

Maximum likelihood estimate of $\theta \cong \frac{Z}{V'}$

$$\frac{Z}{V'} = \frac{13.258}{21.290} = 0.623$$

Standard error of $\frac{Z}{V'}$ is $\frac{1}{\sqrt{V'}}$

$$\frac{1}{\sqrt{V'}} = \frac{1}{\sqrt{21.290}} = 0.217$$

- Approximate 95% confidence interval for θ

$$\frac{Z}{V'} \pm 1.96 \frac{1}{\sqrt{V'}}$$

0.623 ± 1.96 (0.217)

(0.198, 1.048)

- Approximate 95% confidence interval for ψ

$$\exp\left[\frac{Z}{V'} \pm 1.96 \frac{1}{\sqrt{V'}}\right]$$

(1.22, 2.85)

2.3 Logistic analysis using SAS Proc Logistic

Model: $$\log\left[\frac{p(z_i)}{1-p(z_i)}\right]=\alpha+\beta z_i$$

where $p(z_i)$ = probability of a favourable outcome

$$z_i=\begin{cases}0: \text{ if treat}=0\ (\text{Control})\\ 1: \text{ if treat}=1\ (\text{Treated})\end{cases}$$

	Control	Treated	Total
Favourable	69	101	170
Unfavourable	96	75	171
Total	165	176	341

SAS Proc Logistic program and output are shown in Supplement 2.1

(1) Estimation of difference

Log odds ratio θ for a favourable outcome Treated: Control

$$\theta = \log\left[\frac{p(1)\{1-p(0)\}}{p(0)\{1-p(1)\}}\right] = (\alpha+\beta\times 1)-(\alpha+\beta\times 0)$$

From 2 x 2 table

$$\hat{\theta} = \log\left(\frac{s_1 f_0}{s_0 f_1}\right) = 0.628 \qquad se(\hat{\theta}) = \left(\frac{1}{s_0}+\frac{1}{s_1}+\frac{1}{f_0}+\frac{1}{f_1}\right)^{\frac{1}{2}} = 0.2194$$

Using Efficient score and Fisher's information

$$\hat{\theta} \cong \frac{Z}{V'} = 0.623 \qquad se(\hat{\theta}) \approx \frac{1}{\sqrt{V'}} = 0.217$$

From SAS

$$\hat{\theta} = \hat{\beta} = 0.628 \qquad se(\hat{\theta}) = 0.2194$$

Odds ratio ψ

$$\hat{\psi} = \frac{101 \times 96}{69 \times 75} = 1.874$$

$$\frac{\text{P(Success; Treated)}}{\text{P(Failure; Treated)}} = 1.874\ \frac{\text{P(Success; Control)}}{\text{P(Failure; Control)}}$$

95% CI for ψ

$$\exp[0.628 \pm 1.96(0.2194)]$$

(1.22, 2.88)

(2) Hypothesis testing of H_0: $\theta = 0$

(a) Likelihood ratio test

$$\begin{aligned} D(0) - D(\hat{\theta}) &= -2\ell(0) - -2\ell(\hat{\theta}) \\ &= 472.723 - 464.433 \\ &= 8.290 \quad (\text{c.f. } \chi_1^2) \end{aligned}$$

(b) Score test

$$\frac{Z^2}{V'} = 8.2562 \quad (\text{c.f. } \chi_1^2)$$

Pearson's chi-square statistic

(c) Wald's chi-square

$$\left(\frac{\hat{\theta}}{se\left(\hat{\theta}\right)}\right)^2 = 8.1885$$

Statistically significant difference between treatments

Response variable

- Proc Logistic models the probability of the first ordered value of the response variable as given in the response profile

- Default ordering of response is on formatted labels (if formatted) *otherwise* actual values

  e.g. **D**ead (2)
  Survival (1)

Option **ORDER = INTERNAL**
on MODEL or PROC LOGISTIC statement forces SAS to take order of actual values

Explanatory variables

- Options on CLASS statement for fitting factors

  ORDER = INTERNAL
  Order on actual values not on the default formatted values

  PARAM = REF
  Reference cell parameterisation. The *level* of the variable to use as the reference level can be specified.
  e.g. treat (ref=‘Control') The default is REF=LAST.

- To fit a continuous covariate, include variable in MODEL statement only

- PROC LOGISTIC offers more control of ordering explanatory variables than PROC GENMOD

2.4 Logistic analysis using SAS Proc Genmod

SAS Proc Genmod program and output are shown in Supplement 2.2

2.5 Why use Logistic analyses?

Why do we use *Logistic analysis* rather than:

simple *Pearson's chi-square*
the *Efficient score and Fisher's information?*

- to give a systematic way of investigating the structure of data using a **linear model**

- so that we may adjust for **covariate** prognostic factors

- so that we get a magnitude and a confidence interval for an effect

2.6 Further example using SAS Proc Logistic

To examine the effect of:

- age
- treatment adjusted for age
  on favourable outcome

SAS Proc Logistic program and output are shown in Supplement 2.3

From Proc Logistic output (Supplement 2.3)

(1) Hypothesis testing

Change in deviance due to age

= 472.723 – 464.600

= 8.123 (c.f. χ^2_1)

Change in deviance due **treat** (adjusted for **age)**

= 464.600 – 454.770

= 9.830 (c.f. χ^2_1)

Analysis of deviance table:

Source	df	Deviance
age	1	8.123
treat (adjusted for age)	1	9.830
residual	338	454.770
total	340	472.723

Effect of baseline age is significant (p = 0.004)

Treatment effect is still significant having adjusted for baseline age

(2) Estimation: calculation of log odds ratios

Model: $$\log\left[\frac{p(\underline{z}_i)}{1-p(\underline{z}_i)}\right]=\alpha+\eta(\underline{z}_i)$$

where $$\eta(\underline{z}_i)=\beta_1 z_{i1}+\beta_2 z_{i2}$$

$$z_{i1}=\text{age}$$

$$z_{i2}=\begin{cases}0: \text{ if treat}=0\ (\text{Control})\\ 1: \text{ if treat}=1\ (\text{Treated})\end{cases}$$

$p(\underline{z}_i)$ is probability of a favourable outcome log odds of survival for patient with baseline age = 20 relative to patient with age = 50 receiving the same treatment: θ

$$\log\left[\frac{p(20, z_{i2})}{1 - p(20, z_{i2})}\right] = \alpha + \beta_1 20 + \beta_2 z_{i2}$$

$$\log\left[\frac{p(50, z_{i2})}{1 - p(50, z_{i2})}\right] = \alpha + \beta_1 50 + \beta_2 z_{i2}$$

$$\hat{\theta} = \text{logit}[p(20, z_{i2})] - \text{logit}[p(50, z_{i2})]$$

$$= \hat{\beta}_1(-30) = -0.0226\,(-30) = 0.678$$

$$\hat{\psi} = e^{0.678} = 1.97$$

Odds of a favourable outcome are greater for younger patients