STA 4504 - 5503: Outline of Lecture Notes, ©Alan Agresti

Categorical Data Analysis

1. Introduction

• Methods for *response* (dependent) variable Y having scale that is a set of categories

• *Explanatory* variables may be categorical or continuous or both

Example

Y = vote in election (Democrat, Republican, Independent)

x's - income, education, gender, race

Two types of categorical variables

Nominal - unordered categories

Ordinal - ordered categories

Examples

Ordinal

patient condition (excellent, good, fair, poor)

government spending (too high, about right, too low)

Nominal

transport to work (car, bus, bicycle, walk, . . .)

favorite music (rock, classical, jazz, country, folk, pop)

We pay special attention to

binary variables (*success* - *fail*)

for which nominal - ordinal distinction unimportant.

Probability Distributions for Categorical Data

The *binomial* distribution (and its *multinomial* distribution generalization) plays the role that the *normal* distribution does for continuous response.

Binomial Distribution

• n Bernoulli trials - two possible outcomes for each (*success, failure*)

• $\pi = P(success),\ 1 - \pi = P(failure)$ for each trial

• Y = number of successes out of n trials

• Trials are *independent*

Y has *binomial distribution*

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y},\ y = 0, 1, 2, \ldots, n$$

$y! = y(y-1)(y-2)\cdots(1)$ with $0! = 1$ (factorial)

Example Vote (Democrat, Republican)

Suppose $\pi = prob(Democrat) = 0.50$.

For random sample size $n = 3$, let $y =$ number of Democratic votes

$$p(y) = \frac{3!}{y!(3-y)!}.5^y.5^{3-y}$$

$$p(0) = \frac{3!}{0!3!}.5^0.5^3 = .5^3 = 0.125$$

$$p(1) = \frac{3!}{1!2!}.5^1.5^2 = 3(.5^3) = 0.375$$

y	$P(y)$
0	0.125
1	0.375
2	0.375
3	0.152
	1.0

Note

- $E(Y) = n\pi$

$Var(Y) = n\pi(1-\pi), \quad \sigma = \sqrt{n\pi(1-\pi)}$

- $p = \frac{Y}{n}$ = proportion of success $\quad$ (*also denoted* $\hat{\pi}$)

$$E(p) = E\left(\frac{Y}{n}\right) = \pi$$

$$\sigma\left(\frac{Y}{n}\right) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- When each trial has > 2 possible outcomes, numbers of outcomes in various categories have *multinomial distribution*

Inference for a Proportion

We conduct inferences about parameters using *maximum likelihood*

Definition: The *likelihood function* is the probability of the observed data, expressed as a function of the parameter value.

Example: Binomial, $n = 2$, observe $y = 1$

$$p(1) = \frac{2!}{1!1!}\pi^1(1-\pi)^1 = 2\pi(1-\pi)$$
$$= \ell(\pi)$$

the likelihood function defined for π between 0 and 1

If $\pi = 0$, probability is $\ell(0) = 0$ of getting $y = 1$

If $\pi = 0.5$, probability is $\ell(0.5) = 0.5$ of getting $y = 1$

Definition The *maximum likelihood* (ML) estimate is the parameter value at which the likelihood function takes its maximum.

Example $\ell(\pi) = 2\pi(1 - \pi)$ maximized at $\hat{\pi} = 0.5$

i.e., $y = 1$ in $n = 2$ trials is most likely if $\pi = 0.5$.

ML estimate of π is $\hat{\pi} = 0.50$.

Note

• For binomial, $\hat{\pi} = \frac{y}{n}$ = proportion of successes.

• If $y_1, y_2, \ldots, y_n$ are independent from normal (or many other distributions, such as Poisson), ML estimate $\hat{\mu} = \bar{y}$.

• In ordinary regression ($Y \sim$ normal) "least squares" estimates are ML.

• For large n for any distribution, ML estimates are optimal (no other estimator has smaller standard error)

• For large n, ML estimators have approximate normal sampling distributions (under weak conditions)

ML Inference about Binomial Parameter

$$\hat{\pi} = p = \frac{y}{n}$$

Recall $E(p) = \pi,\ \ \sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$.

• Note $\sigma(p)\ \downarrow$ as $n\ \uparrow$, so
$p \to \pi$ (law of large numbers, true in general for ML)

• p is a sample mean for (0,1) data, so by Central Limit Theorem, sampling distribution of p is approximately normal for large n (True in general for ML)

Significance Test for binomial parameter

$H_o : \pi = \pi_o$

$H_a : \pi \neq \pi_o$ (or 1-sided)

Test statistic

$$z = \frac{p - \pi_o}{\sigma(p)} = \frac{p - \pi_o}{\sqrt{\frac{\pi_o(1-\pi_o)}{n}}}$$

has large-sample standard normal (denoted $N(0, 1)$) null distribution. (Note use null SE for test)

p-value = two-tail probability of results at least as extreme as observed (if null were true)

Confidence interval (CI) for binomial parameter

Definition Wald CI for a parameter θ is
$\hat{\theta} \pm z_{\frac{\alpha}{2}}\ (SE)$

(e.g, for 95% confidence, estimate plus and minus 1.96 estimated standard errors, where $z_{.025} = 1.96$)

Example $\theta = \pi, \quad \hat{\theta} = \hat{\pi} = p$

$\sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$ estimated by

$SE = \sqrt{\frac{p(1-p)}{n}}$

95% CI is $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$

Note Wald CI often has poor performance in categorical data analysis unless n quite large.

Example: Estimate π = population proportion of vegetarians

For $n = 20$, we get $y = 0$

$p = \frac{0}{20} = 0.0$

95% CI: $0 \pm 1.96\sqrt{\frac{0\times 1}{20}}$

$= 0 \pm 0,$

$= (0, 0)$

• Note what happens with Wald CI for π if $p = 0$ or 1

• *Actual* coverage probability much less than 0.95 if π near 0 or 1.

• Wald 95% CI = set of π_o values for which p-value $>$ .05 in testing

$$H_o : \pi = \pi_o \qquad H_a : \pi \neq \pi_o$$

using

$z = \frac{p - \pi_o}{\sqrt{\frac{p(1-p)}{n}}}$ (denominator uses *estimated* SE)

Definition *Score test, score CI* use null SE

e.g. Score 95% CI = set of π_o values for which p-value > 0.05 in testing

$$H_o : \pi = \pi_o \quad H_a : \pi \neq \pi_o$$

using

$$z = \frac{p - \pi_o}{\sqrt{\frac{\pi_o(1-\pi_o)}{n}}} \quad \leftarrow \text{note null } SE \text{ in denominator}$$

(known, not estimated)

Example π = probability of being vegetarian

$y = 0, \qquad n = 20, \qquad p = 0$

What π_o satisfies

$$\pm 1.96 = \frac{0 - \pi_o}{\sqrt{\frac{\pi_o(1-\pi_o)}{20}}} \; ?$$

$$1.96\sqrt{\frac{\pi_o(1-\pi_o)}{20}} = |0 - \pi_o|$$

$\pi_o = 0$ is one solution

solve quadratic $\rightarrow \pi_o = .16$ other solution

95% score CI is (0. 0.16), more sensible than Wald CI of (0, 0).

• When solve quadratic, can show midpoint of 95% CI is

$$\frac{y+1.96^2/2}{n+1.96^2} \approx \frac{y+2}{n+4}$$

• Wald CI $p \pm 1.96\sqrt{p(1-p)/n}$ also works well if add 2 successes, add 2 failures before applying (this is the "Agresti-Coull method")

• For inference about proportions, *score* method tends to perform better than *Wald* method, in terms of having actual error rates closer to the advertised levels.

• Another good test, CI uses the *likelihood function*

(e.g. CI = values of π for which $\ell(\pi)$ close to $\ell(\hat{\pi})$

= values of π_o not rejected in "likelihood-ratio test")

• For small n, inference uses actual binomial sampling dist. of data instead of normal approx. for that dist.

2. Two-Way Contingency Tables

Example: Physicians Health Study *(5 year)*

		HEART ATTACK		
		Yes	No	Total
GROUP	Placebo	189	10,845	11,034
	Aspirin	104	10,933	11,037

↗

2x2 table

Contingency table - cells contain counts of outcomes. $I \times J$ table has I rows, J columns.

A *conditional dist* refers to prob. dist. of Y at fixed level of x.

Example:

		Y		
		Yes	No	Total
X	Placebo	.017	.983	1.0
	Aspirin	.009	.991	1.0

Sample conditional dist. for placebo group is

$$.017 = \frac{189}{11,034}, \quad .983 = \frac{10,845}{11,034}$$

Natural way to look at data when

Y = response var.

X = explanatory var.

Example: Diagnostic disease tests

$Y =$ outcome of test: $\quad 1 =$ positive $\quad 2 =$ negative

$X =$ reality: $\quad 1 =$ diseased $\quad 2 =$ not diseased

		Y	
		1	2
X	1		
	2		

sensitivity $= P(Y = 1|X = 1)$

specificity $= P(Y = 2|X = 2)$

If you get positive result, more relevant to you is $P(X = 1|Y = 1)$. This may be low even if sensitivity, specificity high. (See pp. 23-24 of text for example of how this can happen when disease is relatively rare.)

What if X, Y both *response* var's?

$\{\pi_{ij}\} = \{P(X = i, Y = j)\}$ form the *joint distribution* of X and Y.

π_{11}	π_{12}	π_{1+}
π_{21}	π_{22}	π_{2+}
π_{+1}	π_{+2}	1.0

marginal probabilities

Sample cell counts $\{n_{ij}\}$

cell proportions $\{p_{ij}\}$

$p_{ij} = \frac{n_{ij}}{n}$ with $n = \sum_i \sum_j n_{ij}$

Definition X and Y are *statistically independent* if true conditional dist. of Y is identical at each level of x.

		Y	
X	.01	.99	
	.01	.99	

Then, $\pi_{ij} = \pi_{i+}\pi_{+j}$ all i, j

i.e., $P(X = i, Y = j) = P(X = i)P(Y = j)$, such as

		Y	
	.28	.42	.7
X	.12	.18	.3
	.4	.6	1.0

Comparing Proportions in 2x2 Tables

		Y	
		S	F
X	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

Conditional Distributions

$\hat{\pi}_1 - \hat{\pi}_2 = p_1 - p_2$

$$SE(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Example: $p_1 = .017$, $p_2 = .009$, $p_1 - p_2 = .008$

$$SE = \sqrt{\frac{.017 \times .983}{11,034} + \frac{.009 \times .991}{11,037}} = .0015$$

95% CI for $\pi_1 - \pi_2$ is $.008 \pm 1.96(.0015) = (.005, .011)$.

Apparently $\pi_1 - \pi_2 > 0$ $(i.e., \pi_1 > \pi_2)$.

Relative Risk $= \frac{\pi_1}{\pi_2}$

Example: Sample $\frac{p_1}{p_2} = \frac{.017}{.009} = 1.82$

Sample proportion of heart attacks was 82% higher for placebo group.

Note

- See p. 58 of text for SE formula
- SAS provides CI for π_1/π_2.

Example: 95% CI is (1.43, 2.31)

- Independence $\Leftrightarrow \frac{\pi_1}{\pi_2} = 1.0$.

Odds Ratio

		S	F
Group	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

The *odds* the response is a S instead of an $F = \frac{prob(S)}{prob(F)}$

$= \pi_1/(1-\pi_1)$ in row 1.

$= \pi_2/(1-\pi_2)$ in row 2.

e.g., if odds $= 3$, S three times as likely as F.

e.g., if odds $= \frac{1}{3}$, F three times as likely as S.

$$\text{Odds} = 3 \Rightarrow P(S) = \tfrac{3}{4},\ P(F) = \tfrac{1}{4}$$

$$P(S) = \frac{odds}{1 + odds}$$

$$\text{odds} = \tfrac{1}{3} \Rightarrow P(S) = \tfrac{1/3}{1+1/3} = \tfrac{1}{4}$$

Definition: Odds Ratio

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Example

		Yes	No	Total
Heart Attack	Placebo	189	10,845	11,034
	Aspirin	104	10,933	11,037

Sample Proportions

p_1	$1-p_1$		.0171 .9829	1.0
		=		
p_2	$1-p_2$		.0094 .9906	1.0

Sample odds =

$$\frac{.0171}{.9829} = \frac{189}{10,845} = .0174, \text{ placebo}$$
$$= \frac{104}{10,933} = .0095, \text{ aspirin}$$

Sample odds Ratio

$$\hat{\theta} = \frac{.0174}{.0095} = 1.83$$

The odds of a heart attack for placebo group were 1.83 time odds for aspirin group (i.e., 83% higher)

Properties of odds ratio

• Each odds $\geqslant 0$, and $\theta \geqslant 0$.

• $\theta = 1$ when $\pi_1 = \pi_2$; i.e., response independent of group

• The farther θ falls from 1, the stronger the association

(For Y = lung cancer, some stufies have $\theta \approx 10$ for X = smoking, $\theta \approx 2$ for X = passive smoking)

• If rows interchanged, or if columns interchanged, $\theta \to 1/\theta$.

e.g. $\theta = 3, \theta = \frac{1}{3}$ represent same strength of association but in opposite directions.

• For counts

S	F
n_{11}	n_{12}
n_{21}	n_{22}

$\hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

$=$ <u>cross-product ratio</u>

(Yule 1900) (strongly criticized by K. Pearson!)

• Treats X, Y symmetrically

		Placebo	Aspirin
Heart Attack	Yes		
	No		

$\rightarrow \hat{\theta} = 1.83$

• $\theta = 1 \quad \Leftrightarrow \quad \log\theta = 0$

log odds ratio symmetric about 0

e.g., $\theta = 2 \Rightarrow \log\theta = .7$

$\theta = 1/2 \Rightarrow \log\theta = -.7$

• Sampling dist. of $\hat{\theta}$ skewed to right, $\approx$ normal only for very large n.

Note: We use "natural logs" (LN on most calculators)

This is the log with base $e = 2.718...$

• Sampling dist. of $\log\hat{\theta}$ is closer to normal, so construct CI for $\log\theta$ and then exponentiate endpoints to get CI for θ.

Large-sample (asymptotic) standard error of $\log\hat{\theta}$ is

$$SE(\log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

CI for $\log\theta$ is

$$\log\hat{\theta} \pm z_{\frac{\propto}{2}} \times SE(\log\hat{\theta})$$

(e^L, e^U) is CI for θ

Example: $\hat{\theta} = \frac{189 \times 10{,}933}{104 \times 10{,}845} = 1.83$

$\log \hat{\theta} = .605$

$$SE(\log \hat{\theta}) = \sqrt{\frac{1}{189} + \frac{1}{10,933} + \frac{1}{104} + \frac{1}{10,845}} = .123$$

95% CI for $\log \theta$ is

$.605 \pm 1.96(.123) = (.365, .846)$

95% CI for θ is

$(e^{.365}, e^{.846}) = (1.44, 2.33)$

Apparently $\theta > 1$

e denotes *exponential* function

$e^0 = 1,\ e^1 = e = 2.718\ldots$

$e^{-1} = \frac{1}{e} = .368$

$e^x > 0$ all x

exp fn. = antilog for natural log scale ℓn

$e^0 = 1$ means $\log_e(1) = 0$

$e^1 = 2.718$ $\log_e(2.718) = 1$

$e^{-1} = .368$ $\log_e(.368) = -1$

$\log_e(2) = .693$ means $e^{.693} = 2$

Notes

• $\hat{\theta}$ not midpoint of CI, because of skew

• If any $n_{ij} = 0$, $\hat{\theta} = 0$ or ∞, and better estimate and SE results by replacing $\{n_{ij}\}$ by $\{n_{ij} + .5\}$.

• When π_1 and π_2 close to 0

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \approx \frac{\pi_1}{\pi_2}$$

the relative risk

Example: Case-control study in London hospitals (Doll and Hill 1950)

X = smoked $\geqslant 1$ cigarette per day for at least 1 year?

Y = Lung Cancer

		Lung Cancer	
		Yes	No
X	Yes	688	650
	No	21	59
		709	709

Case control studies are "retrospective." Binomial sampling model applies to X (sampled within levels of Y), not to Y .

Cannot estimate $P(Y = yes|x)$,

or $\pi_1 - \pi_2 =$
$P(Y = yes|X = yes) - P(Y = yes|X = no)$

or π_1/π_2

We *can* estimate $P(X|Y)$, so can estimate θ.

$$\begin{aligned}\hat{\theta} &= \frac{\hat{P}(X = yes|Y = yes)/\hat{P}(X = no|Y = yes)}{\hat{P}(X = yes|Y = no)/\hat{P}(X = no|Y = no)} \\ &= \frac{(688/709)/(21/709)}{(650/709)/(59/709)} \\ &= \frac{688 \times 59}{650 \times 21} = 3.0\end{aligned}$$

Odds of lung cancer for smokers were 3.0 times odds for non-smokers.

In fact, if $P(Y = yes|X)$ is near 0, then $\theta \approx \pi_1/\pi_2 =$ rel. risk, and can conclude that *prob.* of lung cancer is $\approx$ 3.0 times as high for smokers as for non-smokers.

Chi - Squared Tests of Independence

Example

JOB SATISFACTION

INCOME	Very Dissat.	Little Satis.	Mod. Satis.	Very Satis	
< 5000	2	4	13	3	22
5000-15,000	2	6	22	4	34
15,000-25,000	0	1	15	8	24
$> 25{,}000$	0	3	13	8	24
	4	14	63	23	104

Data from General Social Survey (1991)

H_o: X and Y independent

H_a: X and Y dependent

H_o means

$$P(X = i, Y = j) = P(X = i)P(Y = j)$$
$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

Expected frequency $\mu_{ij} = n\pi_{ij}$
= mean of dist. of cell count n_{ij}
= $n\pi_{i+}\pi_{+j}$ under H_o.

ML estimates $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$

$= n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$

called *estimated expected frequencies*

Test Statistic

$$X^2 = \sum_{all\ \ cells} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

called Pearson chi - squared statistic (Karl Pearson, 1900)

X^2 has large-sample chi-squared dist. with $df = (I-1)(J-1)$

I = number of rows, J = number of columns

P-value = $P(X^2 \geqslant X^2$ observed)

= right - tail prob.

(Table on p. 343 text)

Example: Job satisfaction and income

$$X^2 = 11.5$$

$$df = (I-1)(J-1) = 3 \times 3 = 9$$

Evidence against H_o is weak.

Plausible that job satisfaction and income are independent.

Note

• Chi-squared dist. has $\mu = df$, $\sigma = \sqrt{2df}$, more bell-shaped as $df \uparrow$

• *Likelihood-ratio* test stat.

$$G^2 = 2 \sum n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right)$$

$$= -2 \log \left[\frac{maximize\ likelihood\ when\ H_o\ true}{maximize\ likelihood\ generally} \right]$$

G^2 also is approx. $\chi^2, df = (I-1)(J-1)$.

Example: $G^2 = 13.47, df = 9$, P-value $= .14$

• df for X^2 test
= no. parameters in general - no. parameters under H_o

Example: Indep. $\pi_{ij} = \pi_{i+}\pi_{+j}$

$df = (IJ - 1) - [(I - 1) + (J - 1)]$

$\sum \pi_{ij} = 1 \qquad \sum \pi_{i+} = 1 \qquad \sum \pi_{+j} = 1$

$= (I - 1)(J - 1) \leftarrow$ Fisher 1922 (not Pearson 1900)

• $X^2 = G^2 = 0$ when all $n_{ij} = \hat{\mu}_{ij}$.

• As $n \uparrow$, $X^2 \to \chi^2$ faster than $G^2 \to \chi^2$, usually close if most $\hat{\mu}_{ij} \geqslant 5$

• These tests treat X, Y as *nominal*. Reorder rows columns, X^2, G^2 unchanged

Sec. 2.5 (we skip) presents *ordinal* tests. We re-analyze with ordinal model in Ch. 6 (more powerful, much smaller P-value).

Standardized (Adjusted) Residuals

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$$

Under H_o: indep., $r_{ij} \approx$ std. normal $N(0,1)$

so $|r_{ij}| > 2$ or 3 represents cell that provides strong evidence against H_o

Example: $n_{44} = 8$, $\hat{\mu}_{44} = \frac{24 \times 23}{104} = 5.31$

$$r_{44} = \frac{8 - 5.31}{\sqrt{5.31(1 - \frac{24}{104})(1 - \frac{23}{104})}} = 1.51$$

None of cells show much evidence of association.

Example

		RELIGIOSITY			
		Very	Mod.	Slightly	Not
GENDER	Female	170	340	174	95
		(3.2)	(1.0)	(-1.1)	(-3.5)
	Male	98	266	161	123
		(-3.2)	(-1.0)	(1.1)	(3.5)

General Social Survey data (variables Sex, Relpersn)

$X^2 = 20.6, G^2 = 20.7, df = 3,$ P-value $= 0.000$

• SAS (PROC GENMOD) also provides "Pearson residuals" (label reschi)

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

which are simpler but less variable than $N(0, 1)$.

$(\sum e_{ij}^2 = X^2)$

SPSS

ANALYZE menu

CROSSTABS suboption

click on STATISTICS

options include X^2 test

click on CELLS

"adjusted standardized"

gives standardized residuals

When enter data as contingency table

Income	Satis.	Count
1	1	2
1	2	4
.	.	.
.	.	.

Select WEIGHT CASES option on DATA menu,

tell SPSS to weight cases by count

STATA and SAS

See www.ats.ucla.edu/stat/examples/icda

for sample programs for examples from 1st edition of text.

R

link to Laura Thompson manual

pp. 35-38 for chi-squared test, standardized residuals,

function chisq.test

Partitioning Chi-squared

$\chi^2_a + \chi^2_b = \chi^2_{a+b}$ for indep. chi-squared stat's.

Example: $G^2 = 13.47$, $X^2 = 11.52$, $df = 9$

Compare income levels on job satisfaction

Income Job Satisfac.

	VD	LS	MS	VS
< 5	2	4	13	3
5 -15	2	6	22	4

	VD	LS	MS	VS
15-25	0	1	15	8
> 25	0	3	13	8

	VD	LS	MS	VS
< 15	4	10	35	7
> 15	0	4	28	16

X^2	G^2	df
.30	.30	3
1.14	1.19	3
10.32	11.98	3
(P = .02	P = .01)	
11.76	13.47	9

Note

• Job satisfaction appears to depend on whether income $>$ or $<$ \$ 15,000

• G^2 exactly partitions, X^2 does not

• Text gives guidelines on how to partition so separate components indep., which is needed for G^2 to partition exactly.

Small-sample test of indep.

2 x 2 case (Fisher 1935)

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n

Exact null dist. of $\{n_{ij}\}$, based on fixed row and column totals, is

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$$

Where $\binom{a}{b} = \frac{a!}{b!(a-b)!}$

Hypergeometric dist.

Example Tea Tasting (Fisher)

		GUESS		
		Milk	Tea	
Poured First	Milk	?		4
	Tea			4
		4	4	8

n_{11} = 0,1,2,3, or 4

$$\begin{array}{|cc|}\hline 4 & 0 \\ 0 & 4 \\ \hline\end{array}$$

has prob.

$$P(4) = \frac{\binom{4}{4}\binom{4}{4-4}}{\binom{8}{4}} = \frac{\left(\frac{4!}{4!0!}\right)\left(\frac{4!}{0!4!}\right)}{\left(\frac{8!}{4!4!}\right)}$$

$$= \frac{4!4!}{8!} = \frac{1}{70} = .014$$

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} = .229$$

n_{11}	$P(n_{11})$
0	.014
1	.229
2	.514
3	.229
4	.014

For 2 x 2 tables,

H_o: indep $\Leftrightarrow$ $H_o : \theta = 1$ for $\theta =$ odds ratio

For $H_o : \theta = 1, \qquad H_a : \theta > 1,$

P-value $= P(\hat{\theta} \geqslant \hat{\theta}_{obs})$
$= .229 + .014 = .243.$

Not much evidence against H_o

Test using hypergeometric called *Fisher's exact test.*

For $H_a : \theta \neq 1$, P-value =

two-tail prob. of outcomes no more likely than observed

Example: P-value $= P(0) + P(1) + P(3) + P(4) = .486$

Note:

• Fisher's exact test extends to $I \times J$ tables (P-value $= .23$ for job sat. and income)

• If make conclusion, e.g., rejecting H_o if $p \leq \alpha = .05$, actual $P(\text{type } I \text{ error}) < .05$ because of discreteness (see text)

Three - Way Contingency Tables

Example: FL death penalty court cases

Victim's Race	Defendant's Race	Death Penalty Yes	No	% Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8

Y = death penalty (response var.)
X = defendant's race (explanatory)
Z = victim's race (control var.)

53	414
11	37

0	16
4	139

are *partial tables*

They *control* (hold constant) Z

The *conditional odds ratios* are:

$Z = \text{white} : \hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = .43$

$Z = \text{black} : \hat{\theta}_{XY(2)} = 0.00$ (.94 after add .5 to cells)

Controlling for victim's race, odds of receiving death penalty were *lower* for white defendants than black defendants.

Add partial tables $\rightarrow XY$ *marginal* table

	Yes	No
W	n_{11}	n_{12}
B	n_{21}	n_{22}

$\hat{\theta}_{XY} = 1.45$

Ignoring victim's race, odds of death penalty *higher* for white defendant's.

Simpson's Paradox: All partial tables show reverse assoc. from that in marginal table.

Cause ?

Moral ? Can be dangerous to "collapse" contingency tables.

Def. X and Y are *conditionally independent given* Z, if they are independent in each partial table.

In $2 \times 2 \times K$ table,

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.0$$

Note Does not imply X and Y indep. in marginal two-way table

Example

Clinic Z	Treatment X	Response Y S	Y F	θ
1	A	18	12	1.0
	B	12	8	
2	A	2	8	1.0
	B	8	32	
Marginal	A	20	20	2.0
	B	20	40	

3. Generalized Linear Models

Components of a GLM

1. *Random Component*

Identify response var. Y

Assume independent observ's $y_1, \ldots, y_n$ from particular form of dist., such as Poisson or binomial

Model how $\mu_i = E(Y_i)$ depends on explanatory var's

2. *Systematic component*

Pick explanatory var's $x_1, \ldots, x_k$ for *linear predictor*

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

3. *Link function*

Model function $g(\mu)$ of $\mu = E(Y)$ using

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

g is the *link function*

Example

- $\log(\mu) = \alpha + \beta_1 x_1 + \ldots$ uses $g(\mu) = \log(\mu)$.

log link often used for a "count" random component, for which $\mu > 0$.

- $\log(\frac{\mu}{1-\mu}) = \alpha + \beta_1 x_1 + \ldots$ uses $g(\mu) = \log(\frac{\mu}{1-\mu})$, the *logit* link.

Often used for binomial, with $\mu = \pi$ between 0 and 1

(logit = log of odds)

- $\mu = \alpha + \beta_1 x_1 + \ldots$ uses $g(\mu) = \mu$, *identity* link

e.g., ordinary regression for normal response.

Note:

• A GLM generalizes ordinary regression by

1. permitting Y to have a dist. other than normal

2. permitting modeling of $g(\mu)$ rather than μ.

• The same ML (max. likelihood) fitting procedure applies to all GLMs. This is basis of software such as PROC GENMOD in SAS.

(Nelder and Wedderburn, 1972)

GLMs for Binary Data

Suppose $Y = 1$ or 0

Let $P(Y = 1) = \pi$, "Bernoulli trial"

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

This is binomial for $n = 1$ trial

$E(Y) = \pi$

$Var(Y) = \pi(1 - \pi)$

For explan. var. x, $\pi = \pi(x)$ varies as x varies.

Linear probability model $\pi(x) = \alpha + \beta x$

This is a GLM for binomial random component and identity link fn.

$Var(Y) = \pi(x)[1 - \pi(x)]$ varies as x varies, so least squares not optimal.

Use ML to fit this and other GLMs.

ex. Y = infant sex organ malformation

1 = present, 0 = absent

x = mother's alcohol consumption
(average drinks per day)

Alcohol	Malformation			Proportion
Consumption	Present	Absent	Total	Present
0	48	17,066	17,114	.0028
< 1	38	14,464	14,502	.0026
1-2	5	788	793	.0063
3-5	1	126	127	.0079
$\geqslant 6$	1	37	38	.0262

Using x scores (0, .5, 1.5, 4.0, 7.0), linear prob. model for $\pi =$ prob. malformation present has ML fit

$$\hat{\pi} = \hat{\alpha} + \hat{\beta}x = .0025 + .0011x$$

At $x = 0$, $\hat{\pi} = \hat{\alpha} = .0025$

$\hat{\pi}$ increases by $\hat{\beta} = .0011$ for each 1-unit increase in alcohol consumption.

Note

• ML estimates $\hat{\alpha}$, $\hat{\beta}$ obtained by iterative numerical optimization.

• To test $H_o : \beta = 0$ (independence), can use

$$z = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

(for large n has approx std. normal dist. under null)

ex. $z = \frac{.0011}{.0007} = 1.50$

For $H_a : \beta \neq 0$, P-value $= 0.13$

Or, z^2 approx. χ^2_1 (ex. $z^2 = 2.24$)

• Could use Pearson X^2 (or G^2) to test indep., but ignores ordering of rows

• Alternative way to apply X^2 (or deviance G^2) is to test fit of model. (see printout)

(compare counts to values predicted by linear model)

• Same fit if enter 5 binomial "success totals" or the 32,574 individual binary responses of 0 or 1.

• Model $\pi(x) = \alpha + \beta x$ can give $\hat{\pi} > 1$ or $\hat{\pi} < 0$

More realistic models are *nonlinear* in shape of $\pi(x)$.

Logistic regression model

$$\log\left[\frac{\pi}{1-\pi}\right] = \alpha + \beta x$$

is GLM for binomial Y with logit link

ex. $logit(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -5.96 + .32x$

$\hat{\pi} \uparrow$ as $x \uparrow$, and P-value $=.012$ for $H_o : \beta = 0$.

(but, $P = .30$ if delete "present" obs. in $\geqslant 6$ drinks!!)

Note

• Chap. 4 studies this model

• For contingency table, one can test H_o : model fits, using estimated expected frequencies that satisfy the model, with X^2, G^2 test stat.'s.

ex. $X^2 = 2.05, G^2 = 1.95$ for H_o : logistic regr. model

$df = 3 = 5$ binom. observ. - 2 parameters (P-value large, no evidence against H_o)

• Both the linear probability model and logistic regression model fit adequately

How can this be?

logistic $\approx$ linear when $\hat{\pi}$ near 0 or near 1 for all observed x.

GLMs for count data

When Y is a count (0,1,2,3,...) traditional to assume *Poisson* dist.

$$P(y) = \frac{e^{-\mu}\mu^y}{y!},\ y = 0, 1, 2, \ldots$$

- $\mu = E(y)$

$\mu = Var(Y),\ \sigma = \sqrt{\mu}$

- In practice often $\sigma^2 > \mu$, greater variation than Poisson predicts (overdispersion)

- *Negative binomial* dist. has separate σ^2 parameter and permits overdispersion.

Poisson regression for count data

Suppose we assume Y has Poisson dist., x an explanatory var.

Model $\mu = \alpha + \beta x$ identity link

or $\log(\mu) = \alpha + \beta x$ log link

loglinear model (Ch. 7 for details about this link)

ex. Y = no. defects on silicon wafer

x = treatment ($1 = B$, $0 = A$) dummy (indicator) var.

10 wafers for each

$A: \ 3, 7, 6, \ldots$ $\bar{y}_A = 5.0$

$B: \ 6, 9, 8, \ldots$ $\bar{y}_B = 9.0$

For model $\mu = \alpha + \beta x$ (identity link)

$\hat{\mu} = 5.0 + 4.0x$

$x = 0: \ \hat{\mu}_A = 5.0 \qquad (= \bar{y}_A)$

$x = 1: \ \hat{\mu}_B = 9.0 \qquad (= \bar{y}_B)$

$\hat{\beta} = 4.0 = \hat{\mu}_B - \hat{\mu}_A$ has $SE = 1.18$

($\leftarrow$ test, CI for β)

For loglinear model $\log(\mu) = \alpha + \beta x$

$\log(\hat{\mu}) = 1.609 + .588x$

$x = 0: \ \log \hat{\mu}_A = 1.609, \ \hat{\mu}_A = e^{1.609} = 5.0$

$x = 1: \ \log \hat{\mu} = 1.609 + .588 = 2.197, \hat{\mu}_B = e^{2.197} = 9.0$

Inference for GLM parameters

CI: $\hat{\beta} \pm z_{\frac{\alpha}{2}}(SE)$

Test: $H_o : \beta = 0$

1. Wald test

$z = \frac{\hat{\beta}}{SE}$ has approx. $N(0,1)$ dist.

For $H_a : \beta \neq 0$, can also use Wald stat.

$z^2 = \left(\frac{\hat{\beta}}{SE}\right)^2$ is approx. χ^2_1.

CI = set of β_o values for $H_o : \beta = \beta_o$ such that

$|\hat{\beta} - \beta_o|/SE \ < z_{\alpha/2}$

2. Likelihood-ratio test

$\ell_0 =$ maximized likelihood when $\beta = 0$

$\ell_1 =$ maximized likelihood for arbitrary β

Test stat. $= -2 \log \left(\frac{\ell_0}{\ell_1} \right)$

$= -2 \log \ell_0 - (-2 \log \ell_1)$

$= -2(L_0 - L_1)$

Where L = maximized *log* likelihood

ex. Wafer defects

Loglinear model $\log(\mu) = \alpha + \beta x$

$\beta = \log \mu_B - \log \mu_A$

$H_O : \mu_A = \mu_B \Leftrightarrow \beta = 0$

Wald test

$z = \frac{\hat{\beta}}{SE} = \frac{.588}{.176} = 3.33$

$z^2 = 11.1,\ df = 1,\ P = .0009$ for $H_a : \beta \neq 0.$

Likelihood-ratio test

$L_1 = 138.2, L_0 = 132.4$

Test stat. $-2(L_0 - L_1) = 11.6,\ df = 1\ P = .0007$

PROC GENMOD reports LR test result with 'type 3' option

Note

• For very large n, Wald test and likelihood ratio test are approx. equivalent, but for small to moderate n the LR test is more reliable and powerful.

• LR stat. also equals difference in "deviances," goodness-of-fit stats.

ex. 27.86 - 16.27 = 11.59

• LR method also extends to CIs:

$100(1-\alpha)\%$ CI = set of β_o in $H_o : \beta = \beta_o$ for which P-value $> \alpha$ in LR test.

(i.e., do not reject H_o at α- level)

GENMOD: LRCI option

$$\beta = \log \mu_B - \log \mu_A = \log \left(\frac{\mu_B}{\mu_A} \right)$$

$$e^{\beta} = \frac{\mu_B}{\mu_A}$$

$$e^{\hat{\beta}} = e^{.5878} = 1.8 = \frac{\hat{\mu}_B}{\hat{\mu}_A}.$$

95% CI for β is $.588 \pm 1.96\ (.176) = (.242,\ .934)$.

95% CI for $e^{\beta} = \frac{\mu_B}{\mu_A}$ is

$(e^{.242},\ e^{.934}) = (1.27,\ 2.54)$.

We're 95% confident that μ_B is between 1.27 and 2.54 times μ_A.

CI based on likelihood-ratio test is (1.28, 2.56).

Deviance of a GLM

The *saturated model* has a separate parameter for each observation and has the perfect fit $\hat{\mu}_i = y_i$

For a model M with maximized log likelihood L_M,

deviance = -2$[L_M - L_S]$, where S = saturated model

= LR stat. for testing that all parameters *not* in M equal 0.

i.e., for

H_o : model holds

H_a : saturated model

For Poisson and binomial models for counts,

$$\text{Deviance} = G^2 = 2\sum y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) \leftarrow \text{for } M$$

When $\hat{\mu}_i$ are large and no. of predictor settings fixed, G^2 and

$$X^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \text{ (Pearson)}$$

are used to test goodness of fit of model
(i.e., H_o : model holds).

They are approx. χ^2 with
df = no. observations - no. model parameters

ex. Wafer defects

$\hat{\mu}_i = 5$ for 10 observ's in Treatment A

$\hat{\mu}_i = 9$ for 10 observ's in Treatment B

For loglinear model, $\log \mu = \alpha + \beta x$

deviance $G^2 = 16.3$

Pearson $X^2 = 16.0$

df = 20-2 = 18

These do not contradict H_O : model holds,

but their use with chi-square dist. is questionable

• $\hat{\mu}_i$ not that large

• theory applies for *fixed* df as $n \uparrow$ (happens with contingency tables)

Note

• For GLMs one can study lack of fit using residuals (later chapters)

• Count data often show *overdispersion* relative to Poisson GLMs.

i.e., at fixed x, sample variance $>$ mean, whereas var. $=$ mean in Poisson.

(often caused by subject *hetrogeneity*)

ex. Y = no. times attended religious services in past year.

Suppose $\mu = 25$. Is $\sigma^2 = 25$ ($\sigma = 5$)?

Negative binomial GLM

More flexible model for count that has

$E(Y) = \mu,\ Var(Y) = \mu + D\mu^2$

where D called a *dispersion para.*

As $D \ \to 0$, neg. bin. $\to$ Poisson.

(Can derive as "gamma dist. mixture" of Poissons, where the Poisson mean varies according to a gamma dist.)

ex. GSS data "In past 12 months, how many people have you known personally that were victims of homicide?"

Y	Black	White
0	119	1070
1	16	60
2	12	14
3	7	4
4	3	0
5	2	0
6	0	1

Model $\log(\mu) = \alpha + \beta x$

Black: $\bar{y} = .52,\ s^2 = 1.15$

White: $\bar{y} = .09,\ s^2 = .16$

For Poisson or neg. bin. model,

$\log \hat{\mu} = -2.38 + 1.73x$

$e^{1.73} = 5.7 = \frac{.522}{.092} = \frac{\bar{y}_B}{\bar{y}_W}$

However, SE for $\hat{\beta} = 1.73$ is .147 for Poisson, .238 for neg. bin.

Wald 95% CI for $e^{\beta} = \mu_B/\mu_W$ is

Poisson: $e^{1.73 \pm 1.96(.147)} = (4.2,\ 7.5)$

Neg bin: $e^{1.73 \pm 1.96(.238)} = (3.5,\ 9.0)$

In accounting for overdispersion, neg. bin. model has wider CIs.

LR CIs are $(e^{1.444},\ e^{2.019}) = (4.2,\ 4.7)$ and $(3.6,\ 9.2)$

For neg. bin. model, estimated dispersion para. $\hat{D} = 4.94 \qquad (SE = 1.0)$

$\hat{Var}(Y) = \hat{\mu} + \hat{D}\hat{\mu}^2 = \hat{\mu} + 4.94\hat{\mu}^2$

strong evidence of overdispersion

When Y is a count, safest strategy is to use negative binomial GLM, especially when dispersion para. is significantly > 0.

Models for Rates

When y_i have different bases

(e.g., no. murders for cities with different popul. sizes)

more relevant to model *rate* at which events occur.

Let $y =$ count with index t

Sample rate $\frac{y}{t}$, $E\left(\frac{Y}{t}\right) = \frac{\mu}{t}$.

Loglinear model $\log\left(\frac{\mu}{t}\right) = \alpha + \beta x$

or $\log(\mu) - \log(t) = \alpha + \beta x$

See text pp. 82-84 for discussion

4. Logistic Regression

$Y = 0$ or 1

$\pi = P(Y = 1)$

$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$

$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right]$

Uses "logit" link for binomial Y. Equivalently,

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

where $\exp(\alpha + \beta x) = e^{\alpha+\beta x}$.

Properties

• Sign of β indicates whether $\pi(x) \uparrow$ or $\downarrow$ as $x \uparrow$

• If $\beta = 0, \pi(x) = \frac{e^{\alpha}}{1+e^{\alpha}}$ constant as $x \uparrow$ ($\pi > \frac{1}{2}$ if $\alpha > 0$)

• Curve can be approximated at fixed x by straight line to describe rate of change.

e.g., at x with $\pi(x) = \frac{1}{2}$, slope $= \beta\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{\beta}{4}$.

at x with $\pi(x) = 0.1$ or 0.9, slope $= \beta\,(0.1)\,(0.9) = 0.09\beta$

Steepest slope where $\pi(x) = \frac{1}{2}$.

• When $\pi = \frac{1}{2}, \log\left[\frac{\pi}{1-\pi}\right] = \log\left[\frac{0.5}{0.5}\right] = \log(1) = 0 = \alpha + \beta x \longrightarrow x = -\frac{\alpha}{\beta}$ is the x value where this happens

• $\frac{1}{\beta} \approx$ distance between x values with $\pi = 0.5$ and $\pi = 0.75$ (or 0.25).

• ML fit obtained with iterative numerical methods.

Ex. Horseshoe crabs

$Y = 1$ if female crab has satellites.

$Y = 0$ if no satellite.

$x =$ weight (kg) $(\bar{x} = 2.44, s = 0.58)$

$n = 173$

ML fit: $\text{logit}[\hat{\pi}(x)] = -3.69 + 1.82x$

or $\hat{\pi}(x) = \frac{\exp(-3.69+1.82x)}{1+\exp(-3.69+1.82x)}$

- $\hat{\beta} > 0$ so $\hat{\pi} \uparrow$ as $x \uparrow$

- At $x = \bar{x} = 2.44$,

$$\hat{\pi} = \frac{\exp(-3.69 + 1.82(2.44))}{1 + \exp(-3.69 + 1.82(2.44))} = \frac{e^{0.734}}{1 + e^{0.734}} = \frac{2.08}{3.08} = 0.676$$

• $\hat{\pi} = \frac{1}{2}$ when $x = -\frac{\hat{\alpha}}{\hat{\beta}} = \frac{3.69}{1.82} = 2.04$.

• At $x = 2.04$, when $x \uparrow 1, \hat{\pi} \uparrow$ approx. $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = \frac{\hat{\beta}}{4} = 0.45$.

However, $s = 0.58$ for weight, and $1-$unit change is too large for this approx. to be good.
(actual $\hat{\pi} = 0.86$ at 3.04)

As $x \uparrow 0.1$ kg, $\hat{\pi} \uparrow$ approx $0.1\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = 0.045$
(actual $\hat{\pi} = 0.547$).

• At $x = 5.20$ (max. value), $\hat{\pi} = 0.997$. As $x \uparrow 0.1$, $\hat{\pi} \uparrow \approx 0.1(1.82)(0.997)(0.003) = 0.0006$.

Rate of changes varies as x does.

Note

• If we assume $Y \sim$ Normal and fitted model $\mu = \alpha + \beta x$,

$$\hat{\mu} = -0.145 + 0.323x$$

At $x = 5.2, \hat{\mu} = 1.53!$ (for estimated prob. of satellite)

• Alternative way to describe effect (not dependent on units) is

$$\hat{\pi}(x_2) - \hat{\pi}(x_1)$$

such as $\hat{\pi}(UQ) - \hat{\pi}(LQ)$

Ex.

For $x =$ weight, $LQ = 2.00, UQ = 2.85$

At $x = 2.00, \hat{\pi} = 0.48$; at $x = 2.85, \hat{\pi} = 0.81$.

$\hat{\pi}$ increases by 0.33 over middle half of x values.

Odds ratio interpretation

Since $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$, the odds

$$\frac{\pi}{1-\pi} = e^{\alpha+\beta x}$$

When $x \uparrow 1$, $\frac{\pi}{1-\pi} = e^{\alpha+\beta(x+1)} = e^{\beta}e^{\alpha+\beta x}$

$\longrightarrow$ odds multiply by e^{β}, which is $\frac{\text{odds at x+1}}{\text{odds at x}}$

$\beta = 0 \Longleftrightarrow e^{\beta} = 1$, odds stay constant.

Ex. $\hat{\beta} = 1.82, e^{\hat{\beta}} = 6.1$

Estimated odds of satellite multiply by 6.1 for 1 kg increase in weight.

If $x \uparrow 0.1, e^{0.1\hat{\beta}} = e^{0.182} = 1.20$.

Estimated odds increase by 20%.

Inference

CI

95% CI for β is

$$\hat{\beta} \pm z_{0.025}(SE) \qquad \text{(Wald method)}$$

$$1.815 \pm 1.96(0.377), \text{ or } (1.08, 2.55)$$

95% CI for e^{β}, multiplicative effect on odds of 1-unit increase in x, is

$$(e^{1.08}, e^{2.55}) = (2.9, 12.8).$$

95% CI for $e^{0.1\beta}$ is

$$(e^{0.108}, e^{0.255}) = (1.11, 1.29).$$

(odds increases at least 11%, at most 29%).

Note:

• For small n, safer to use likelihood-ratio CI than Wald CI (can do with LRCI option in SAS GENMOD)

Ex. LR CI for e^{β} is

$$(e^{1.11}, e^{2.60}) = (3.0, 13.4)$$

• For binary observation ($y = 0$ or 1), SAS (PROC GENMOD) can use model statement

$$\text{model } y = \text{weight/dist} = \text{bin} \ldots$$

but SAS forms logit as $\log\left[\frac{P(Y=0)}{P(Y=1)}\right]$ instead of $\log\left[\frac{P(Y=1)}{P(Y=0)}\right]$ unless use "descending" option.

e.g., get $\text{logit}(\hat{\pi}) = 3.69 - 1.82x$ instead of $\text{logit}(\hat{\pi}) = -3.69 + 1.82x$.

• Software can also construct CI for $\pi(x)$ (in SAS, PROC GENMOD or PROC LOGISTIC)

Ex. At $x = 3.05$ (value for 1^{st} crab), $\hat{\pi} = 0.863$. 95% CI for π is

$$(0.766, 0.924)$$

Significance Test

$H_0 : \beta = 0$ states that Y indep. of X (i.e., $\pi(x)$ constant)

$H_0 : \beta \neq 0$

$$z = \frac{\hat{\beta}}{SE} = \frac{1.815}{0.377} = 4.8$$

or Wald stat. $z^2 = 23.2$, $df = 1$ (chi-squared)
P-value< 0.0001

Very strong evidence that weight has positive effect on π.

Likelihood-ratio test

When $\beta = 0$, $L_0 = -112.88$ (log-likelihood under null)

When $\beta = \hat{\beta}$, $L_1 = -97.87$

Test stat.

$$-2(L_0 - L_1) = 30.0$$

Under H_0, has approx. χ^2 dist. $df = 1$ ($P < 0.0001$)

(can get using TYPE3 option in GENMOD)

Note: Recall for a model M,

$$\text{deviance } = -2(L_M - L_s)$$

L_s means the log-likelihood under saturated (perfect fit) model.

To compare model M_0 with a more complex model M_1,

$$\begin{aligned} \text{LR stat .} &= -2(L_0 - L_1) \\ &= -2(L_0 - L_s) - [-2(L_1 - L_s)] \\ &= \text{ diff. of deviances} \end{aligned}$$

Ex. $H_0 : \beta = 0$ in $\text{logit}[\pi(x)] = \alpha + \beta x$ (This is M_1).

$M_0 : \text{logit}[\pi(x)] = \alpha$

Diff. of deviances $= 225.76 - 195.74 = 30.0 =$ LR statistic.

Multiple Logistic Regression

Y binary, $\pi = P(Y = 1)$

$x_1, x_2, \ldots, x_k$ can be quantitative, qualitative (using dummy var's), or both.

Model form is

$$\text{logit}\,[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_k x_k$$

or, equivalently

$$\pi = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2 \cdots+\beta_k x_k}}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2 \cdots+\beta_k x_k}}$$

β_i = partial effect of x_i, controlling for other var's in model.

e^{β_i} = conditional odds ratio between Y and x_i (1-unit change) keeping other predictors fixed.

Ex. Horseshoe crab data

Sampled female has: $Y = 1$, at least 1 "satellite", $Y = 0$, no satellites.

Let $x =$ weight, $c =$ color (qualitative 4 cat's).

$c_1 = 1$	medium light
$c_1 = 0$	otherwise
$c_2 = 1$	medium
$c_2 = 0$	otherwise
$c_3 = 1$	medium dark
$c_3 = 0$	otherwise

For dark crabs, $c_1 = c_2 = c_3 = 0$.

CLASS COLOR statement in SAS asks SAS to set up dummy variables (indicator) for COLOR (need 3 dummies for 4 categories).

Model:

$$\text{logit}\,[P(Y=1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x_4$$

has ML fit

$$\text{logit}\,(\hat{\pi}) = -4.53 + 1.27c_1 + 1.41c_2 + 1.08c_3 + 1.69x$$

e.g., for dark crabs, $c_1 = c_2 = c_3 = 0$,

$$\text{logit}\,(\hat{\pi}) = -4.53 + 1.69x$$

At $x = \bar{x} = 2.44$,

$$\hat{\pi} = \frac{e^{-4.53+1.69(2.44)}}{1 + e^{-4.53+1.69(2.44)}} = 0.40$$

For medium light crabs ($c_1 = 1, c_2 = c_3 = 0$),

$$\text{logit}\,(\hat{\pi}) = -4.53 + 1.27(1) + 1.69x = -3.26 + 1.69x$$

At $x = \bar{x} = 2.44$, $\hat{\pi} = 0.70$

• At each weight, medium-light crabs are more likely than dark crabs to have satellites.

$$\hat{\beta} = 1.27, e^{1.27} = 3.6$$

At a given weight, estimated odds a med-light crab has satellite are 3.6 times estimated odds for dark crab.

e.g., at $x = 2.44$,

$$\frac{\text{odds for med-light}}{\text{odds for dark}} = \frac{0.70/0.30}{0.40/0.60} = 3.6$$

How could you get an estimated odds radio comparing ML to M or MD?

Compare ML ($c_1 = 1$) to M ($c_2 = 1$)

$$1.27 - 1.41 = -0.14, \quad e^{-0.14} = 0.9$$

At any given weight, estimated odds a ML crab has satellite are 0.9 times estimated odds a M crab has satellite.

Note

• Model assumes lack of interaction between color and weight in effects on π. This implies coefficient of x = weight is same for each color ($\hat{\beta}_4 = 1.69$).

i.e., shape of curve for effect of x on π is same for each color.

Inference: Do we need color in model?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

Given weight, Y is indep. of color.

Likelihood-ratio test statistic

$$\begin{aligned} -2(L_0 - L_1) &= -2[(-97.9) - (-94.3)] = \text{ diff. of deviances} \\ &= 195.7 - 188.5 = 7.2 \\ \text{df} &= 171 - 168 = 3, \quad P = 0.07 \end{aligned}$$

Some evidence (but not strong) of a color effect, given weight (only 22 "dark" crabs).

Is strong evidence of weight effect ($\hat{\beta} = 1.69$ has SE= 0.39).

Given color, estimated odds of satellite at weight $x + 1$ equal $e^{1.69} = 5.4$ times estimated odds at weight x.

Note Other simple models also adequate.

Ex. for nominal model, color estimates

$$\begin{array}{cccc} (1.27, & 1.41, & 1.08, & 0) \\ \uparrow & \uparrow & \uparrow & \uparrow \\ \text{ML} & \text{M} & \text{MD} & \text{D} \end{array}$$

suggest

$$\text{logit}\,[P(Y=1)] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

where $x_2 = 0,$ dark , $x_2 = 1,$ other color.

For it, $\hat{\beta}_2 = 1.295$ (SE= 0.522)

Given weight, estimated odds of satellite for nondark crabs $= e^{1.295} = 3.65$ times estimated odds for dark crabs.

Does model with 4 separate colors estimates fit better?

$$H_0: \text{ simple model (1 dummy)}$$

$$H_a: \text{ more complex model (3 dummies)}$$

Note; H_0 is $\beta_1 = \beta_2 = \beta_3 = 0$ in more complex model,

$$\text{logit}[P(Y=1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x$$

$$\begin{aligned} \text{LR stat.} &= \text{diff. in deviance} \\ &= 189.17 - 188.54 = 0.6 \ (\text{df} = 2) \end{aligned}$$

Simple model is adequate.

How about model allowing interaction?

$$\begin{aligned}\text{logit}\,[P(Y=1)] &= \alpha+\beta_1 c_1+\beta_2 c_2+\beta_3 c_3+\beta_4 x \\ &+ \beta_5 c_1 x+\beta_6 c_2 x+\beta_7 c_3 x\end{aligned}$$

Color	Weight effect	(coeff. of x)
dark	β_4	$(c_1=c_2=c_3=0)$
med-light	$\beta_4+\beta_5$	$(c_1=1)$
medium	$\beta_4+\beta_6$	$(c_2=1)$
med-dark	$\beta_4+\beta_7$	$(c_3=1)$

For H_0: no interaction ($\beta_5=\beta_6=\beta_7=0$)

LR stat.$=-2(L_0-L_1)=6.88$, df$=3$, P-value$=0.08$.

Weak evidence of interaction.
For easier interpretation, use simpler model.

Ordinal factors

Models with dummy var's treat color as qualitative (nominal).

To treat as quantitative, assign scores such as $(1, 2, 3, 4)$ and model trend.

$$\text{logit } = \alpha + \beta_1 x_1 + \beta_2 x_2$$

x_1 : weight, x_2 : color.

ML fit:

$$\text{logit } (\hat{\pi}) = -2.03 + 1.65 x_1 - 0.51 x_2$$

SE for $\hat{\beta}_1$ is 0.38, SE for $\hat{\beta}_2$ is 0.22.

$\hat{\pi} \downarrow$ as color $\uparrow$(more dark), controlling for weight.

$$e^{-0.51} = 0.60$$

which is estimated odds ratio for 1-category increase in darkness.

Does model treating color as nominal fit better?

$$H_0: \text{ simpler (ordinal) model holds}$$
$$H_a: \text{ more complex (nominal) model holds}$$

$$\begin{aligned} \text{LR stat.} &= -2(L_0 - L_1) \\ &= \text{ diff. in deviances} \\ &= 190.27 - 188.54 = 1.7, \quad \text{df } = 2 \end{aligned}$$

Do not reject H_0.

Simpler model is adequate.

Qualitative predictors

Ex. FL death penalty revisited

Victims' race	Defendant's race	Death Penalty Yes	No	n
B	B	4	139	143
	W	0	16	16
W	B	11	37	48
	W	53	414	467

$$\pi = P(Y = \text{ yes })$$

v = 1 victims black
0 victims white
d= 1 defendant black
0 defendant white

Model

$$\text{logit } (\pi) = \alpha + \beta_1 d + \beta_2 v$$

has ML fit

$$\text{logit } (\hat{\pi}) = -2.06 + 0.87d - 2.40v$$

e.g., controlling for victim's race, estimated odds of death penalty for black def's equal $e^{0.87} = 2.38$ times estimated odds for white def's

95% CI is:

$$e^{0.87 \pm 1.96(0.367)} = (1.16, 4.89)$$

Note

• Lack of interaction term means estimated odds ratio between Y and

$$d \text{ same at each level of } v \quad (e^{0.87} = 2.38)$$

$$v \text{ same at each level of } d \quad (e^{-2.40} = 0.09)$$

$$(e^{2.40} = \frac{1}{0.09}) = 11.1$$

i.e., cont. for d, est. odds of death pen. when $v =$ white were 11.1 times est. odds when $v =$ black.

(homogeneous association) means same odds ratio at each level of other var.

• $H_0 : \beta_1 = 0$ (Y conditional indep. of d given v)

$H_a : \beta_1 \neq 0$

$$z = \frac{\hat{\beta}_1}{\text{SE}} = \frac{0.868}{0.367} = 2.36$$

or Wald stat. $z^2 = 5.59,\ df = 1,\ P = 0.018.$

Evidence that death penalty more likely for black def's, controlling for v.

Likelihood-ratio test

Test of $H_0 : \beta_1 = 0$. Compares models

$$H_0 : \text{ logit } (\pi) = \alpha + \beta_2 v$$

$$H_a : \text{ logit } (\pi) = \alpha + \beta_1 d + \beta_2 v$$

$$\begin{aligned} \text{LR stat.} &= -2(L_0 - L_1) \\ &= 2(211.99 - 209.48) = 5.0 \\ &= \text{ diff. of deviances} \\ &= 5.39 - 0.38 = 5.01, \quad \text{df } = 1 \ \ (P = 0.025) \end{aligned}$$

Exercise

- Conduct Wald, LR, test of $H_0 : \beta_2 = 0$

- Get point and interval estimate of odds ratio for effect of victim's race, controlling for d.

what if $v = 1$ is white, $v = 0$ is black?

Note

• A common application for logistic regression having multiple 2×2 tables is multi-center clinical trials.

Center	Treatment	Response S	Response F
1	1		
	2		
2	1		
	2		
⋮	⋮	⋮	
K	1		
	2		

$$\text{logit}\,[P(Y=1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_{K-1} c_{K-1} + \beta x$$

Assumes odds ratio $= e^{\beta}$ in each table.

A model like this with several dummy var's for a factor is often expressed as

$$\text{logit}\,[P(Y=1)] = \alpha + \beta_i^c + \beta x$$

β_i^c is effect for center i (relative to last center).

To test $H_0 : \beta = 0$ about treatment effect for several 2×2 tables, could use

- likelihood-ratio test
- Wald test
- Cochran-Mantel-Haenszel test (p. 114)
- Small-sample generalization of Fisher's exact test (pp. 158–159)

Ex. Exercise 4.19

Y = support current abortion laws(1 = yes, 0 = no).

$$\text{logit } [P(Y = 1)] = \alpha + \beta_h^G + \beta_i^R + \beta_j^P + \beta x,$$

where β_h^G is for gender, β_i^R is for religion, and β_j^P is for party affil.

For religion (Protestant, Catholic, Jewish)

$$\hat{\beta}_1^R = -0.57,\ \hat{\beta}_2^R = -0.66,\ \hat{\beta}_3^R = 0.0$$

β_i^R represents terms

$$\hat{\beta}_1^R r_1 + \hat{\beta}_2^R r_2 = -0.57 r_1 - 0.66 r_2,$$

where $r_1 = 1$, Prot.; $r_1 = 0$, other, and $r_2 = 1$, Cath.; $r_2 = 0$, other.

Chapter 5: Building Logistic Regression Models

- Model selection

- Model checking

- Be careful with "sparse" categorical data (infinite estimates possible).

Model Selection with Many Predictors

Ex. Horseshoe crab study

Y = whether female crab has satellites (1 = yes, 0 = no).

Explanatory variables

- Weight

- Width

- Color(ML, M, MD, D), dummy var's c_1, c_2, c_3.

- Spine condition (3 categories), dummy var's s_1, s_2.

Consider model for crabs:

$$\begin{aligned}\text{logit}[P(Y=1)] &= \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 \\ &+ \beta_4 s_1 + \beta_5 s_2 + \beta_6(weight) + \beta_7(width)\end{aligned}$$

LR test of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$ has test stat.

$$\begin{aligned}-2(L_0 - L_1) &= \text{ difference of deviances} \\ &= 225.8 - 185.2 = 40.6 \\ &\text{df } = 7 \ (P < 0.0001)\end{aligned}$$

Strong evidence at least one predictor has an effect.

But,. . . , look at Wald tests of individual effects! (e.g., weight)

Multicollinearity (strong correlations among predictors) also plays havoc with GLMs.

e.g. corr(weight, width)=0.89

Partial effect of either relevant? Sufficient to pick one of these for a model.

Ex. Using backward elimination

• Use W = width, C = color, S = spline as predictors.

• Start with complex model, including all interactions.

• Drop "least significant" (e.g., largest P-value) variable among highest-order terms.

• Refit model

• Continue until all variables left are significant.

Note: If testing many interactions, simpler and perhaps better to test at one time as a group of terms.

Ex. H_0 : Model $C + S + W$ has 3 parameters for C, 2 parameters for S, 1 parameter for W.

H_a : Model

$$\begin{aligned} C * S * W &= C + S + W + C \times S \\ &+ C \times W + S \times W + C \times S \times W \end{aligned}$$

$$\begin{aligned} \text{LR stat.} &= \text{diff. in deviances} \\ &= 186.6 - 170.4 = 16.2 \\ \text{df} &= 166 - 152 = 14 \ \ (P = 0.30) \end{aligned}$$

Simpler model $C + S + W$ is adequate.

At next stage, S can be dropped from model $C+S+W$.

$$\text{diff. in deviances} = 187.5 - 186.6 = 0.9, \ \text{df} = 2.$$

Results in model fit (see text for details)

$$\text{logit}(\hat{\pi}) = -12.7 + 1.3c_1 + 1.4c_2 + 1.1c_3 + 0.47(\text{width})$$

Setting $\beta_1 = \beta_2 = \beta_3$ gives

$$\text{logit}(\hat{\pi}) = -13.0 + 1.3c + 0.48(\text{width})$$

where $c = 1$, ML, M, MD; $c = 0$, D.

Conclude

• Given width, estimated odds of satellite for nondark crabs equal $e^{1.3} = 3.7$ times est. odds for dark crabs.

95% CI: $e^{1.3 \pm 1.96(0.525)} = (1.3, 10.3)$
(wide CI reflects small number of dark crabs in sample)

• Given color, estimated odds of satellite multiplied by $e^{0.48 \pm 1.96(0.104)} = (1.3, 2.0)$ for each 1 cm increase in width.

Criteria for selecting a model

• Use theory, other research as guide.

• Parsimony (simplicity) is good.

• Can use some criterion to choose among set of models. Most popular criterion is *Akaike information criterion (AIC)* :
Chooses model with minimum

$$\text{AIC } = -2(L - \text{ no. model parameters})$$

where $L = \log$ likelihood.

• For exploratory purpose, can use automated procedure such as backward elimination.

• Ideally should have ≥ 10 outcomes of each type per predictor.

Ex. $n = 1000$, $(Y = 1)$ 30 times, $(Y = 0)$ 970 times. Model should contain ≤ 3 predictors.

Ex. $n = 173$ horseshoe crabs. $(Y = 1)$: 111 crabs; $(Y = 0)$: 62 crabs. Use ≤ 6 predictors.

Note

• Some software (e.g. PROC LOGISTIC in SAS) has options for stepwise selection procedures.

• Can further check fit with residuals for grouped data, influence measures, cross validation.

• To summarize predictive power, can use correlation$(Y, \hat{\pi})$.

Predictors	Correlation
color	0.28
width	0.40
color+width	0.452
color=dark+width	0.447

Another summary: Classification table

Predict $\hat{Y} = 1$ if $\hat{\pi} > 0.50$ and $\hat{Y} = 0$ if $\hat{\pi} < 0.50$

	Prediction		
Actual	$\hat{Y} = 1$	$\hat{Y} = 0$	
$Y = 1$	94	17	111
$Y = 0$	34	28	62

Sensitivity=P$(\hat{Y} = 1 | Y = 1) = \frac{94}{94+17} = 0.85$

Specificity=P$(\hat{Y} = 0 | Y = 0) = \frac{28}{28+34} = 0.45$

SAS: Get with CTABLE option in PROC LOGISTIC, for various "cutpoints".

SPSS: Get with BINARY LOGISTIC choice on REGRESSION menu.

Model checking

Is the chosen model adequate?

• Goodness of fit test
But, tests using deviance G^2, X^2 limited to "nonsparse" contingency tables.

• Check whether fit improves by adding other predictors, interactions between predictors.

LR stat. = change in deviance is useful for comparing models even when G^2 not valid as overall test of fit.

EX. Florida death penalty data

Victim Race	Defendant Race	Death penalty (Y) Yes	No	n
B	B	4	139	143
	W	0	16	16
W	B	11	37	48
	W	53	414	467

$$\pi = P(\text{Y} = \text{yes})$$

Goodness of fit

For model fit with $d = 1$ (black def.) or 1 (white def.) and $v = 1$ (black vic.) and $v = 0$ (white vic.),

$$\text{logit}\,(\hat{\pi}) = -2.06 + 0.87d - 2.40v,$$

$$\hat{\pi} = \frac{e^{-2.06+0.87d-2.40v}}{1 + e^{-2.06+0.87d-2.40v}}$$

e.g., for 467 cases with white def., victim, $d = v = 0$, $\hat{\pi} = \frac{e^{-2.06}}{1+e^{-2.06}} = 0.113$.

Fitted count "Yes"$= 467(0.113) = 52.8$
Fitted count "No"$= 467(0.887) = 414.2$

Observed counts = 53 and 414

Summarizing fit over 8 cells,

$$X^2 = \sum \frac{(\text{ obs } - \text{ fit })^2}{\text{fit}} = 0.20$$

$$G^2 = 2\sum \text{ obs } \log \frac{\text{obs}}{\text{fit}} = 0.38$$

$$= \text{deviance for model}$$

$$\text{df} = 4 - 3 = 1$$

4 = no. binomial observ's, 3 = no. model parameters.

For $G^2 = 0.38$, P= 0.54 for H_0 : model holds
(no evidence of lack of fit).

Note

• Model assumes lack of interaction between d and v in effects on Y, so goodness of fit test in this example is a test of H_0: no interaction.

• For continuous predictors or many predictors with small $\hat{\mu}_i$, X^2 and G^2 are not approx. χ^2. For better approx., can group data before applying X^2, G^2.

Hosmer-Lemeshow test groups using ranges of $\hat{\pi}$ values

(available in PROC LOGISTIC).

Or, can try to group predictor values (if only 1 or 2 predictors).

Residuals for Logistic Regression

At setting i of explanatory var's, let

$$y_i = \text{no. successes}$$

$$n_i = \text{ no. trials (preferably "large")}$$

$$\hat{\pi} = \text{estimated prob. of success,}$$

based on ML model fit

For a binomial GLM, Pearson residuals are

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

$$(X^2 = \sum_i e_i^2)$$

e_i (called Reschi in SAS GENMOD) is approx. $N(0, v)$, when model holds, but $v < 1$.

Standardized Pearson residual (adjusted residual in some books, SPSS)

$$r_i = \frac{y_i - n_i\hat{\pi}_i}{\text{SE}} = \frac{e_i}{\sqrt{1-h_i}}$$

where h_i is called "leverage" (r_i labelled StReschi in SAS).

r_i is approx. $N(0,1)$ when model holds.

$|r_i| > 2$ or 3 (approx.) suggests lack of fit.

EX. Y = admitted into graduate school at Berkeley (1=yes, 0=no). Data on p. 237 of text.

G =gender ($g = 0$ female, $g = 1$ male).

D = department (A, B, C, D, E, F).

$d_1 = 1,$ dept. A; $d_1 = 0,$ otherwise

.....

$d_5 = 1,$ dept. E; $d_5 = 0,$ otherwise

For department F, $d_1 = d_2 = \cdots = d_5 = 0$.

• Model

$$\text{logit}\,[P(Y=1)] = \alpha + \beta_1 d_1 + \cdots + \beta_5 d_5 + \beta_6 g$$

seems to fit poorly (deviance $G^2 = 20.2$, df=5, P=0.01)

• Simpler models fit poorly. e.g., model with $\beta_6 = 0$ assumes Y indep. of G, controlling for D, has

$$G^2 = 21.7, \text{df} = 6, P = 0.001$$

Apparently, there is a gender $\times$ dept. interaction.

Residual analysis indicates lack of fit only for dept. A (Standardized Pearson residual= ± 4.15 in model 2).

In other depts., model with no gender effect is adequate.

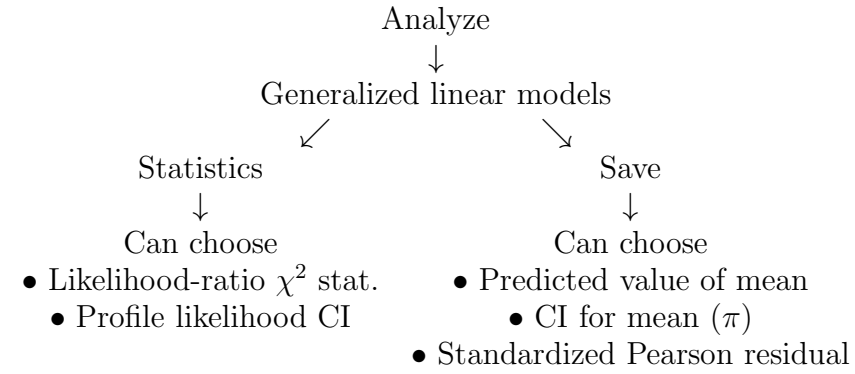
Note • In dept. A, $\hat{\theta} = 0.35$ (odds of admission lower for males)

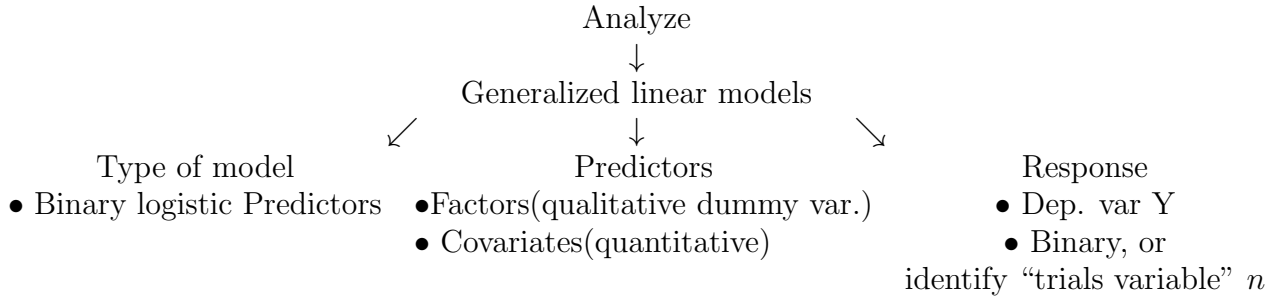
• Alternative way to express model with qualitative factor is

$$\text{logit } = [P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z,$$

where β_i^X is effect of classification in category i of X.

- In SPSS (version 16.0)





Sparse data

Caution: Parameter estimates in logistic regression can be infinite.

Ex.

	S	F
1	8	2
0	10	0

Model

$$\log\left[\frac{P(S)}{P(F)}\right] = \alpha + \beta x$$

$e^{\hat{\beta}} =$ odds ratio $= \frac{8\times 0}{2\times 10} = 0$

$\hat{\beta} =$ log odds ratio $= -\infty$.

Ex. Text p.155 for multi-center clinical trial (5 centers, each with 2×2 table)

Ex. $y = 0$ for $x < 50$ and $y = 1$ for $x > 50$.

$$\text{logit}[P(Y = 1)] = \alpha + \beta x$$

has $\hat{\beta} = \infty$. Software may not realize this!

PROC GENMOD
$\hat{\beta} = 3.84$, SE= 15601054.

PROC LOGISTIC gives warning

SPSS
$\hat{\beta} = 1.83$, SE= 674.8.

Infinite estimates exists when can separate x-values where $y = 1$ from x-values where $y = 0$ (perfect discrimination).

Ch 6: Multicategory Logit Models

Y has J categories, $J > 2$.

Extensions of logistic regression for nominal and ordinal Y assume a multinomial distribution for Y.

Model for Nominal Responses

Let $\pi_j = P(Y = j), \quad j = 1, 2, \ldots, J$

Baseline-category logits are

$$\log\left(\frac{\pi_j}{\pi_J}\right), \quad j = 1, 2, \ldots, J-1.$$

Baseline-category logit model has form

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \quad j = 1, 2, \ldots, J-1.$$

i.e., separate set of parameters (α_j, β_j) occurs for each logit (for each predictor).

Note:

- Which category we use as the baseline category (i.e., cat. J) is arbitrary (For nominal variables, the order of the categories is arbitrary).

- $\exp(\hat{\beta}_j)$ is the multiplicative impact of a 1-unit increase in x on the odds of making response j instead of response J.

- Can use model with ordinal response variables also, but then you ignore information about ordering.

Ex. Income and job satisfaction (1991 GSS data)

INCOME	JOB SATISFACTION			
($ 1000)	Very dissatisfied	Little dissatisfied	Moderate satisfied	Very satisfied
< 5	2	4	13	3
5-15	2	6	22	4
15-25	0	1	15	8
> 25	0	3	13	8

Using x =income scores $(3, 10, 20, 30)$, we use SAS (PROC LOGISTIC) to fit model

$$\log\left(\frac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x, \quad j = 1, 2, 3,$$

for $J = 4$ job satisfaction categories

SPSS: fit using MULTINOMIAL LOGISTIC suboption under REGRESSION option in ANALYZE menu

Prediction equations

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) = 0.56 - 0.20x$$

$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right) = 0.65 - 0.07x$$

$$\log\left(\frac{\hat{\pi}_3}{\hat{\pi}_4}\right) = 1.82 - 0.05x$$

Note

- For each logit, odds of being in less satisfied category (instead of very satisfied) decrease as x = income $\uparrow$.

- The estimated odds of being "very dissatisfied" instead of "very satisfied" multiplies by $e^{-0.20} = 0.82$ for each 1 thousand dollar increase in income.

For a 10 thousand dollar increase in income, (e.g., from row 2 to row 3 or from row 3 to row 4 of table), the estimated odds multiply by

$$e^{10(-0.20)} = e^{-2.0} = 0.14.$$

e.g, at at $x = 30$, the estimated odds of being "very dissatisfied" instead of "very satisfied" are just 0.14 times the corresponding odds at $x = 20$.

- Model treats income as quantitative, Y = job satisfaction as qualitative (nominal), but Y is ordinal. (We later consider a model that treats job satisfaction as ordinal.)

Estimating response probabilities

Equivalent form of model is

$$\pi_j = \frac{e^{\alpha_j+\beta_j x}}{1 + e^{\alpha_1+\beta_1 x} + \cdots + e^{\alpha_{J-1}+\beta_{J-1} x}},\ j = 1, 2, \ldots, J-1$$

$$\pi_J = \frac{1}{1 + e^{\alpha_1+\beta_1 x} + \cdots + e^{\alpha_{J-1}+\beta_{J-1} x}}$$

Then,

$$\frac{\pi_j}{\pi_J} = e^{\alpha_j+\beta_j x}$$

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x$$

Note $\sum \pi_j = 1$.

Ex. Job satisfaction data

$$\hat{\pi}_1 = \frac{e^{0.56-0.20x}}{1 + e^{0.56-0.20x} + e^{0.65-0.07x} + e^{1.82-0.05x}}$$

$$\hat{\pi}_2 = \frac{e^{0.65-0.07x}}{1 + e^{0.56-0.20x} + e^{0.65-0.07x} + e^{1.82-0.05x}}$$

$$\hat{\pi}_3 = \frac{e^{1.82-0.05x}}{1 + e^{0.56-0.20x} + e^{0.65-0.07x} + e^{1.82-0.05x}}$$

$$\hat{\pi}_4 = \frac{1}{1 + e^{0.56-0.2x} + e^{0.65-0.07x} + e^{1.82-0.05x}}$$

e.g. at $x = 30$, estimated prob. of "very satisfied" is

$$\hat{\pi}_4 = \frac{1}{1 + e^{0.56-0.20x} + e^{0.65-0.07x} + e^{1.82-0.05x}} = 0.365.$$

Likewise, $\hat{\pi}_1 = 0.002,\ \hat{\pi}_2 = 0.084,\ \hat{\pi}_3 = 0.550$

$\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 + \hat{\pi}_4 = 1.0$

• ML estimates determine effects for all pairs of categories, e.g.

$$\begin{aligned}\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) &= \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) - \log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right) \\ &= (0.564 - 0.199x) - (0.645 - 0.070x) \\ &= -0.081 - 0.129x\end{aligned}$$

• Contingency table data, so can test goodness of fit

The deviance is the LR test statistic for testing that all parameters not in model = 0.

Deviance = $G^2 = 4.18$, $df = 6$, P-value = 0.65 for H_0 : Model holds with linear trends for income

(Also, Pearson $X^2 = 3.6$, $df = 6, P = 0.73$ for same hypothesis)

Model has 12 logits (3 at each of 4 income levels), 6 parameters, so $df = 12 - 6 = 6$ for testing fit.

Note: Inference uses usual methods

- Wald CI for β_j is $\hat{\beta}_j \pm z(SE)$

- Wald test of $H_0 : \beta_j = 0$ uses $z = (\hat{\beta}_j/SE)$ or $z^2 \sim \chi^2_1$

- For small n, better to use LR test, LR CI

Ex. Overall "global" test of income effect

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

SAS reports Wald statistic $= 7.03$, $df = 3$, $P = 0.07$

Weak evidence, but ignores ordering of satisfaction categories.

(With many parameters, Wald stat. = quadratic form $\hat{\beta}'[Cov(\hat{\beta})]^{-1}\hat{\beta}$)

Can get LR statistic by comparing deviance with simpler "independence model"

LR stat. $= 9.29$, $df = 3$, $P = 0.03$.

Model for Ordinal Responses

The cumulative probabilities are

$$P(Y \leq j) = \pi_1 + \ldots + \pi_j, \quad j = 1, 2, \ldots, J.$$

Cumulative logits are

$$\text{logit}\,[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right]$$

$$= \log\left[\frac{P(Y \leq j)}{P(Y > j)}\right] = \log\left[\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right]$$

for

$$j = 1, 2, \ldots, J - 1$$

Cumulative logit model has form

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \beta x$$

- separate intercept α_j for each cumulative logit
- same slope β for each cumulative logit

Note

• e^{β} = multiplicative effect of 1-unit change in x on odds that $(Y \leq j)$ (instead of $(Y > j)$)

•

$$\frac{\text{odds of } (Y \leq j) \text{ at } x_2}{\text{odds of } (Y \leq j) \text{ at } x_1} = e^{\beta(x_2 - x_1)}$$

$$= e^{\beta} \quad \text{when } x_2 = x_1 + 1$$

Also called proportional odds model.

Software notes

• SAS: ML fit with PROC LOGISTIC,
PROC GENMOD (dist=mult, link=clogit)

PROC LOGISTIC default for dummy variable is 1 in category, -1 if in last category, 0 otherwise.

To use usual form of 1 in category, 0 otherwise, use param = ref option, e.g.,

CLASS race gender / param = ref ;

- SPSS:

ANALYZE$\longrightarrow$ REGRESSION $\longrightarrow$ ORDINAL to get cumulative logit model but estimates $\hat{\beta}$ have opposite sign as in SAS (as in modeling $\log[P(Y > j)/P(Y \leq j)]$)

Ex. Job satisfaction and income

$$\text{logit}\,[\hat{P}(Y \leq j)] = \hat{\alpha}_j + \hat{\beta}x = \hat{\alpha}_j - 0.056x, \quad j = 1, 2, 3$$

Odds of response at low end of job satisfaction scale $\downarrow$ as $x =$ income $\uparrow$

$$e^{\hat{\beta}} = e^{-0.056} = 0.95$$

Estimated odds of satisfaction below any given level (instead of above it) multiplies by 0.95 for 1-unit increase in x (but, $x = 3, 10, 20, 30$)

For $\$10,000$ increase in income, estimated odds multiply by

$$e^{10\hat{\beta}} = e^{10(-0.056)} = 0.57$$

e.g., estimated odds of satisfaction being below (instead of above) some level at $\$30,000$ income equal 0.57 times the odds at $\$20,000$.

Note

- If reverse order, $\hat{\beta}$ changes sign but has same SE.

Ex. Category 1 = Very satisfied, 2 = Moderately satisfied, 3 = little dissatisfied, 4 = Very dissatisfied

$$\hat{\beta} = 0.056, e^{\hat{\beta}} = 1.06 = 1/0.95$$

(Response more likely at "very satisfied" end of scale as $x \uparrow$)

- $H_0 : \beta = 0$ (job satisfaction indep. of income) has

$$\text{Wald stat.} = \left(\frac{\hat{\beta} - 0}{SE}\right)^2 = \left(\frac{-0.056}{0.021}\right)^2 = 7.17$$

($df = 1, P = 0.007$)

LR statistic = 7.51 ($df = 1, P = 0.006$)

These tests give stronger evidence of association than if treat:

- Y as nominal (BCL model),

$$\log\left(\frac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x$$

(Recall $P = 0.07$ for Wald test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$)

- X, Y as nominal

Pearson test of indep. has $X^2 = 11.5$, $df = 9, P = 0.24$ (analogous to testing all $\beta_j = 0$ in BCL model with dummy predictors).

With BCL or cumulative logit models, can have quantitative and qualitative predictors, interaction terms, etc.

Ex. Y = political ideology (GSS data)
(1= very liberal, . . . , 5 = very conservative)

x_1 = gender (1 = F, 0 = M)

x_2 = political party (1 = Democrat, 0 = Republican)

ML fit

$$\text{logit}\,[\hat{P}(Y \leq j)] = \hat{\alpha}_j + 0.117x_1 + 0.964x_2$$

For $\hat{\beta}_1 = 0.117$, $SE = 0.127$
For $\hat{\beta}_2 = 0.964$, $SE = 0.130$

For each gender, estimated odds a Democrat's response is in liberal rather than conservative direction (i.e., $Y \leq j$ rather than $Y > j$) are $e^{0.964} = 2.62$ times estimated odds for Republican's response.

- 95% CI for true odds ratio is

$$e^{0.964 \pm 1.96(0.130)} = (2.0, 3.4)$$

- LR test of $H_0 : \beta_2 = 0$ (no party effect, given gender) has test stat. $= 56.8$, $df = 1$ ($P < 0.0001$)

Very strong evidence that Democrats tend to be more liberal that Republicans (for each gender)

Not much evidence of gender effect (for each party)

But, is there interaction?

ML fit of model permitting interaction is

$$\text{logit}\,[\hat{P}(Y \leq j)] = \hat{\alpha}_j + 0.366x_1 + 1.265x_2 - 0.509x_1x_2$$

For $H_0 : \beta_3 = 0$, LR stat. $= 3.99$, $df = 1$ ($P = 0.046$)

Estimated odds ratio for party effect (x_2) is

$$e^{1.265} = 3.5 \text{ when } x_1 = 0 \text{ (M)}$$
$$e^{1.265-0.509} = 2.2 \text{ when } x_1 = 1 \text{ (F)}$$

Estimated odds ratio for gender effect (x_1) is

$$e^{0.366} = 1.4 \text{ when } x_2 = 0 \text{ (Republican)}$$
$$e^{0.366-0.509} = 0.9 \text{ when } x_2 = 1 \text{ (Democrat)}$$

i.e., for Republicans, females ($x_1 = 1$) tend to be more liberal that males.

Find $\hat{P}(Y = 1)$ (very liberal) for male Republicans, female Republicans.

$$\hat{P}(Y \leq j) = \frac{e^{\hat{\alpha}_j + 0.366x_1 + 1.265x_2 - 0.509x_1x_2}}{1 + e^{\hat{\alpha}_j + 0.366x_1 + 1.265x_2 - 0.509x_1x_2}}$$

For $j = 1$, $\hat{\alpha}_1 = -2.674$

Male Republicans $(x_1 = 0, x_2 = 0)$

$$\hat{P}(Y = 1) = \frac{e^{-2.674}}{1 + e^{-2.674}} = 0.064$$

Female Republicans $(x_1 = 1, x_2 = 0)$

$$\hat{P}(Y = 1) = \frac{e^{-2.674 + 0.366}}{1 + e^{-2.674 + 0.366}} = 0.090$$

(weak gender effect for Republicans, likewise for Democrats but in opposite direction)

Similarly, $\hat{P}(Y = 2) = \hat{P}(Y \leq 2) - \hat{P}(Y \leq 1)$, etc.

Note $P(Y = 5) = P(Y \leq 5) - P(Y \leq 4) =$
$1 - P(Y \leq 4)$ (use $\hat{\alpha}_4 = 0.879$)

Note

- If reverse order of response categories

(very lib., slight lib., moderate, slight cons., very cons.)$\longrightarrow$

(very cons., slight cons., moderate, slight lib., very liberal)

estimates change sign, odds ratio $\longrightarrow$ 1/(odds ratio)

- For ordinal response, other orders not sensible.

Ex. categories (liberal, moderate, conservative)

Enter into SAS as 1, 2, 3

or PROC GENMOD ORDER=DATA;

or else SAS will alphabetize as
(conservative, liberal, moderate)
and treat that as ordering for the cumulative logits

Ch 8: Models for Matched Pairs

Ex. Crossover study to compare drug with placebo.

86 subjects randomly assigned to receive drug then placebo or else placebo then drug.

Binary response (S, F) for each

Treatment	S	F	Total
Drug	61	25	86
Placebo	22	64	86

Methods so far (e.g., X^2, G^2 test of indep., CI for θ, logistic regression) assume independent samples, they are inappropriate for dependent samples (e.g., same subjects in each sample, which yield matched pairs).

To reflect dependence, display data as 86 observations rather than 2×86 observations.

		Placebo S	Placebo F	
Drug	S	12	49	61
	F	10	15	25
		22	64	86

Population probabilities

	S	F	
S	π_{11}	π_{12}	π_{1+}
F	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	1.0

Compare dependent samples by making inference about $\pi_{1+} - \pi_{+1}$.

There is *marginal homogeneity* if $\pi_{1+} = \pi_{+1}$.

Note:

$$\begin{aligned} \pi_{1+} - \pi_{+1} &= (\pi_{11} + \pi_{12}) - (\pi_{11} - \pi_{21}) \\ &= \pi_{12} - \pi_{21} \end{aligned}$$

So, $\pi_{1+} = \pi_{+1} \Longleftrightarrow \pi_{12} = \pi_{21}$ $(symmetry)$

Under H_0 : marginal homogeneity,

$$\frac{\pi_{12}}{\pi_{12} + \pi_{21}} = \frac{1}{2}$$

Each of $n^* = n_{12} + n_{21}$ observations has probability $\frac{1}{2}$ of contributing to n_{12}, $\frac{1}{2}$ of contributing to n_{21}.

$n_{12} \sim bin(n^*, \frac{1}{2})$, mean $= \frac{n^*}{2}$, std. dev. $= \sqrt{n^*(\frac{1}{2})(\frac{1}{2})}$.

By normal approximation to binomial, for large n^*

$$z = \frac{n_{12} - n^*/2}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} \sim N(0,1)$$

$$= \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

Or,

$$z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2_1$$

called McNemar's test

Ex.

		Placebo S	Placebo F	
Drug	S	12	49	61 (71%)
	F	10	15	
		22 (26%)		86

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{49 - 10}{\sqrt{49 + 10}} = 5.1$$

$P < 0.0001$ for $H_0 : \pi_{1+} = \pi_{+1}$ vs $H_a : \pi_{1+} \neq \pi_{+1}$
Extremely strong evidence that probability of success is higher for drug than placebo.

CI for $\pi_{1+} - \pi_{+1}$

Estimate $\pi_{1+} - \pi_{+1}$ by $p_{1+} - p_{+1}$, difference of sample proportions.

$$Var(p_{1+} - p_{+1}) = Var(p_{1+}) + Var(p_{+1}) - 2Cov(p_{1+}, p_{+1})$$

$$SE = \sqrt{\hat{V}ar(p_{1+} - p_{+1})}$$

n_{11}	n_{12}
n_{21}	n_{22}

n

12	49
10	15

86

$$\begin{aligned} p_{1+} - p_{+1} &= \frac{n_{11} + n_{12}}{n} - \frac{n_{11} + n_{21}}{n} \\ &= \frac{n_{12} - n_{21}}{n} = \frac{49 - 10}{86} = 0.453 \end{aligned}$$

The standard error of $p_{1+} - p_{+1}$ is

$$\frac{1}{n}\sqrt{(n_{12} + n_{21}) - \frac{(n_{12} - n_{21})^2}{n}}$$

For the example, this is

$$\frac{1}{86}\sqrt{(49 + 10) - \frac{(49 - 10)^2}{86}} = 0.075$$

95% CI is $\quad 0.453 \pm 1.96(0.075) = (0.31, 0.60)$.

Conclude we're 95% confident that probability of success is between 0.31 and 0.60 higher for drug than for placebo.

Measuring agreement (Section 8.5.5)

Ex. Movie reviews by Siskel and Ebert

		Ebert			
		Con	Mixed	Pro	
Siskel	Con	24	8	13	45
	Mixed	8	13	11	32
	Pro	10	9	64	83
		42	30	88	160

How strong is their agreement?

Let $\pi_{ij} = P(S = i,\ E = j)$

$$P(\text{agreement}) = \pi_{11} + \pi_{22} + \pi_{33} = \sum \pi_{ii}$$
$$= 1 \quad \text{if perfect agreement}$$

If ratings are independent, $\pi_{ii} = \pi_{i+}\pi_{+i}$

$$P(\text{ agreement }) = \sum \pi_{ii} = \sum \pi_{i+}\pi_{+i}$$

Kappa $\quad \kappa = \dfrac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}}$

$$= \frac{P(\text{agree}) - P(\text{agree}|\text{independent})}{1 - P(\text{agree}|\text{independent})}$$

Note

•$\kappa = 0$ if agreement only equals that expected under independence.

•$\kappa = 1$ if perfect agreement.

•Denominator = maximum difference for numerator, if perfect agreement.

Ex.

$$\sum \hat{\pi}_{ii} = \frac{24 + 13 + 64}{160} = 0.63$$

$$\sum \hat{\pi}_{i+}\hat{\pi}_{+i} = \left(\frac{45}{160}\right)\left(\frac{42}{160}\right) + \cdots + \left(\frac{83}{160}\right)\left(\frac{88}{160}\right) = 0.40$$

$$\hat{\kappa} = \frac{0.63 - 0.40}{1 - 0.40} = 0.389$$

The strength of agreement is only moderate.

•95% CI for κ: $0.389 \pm 1.96(0.06) = (0.27,\ 0.51)$.

•For $H_0 : \kappa = 0$,

$$z = \frac{\hat{\kappa}}{SE} = \frac{0.389}{0.06} = 6.5$$

There is extremely strong evidence that agreement is better than "chance".

•In SPSS,

Analyze → Descriptive statistics → Crosstabs

Click statistics, check Kappa
(McNemar also is an option).

If enter data as contingency table (e.g. one column called "count")
Data → Weight cases by count

Ch 9: Models for Correlated, Clustered Responses

Usual models apply (e.g., logistic regr. for binary var., cumulative logit for ordinal) but model fitting must account for dependence (e.g., from repeated measures on subjects).

Generalized Estimating Equation (GEE) approach to repeated m

•Specify model in usual way.

•Select a "working correlation" matrix for best guess about correlation pattern between pairs of observations.

Ex. For T repeated responses, exchangeable correlation is

$$\begin{array}{cc} & \text{Time} \\ & \begin{array}{cccc} 1 & 2 & \cdots & T \end{array} \\ \begin{array}{c} \\ 1 \\ \text{Time} \\ 2 \\ \\ T \end{array} & \begin{array}{cccc} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ & \vdots & & \\ \rho & \rho & \cdots & 1 \end{array} \end{array}$$

•Fitting method gives estimates that are good even if misspecify correlation structure.

•Fitting method uses empirical dependence to adjust standard errors to reflect actual observed dependence.

•Available in SAS (PROC GENMOD) using REPEATED statement, identifying by

$$\text{SUBJECT} \;=\; \text{var}$$

the variable name identifying sampling units on which repeated measurements occur.

•In SPSS, Analyze → generalized linear models → generalized estimating equations
Menu to identify subject variable, working correlation matrix

Ex. Crossover study

		Placebo: S	Placebo: F	
Drug	S	12	49	61
	F	10	15	25
		22	64	86

Model

$\text{logit}[P(Y_t = 1)] = \alpha + \beta t, \;\; t = 1, \text{ drug }, \;\; t = 0, \text{ placebo}$

GEE fit

$$\text{logit}[\hat{P}(Y_t = 1)] = -1.07 + 1.96t$$

Estimated odds of S with drug equal $e^{1.96} = 7.1$ times estimated odds with placebo. 95% CI for odds ratio (for marginal probabilities) is

$$e^{1.96 \pm 1.96(0.377)} = (3.4, 14.9)$$

Note

•Sample marginal odds ratio

$$= \frac{61 \times 64}{25 \times 22} = 7.1 \quad (\log \hat{\theta} = 1.96)$$

(model is saturated)

	S	F
D	61	25
P	22	64

•With GEE approach, can have also "between-subject" explanatory var's, such as gender, order of treatments, etc.

•With identity link,

$$\hat{P}(Y_t = 1) = 0.26 + 0.45t$$

i.e., $0.26 = \frac{22}{86}$ = estimated prob. of success for placebo.
$0.26 + 0.45 = 0.71 = \frac{61}{86}$ = est. prob. of success for drug.

$\hat{\beta} = 0.45 = 0.71 - 0.26$ = estimated difference of proportions.

95% CI: $0.45 \pm 1.96(0.075) = (0.307, 0.600)$ for true diff.

Note

GEE is a "quasi-likelihood" method

•Assumes dist. (e.g. binomial) for Y_1, for $Y_2, \cdots$, for Y_T, (marginal dist.'s)

•No dist. assumed for joint dist. of $(Y_1, Y_2, \cdots, Y_T)$.

•No likelihood function

No LR inference (LR test, LR CI)

•For responses $(Y_1, Y_2, \cdots, Y_T)$ at T times, we consider marginal model that describes each Y_t in terms of explanatory var's.

•Alternative conditional model puts terms in model for subjects, effects apply conditional on subject. e.g.

$$\text{logit}[P(Y_{it} = 1) = \alpha_i + \beta t]$$

$\{\alpha_i\}$ (effect for subject i) commonly treated as "random effects" having a normal dist. (Ch 10)

Ex y = response on mental depression (1= normal, 0= abnormal)
three times (1,2,4 weeks)
two drug treatments (standard, new)
two severity of initial diagnosis groups (mild, severe)

Is the rate of improvement better with the new drug?

The data are a $2 \times 2 \times 2 = 2^3$ table for profile of reponses on (Y_1, Y_2, Y_3) at each of 4 combinations of drug and diagnosis severity.

		Response at Three Times							
Diag	Drug	nnn	nna	nan	naa	ann	ana	aan	aaa
Mild	Stan	16	13	9	3	14	4	15	6
Mild	New	31	0	6	0	22	2	9	0
Sev	Stan	2	2	8	9	9	15	27	28
Sev	New	7	2	5	2	31	5	32	6

		Sample Proportion Normal		
Diagnosis	Drug	Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New	0.18	0.50	0.83

e.g., $0.51 = (16+13+9+3)/(16+13+9+3+14+4+15+6)$

Let Y_t = response of randomly selected subject at time t, (1 = normal, 0 = abnormal)

s = severity of initial diagnosis (1 = severe, 0 = mild)

d = drug treatment (1 = new, 0 = standard)

t = time (0, 1, 2), which is $\log_2$(week number).

Model

$$\log\left[\frac{P(Y_t = 1)}{P(Y_t = 0)}\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

assumes same rate of change β_3 over time for each (s, d) combination.

Unrealistic?

More realistic model

$$\log\left[\frac{P(Y_t=1)}{P(Y_t=0)}\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4(d \times t)$$

permits time effect to differ by drug

$d = 0$ (standard), time effect $= \beta_3$ for standard drug,

$d = 1$ (new) time effect $= \beta_3 + \beta_4$ for new drug.

GEE estimates			
$\hat{\beta}_1$	=	-1.31	s
$\hat{\beta}_2$	=	-0.06	d
$\hat{\beta}_3$	=	0.48	t
$\hat{\beta}_4$	=	1.02	$d \times t$

Test of H_0 : no interaction ($\beta_4 = 0$) has

$$z = \frac{\hat{\beta}_4}{SE} = \frac{1.02}{0.188} = 5.4$$

Wald stat. $z^2 = 29.0$ (P<0.0001)

Very strong evidence of faster improvement for new drug.

Could also add $s \times d$, $s \times t$ interactions, but they are not significant.

•When diagnosis= severe, estimated odds of normal response are $e^{-1.31} = 0.27$ times estimated odds when diagnosis = mild, at each $d \times t$ combination.

•$\hat{\beta}_2 = -0.06$ is drug effect only at $t = 0$, $e^{-0.06} = 0.94 \approx 1.0$, so essentially no drug effect at $t = 0$ (after 1 week). Drug effect at end of study ($t = 2$) estimated to be $e^{\hat{\beta}_2 + 2(\hat{\beta}_4)} = 7.2$.

•Estimated time effects are

$$\hat{\beta}_3 = 0.48, \quad \text{standard treatment } (d = 0)$$

$$\hat{\beta}_3 + \hat{\beta}_3 = 1.50, \quad \text{new treatment } (d = 1)$$

Cumulative Logit Modeling of Repeated Ordinal Responses

For multicategory responses, recall popular logit models use logits of cumulative probabilities (ordinal response)

$$\log[P(Y \leq j)/P(Y > j)] \quad \text{cumulative logits}$$

or logits comparing each probability to a baseline (nominal response)

$$\log[P(Y = j)/P(Y = I)] \quad \text{baseline-category logits}$$

GEE for cumulative logit models presented by Lipsitz et al. (1994)

SAS (PROC GENMOD) provides with independence working correlations

Ex. Data from randomized, double-blind clinical trial comparing hypnotic drug with placebo in patients with insomnia problems

	Time to Falling Asleep				
	Initial	Follow-up			
Treatment		<20	20–30	30–60	>60
Active	<20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	>60	9	17	13	8
Placebo	<20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	>60	4	11	14	22

Sample marginal distributions

		Response			
Treatment	occasion	<20	20–30	30–60	>60
Active	Initial	0.1	0.17	0.34	0.39
	Follow-up	0.34	0.41	0.16	0.09
Placebo	Initial	0.12	0.17	0.29	0.42
	Follow-up	0.26	0.24	0.29	0.21

Ex.

$$0.10 = \frac{7+4+1+0}{7+4+1+0+\cdots+9+17+13+8}$$

$$0.34 = \frac{7+11+13+9}{7+4+1+0+\cdots+9+17+13+8}$$

```
data francom;
 input case treat occasion outcome;
datalines;
   1       1     0        1
   1       1     1        1
   2       1     0        1
   2       1     1        1
   3       1     0        1
   3       1     1        1
   4       1     0        1
   4       1     1        1
   5       1     0        1
   5       1     1        1
   6       1     0        1
   6       1     1        1
   7       1     0        1
   7       1     1        1
   8       1     0        1
   8       1     1        2
   9       1     0        1
   9       1     1        2
  10       1     0        1
  10       1     1        2

 239       0     0        4
 239       0     1        4
 ;
proc genmod data=insomnia;
  class case;
  model outcome = treat occasion treat*occasion /
     dist=multinomial link=cumlogit ;
  repeated subject=case / type=indep corrw;
run;
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1	-2.2671	0.2188	-2.6959	-1.8383	-10.36	<.0001
Intercept2	-0.9515	0.1809	-1.3061	-0.5969	-5.26	<.0001
Intercept3	0.3517	0.1784	0.0020	0.7014	1.97	0.0487
treat	0.0336	0.2384	-0.4337	0.5009	0.14	0.8879
occasion	1.0381	0.1676	0.7096	1.3665	6.19	<.0001
treat*occasion	0.7078	0.2435	0.2305	1.1850	2.91	0.0037

Y_t = time to fall asleep
x = treatment (0 = placebo, 1 = active)
t = occasion (0 = initial, 1 = follow-up after 2 weeks)

Model

$$\text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \times x), \quad j = 1, 2, 3$$

GEE estimates:
$\hat{\beta}_1 = 1.04$ ($SE = 0.17$), placebo occasion effect
$\hat{\beta}_2 = 0.03$ ($SE = 0.24$), treatment effect initially
$\hat{\beta}_3 = 0.71$ ($SE = 0.24$), interaction

Considerable evidence that distribution of time to fall asleep decreased more for treatment than placebo group.

$H_0 : \beta_3 = 0$ has $z = \frac{\hat{\beta}_3}{SE} = \frac{0.71}{0.24} = 2.9, \quad P = 0.004$

The treatment effect is $\hat{\beta}_2 = 0.03$ at $t = 0$

$$\hat{\beta}_2 + \hat{\beta}_3 = 0.03 + 0.71 = 0.74 \text{ at } t = 1$$

For the active group, the odds of response$\leq j$ (e.g. falling asleep in ≤ 60 minutes) are estimated to be

- $e^{0.03} = 1.03$ times the odds for placebo, at initial time ($t = 0$)

- $e^{0.74} = 2.1$ times the odds for placebo, at follow-up time ($t = 1$)

Handling repeated measurement and other forms of clustered d

Observations $(Y_1, Y_2, \cdots, Y_T)$ (e.g., T times)

1. Marginal models (Ch. 9)

Simultaneously model each $E(Y_t)$ $t = 1, \cdots, T$
get standard errors that account for the actual dependence using method such as GEE (generalized estimating equations)
e.g. REPEATED statement in PROC GENMOD (SAS)

Ex. binary data $Y_t = 0$ or 1, $t = 1, 2$ (matched pair)

$$E(Y_t) = P(Y_t = 1)$$

Model $\text{logit}[P(Y_t = 1)] = \alpha + \beta x_t$, x_t is the value of explan. var. for observ. t

depression data (matched triplets) $\rightarrow$ (some explan. var's constant across t, others vary)

Note: In practice, missing data is a common problem in longitudinal studies. (no problem for software, but are observations "missing at random"?)

2. Random effects models (Ch. 10)

Account for having multiple responses per subject (or "cluster") by putting a subject term in model

Ex. binary data $Y_t = 0$ or 1

Now let Y_{it} = response by subject i at time t

Model

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_t$$

intercept α_i varies by subject

Large positive α_i

$$\text{large } P(Y_{it} = 1) \text{ each } t$$

Large negative α_i

$$\text{small } P(Y_{it} = 1) \text{ each } t$$

These will induce dependence, averaging over subjects.

Heterogeneous popul. $\Rightarrow$ highly variable $\{\alpha_i\}$

but number of parameters > number of subjects

Solution • Treat $\{\alpha_i\}$ as random rather than parameters (fixed)

• Assume dist. for $\{\alpha_i\}$. e.g. $\{\alpha_i\} \sim N(\alpha, \sigma)$ (2 para.) or $\alpha_i = \alpha + u_i$ where is α a parameter

random effects $\rightarrow \{u_i\} \sim N(0, \sigma)$

Model

$$\text{logit}[P(Y_{it} = 1)] = u_i + \alpha + \beta x_t$$

$\{u_i\}$ are random effects. Parameters such as β are fixed effects.

A generalized linear mixed model (GLMM) is a GLM with both fixed and random effects.

SAS: PROC NLMIXED (ML), PROC GLIMMIX (not ML)

Software must "integrate out" the random effects to get the likelihood fn., ML est. $\hat{\beta}$ and std. error.

Also estimate σ and can predict $\{u_i\}$.

Ex. depression study

		Response at Three Times							
Diag	Drug	nnn	nna	nan	naa	ann	ana	aan	aaa
Mild	Stan	16	13	9	3	14	4	15	6
Mild	New	31	0	6	0	22	2	9	0
Sev	Stan	2	2	8	9	9	15	27	28
Sev	New	7	2	5	2	31	5	32	6

		Sample Proportion Normal		
Diagnosis	Drug	Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New	0.18	0.50	0.83

We used GEE to fit "marginal model"

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t)$$

$Y_t = 1$ (normal), s : severity (=1 severe), d : drug (=1 new), t : time (= 0,1,2).

Now we use ML to fit "random effects" model

$$\text{logit}[P(Y_{it} = 1)] = u_i + \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t)$$

assume $\{u_i\}$ has $N(0, \sigma)$ (need to estimate σ).

$$\hat{\beta}_1 = -1.32 \quad \text{severity effect}$$

$$\hat{\beta}_2 = -0.06 \quad \text{drug effect at } t = 0$$

$\hat{\beta}_3 = 0.48$ time effect for standard drug ($d = 0$)

$\hat{\beta}_4 = 1.02$ add to $\hat{\beta}_3$ to get time effect for new drug ($d = 1$)

$\hat{\sigma} = 0.07$ est. std. dev. of random effects

Note

- Similar conclusions as with marginal model (e.g., significant interaction)

- When $\hat{\sigma} = 0$, estimates and std. errors same as treating repeated observ's as independent

- Details: Ch.10 of textbook

- When $\hat{\sigma}$ is large, estimates from random effects logit model tend to be larger than estimates from marginal logit model.

Graph here

Ch 7: Loglinear Models

- Logistic regression distinguishes between response variable Y and explanatory variables $x_1, x_2, \ldots$.

- Loglinear models treat all variables as response variables (like correlation analysis)

Ex. (text) Survey of high school students

Y_1: used marijuana? (yes, no)

Y_2: alcohol? (yes, no)

Y_3: cigarettes? (yes, no)

Any variables independent?
Strength of association?
Interaction?

Loglinear models treat cell counts as Poisson and use the log link function

Motivation: In $I \times J$ table, X and Y are independent if

$$P(X = i, Y = j) = P(X = i)P(Y = j) \text{ for all } i, j.$$

$$\pi_{ij} = \pi_{i+}\,\pi_{+j}$$

For expected frequencies

$$\mu_{ij} = n\,\pi_{ij}$$

$$\mu_{ij} = n\,\pi_{i+}\,\pi_{+j}$$

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

λ_i^X: effect of X falling in row i
λ_j^Y: effect of Y falling in column j

This is loglinear model of independence

Treats X and Y symmetrically (differs from logistic regression, which distinguishes between Y =response and X =explanatory)

Ex. Income and job satisfaction

Income ($1000)	Job Satisfaction (Y)			
	Very dissat.	Little dissat.	Moderately satis.	Very satisfied
< 5	2	4	13	3
$5 - 15$	2	6	22	4
$15 - 25$	0	1	15	8
> 25	0	3	13	8

Using x =income scores $(3, 10, 20, 30)$,
we used SAS (PROC LOGISTIC) to fit model

$$\log\left(\frac{\pi_i}{\pi_4}\right) = \alpha_j + \beta_j\, x, \quad j = 1, 2, 3$$

EX. Income (I) and job satisfaction (S)

(We analyzed this using multinomial logit models in Ch. 6)

Model

$$\log(\mu_{ij}) = \lambda + \lambda_i^I + \lambda_j^S$$

can be expressed as

$$\log(\mu_{ij}) = \lambda + \lambda_1^I z_1 + \lambda_2^I z_2 + \lambda_3^I z_3 + \lambda_1^S w_1 + \lambda_2^S w_2 + \lambda_3^S w_3$$

where

z_1 =1, income cat. 1 $\quad$ w_1 =1, sat. cat. 1

0, otherwise $\quad$ 0, otherwise

. . . . . $\quad$ . . . . .

z_3 =1, income cat. 3 $\quad$ w_3 =1, sat. cat. 3

0, otherwise $\quad$ 0, otherwise

Parameter	No. nonredundant	
λ	1	
λ_i^X	$I-1$	(can set $\lambda_I^X = 0$)
λ_j^Y	$J-1$	(can set $\lambda_J^Y = 0$)
λ_{ij}^{XY}	$(I-1)(J-1)$	(no. of products of dummy var's)

Note.

For a Poisson loglinear model

$$\text{df} = \text{no. Poisson counts} - \text{ no. parameters}$$

(no. Poisson counts = no. cells)

Ex. Independence model, $I \times J$ table

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

$$\text{df} = IJ - [1 + (I-1) + (J-1)] = (I-1)(J-1)$$

Test of indep. using Pearson X^2 or like-ratio G^2 is a goodness-of-fit test of the indep. loglinear model.

The model allowing association

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

has df= 0 (saturated), giving a perfect fit.

Ex. Recall 4×4 table

Independence model

$$\log(\mu_{ij}) = \lambda + \lambda_i^I + \lambda_j^S$$

has $X^2 = 11.5$, $G^2 = 13.5$, df$= 9$.

Saturated model: $X^2 = G^2 = 0$, df$= 0$. (All $\hat{\mu}_{ij} = n_{ij}$)

Estimated odds ratio using highest and lowest categories is

$$\begin{aligned}\frac{\hat{\mu_{11}}\hat{\mu_{44}}}{\hat{\mu_{14}}\hat{\mu_{41}}} &= \exp\left[\hat{\lambda_{11}^{IS}} + \hat{\lambda_{44}^{IS}} - \hat{\lambda_{14}^{IS}} - \hat{\lambda_{41}^{IS}}\right] \\ &= \exp(24.288) = 35,294,747,720 \quad \text{(GENMOD)} \\ &= \frac{n_{11}n_{44}}{n_{14}n_{41}} = \frac{2\times 8}{3\times 0} = \infty\end{aligned}$$

since model is saturated

(software doesn't quite get right answer when ML est.=∞)

Loglinear Models for Three-way Tables

Two-factor terms represent conditional log odds ratios, at fixed level of third var.

Ex. $2 \times 2 \times 2$ table

Let μ_{ijk} denotes expected freq.; λ_{ik}^{XZ} and λ_{jk}^{YZ} denote assoc. para.'s.

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

satisfies

- $\log \theta_{XY(Z)} = 0$ (X and Y cond. indep., given Z)
- $$\begin{aligned} \log \theta_{X(j)Z} &= \lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ} \\ &= 0 \quad \text{if } \{\lambda_{ij}^{XZ} = 0\} \end{aligned}$$

  i.e. the XZ odds ratio is same at all levels of Y

  Denote by (XZ, YZ),

called model of XY conditional independence.

Ex.

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

called model of homogeneous association

Each pair of var's has association that is identical at all levels of third var.

Denote by (XY, XZ, YZ).

Ex. Berkeley admissions data ($2 \times 2 \times 6$)
Gender(M,F) $\times$ Admitted(Y, N) $\times$ Department(1,2,3,4,5,6)

Recall marginal 2×2 AG table has $\hat{\theta} = 1.84$

- Model (AD, DG)

A and G cond. indep., given D.
e.g. for Dept. 1,

$$\hat{\theta}_{AG(1)} = \frac{531.4 \times 38.4}{293.6 \times 69.6} = 1.0$$
$$= \hat{\theta}_{AG(2)} = \ldots = \hat{\theta}_{AG(6)}$$

But model fits poorly: $G^2 = 21.7$, $X^2 = 19.9$, $df = 6$ ($P < .0001$) for H_0 : model (AD, DG) holds.

Conclude A and G not cond. indep given D.

• Model (AG, AD, DG)

Also permits AG assoc., with same odds ratio for each dept.

e.g. for Dept. 1,

$$\begin{aligned}\hat{\theta}_{AG(1)} &= \frac{529.3 \times 36.3}{295.7 \times 71.7} = 0.90 \\ &= \hat{\theta}_{AG(2)} = \ldots = \hat{\theta}_{AG(6)} \\ &= \exp(\hat{\lambda}^{AG}_{11} + \hat{\lambda}^{AG}_{22} - \hat{\lambda}^{AG}_{12} - \hat{\lambda}^{AG}_{21}) \\ &= \exp(-.0999) = .90\end{aligned}$$

Control for dept., estimated odds of admission for males equal .90 times est. odds for females.

$\hat{\theta} = 1.84$ ignores dept. (Simpson's paradox)

But this model also fits poorly: $G^2 = 20.2$, $X^2 = 18.8$, $df = 5$ ($P < .0001$) for H_0 : model (AG, AD, DG) holds.

i.e. true AG odds ratio not identical for each dept.

• Adding 3-factor interaction term λ^{GAD}_{ijk} gives saturated model ($1 \times 1 \times 5$ cross products of dummies)

Residual analysis

For model (AD, DG) or (AD, AG, DG), only Dept.1 has large adjusted residuals. (≈ 4 in abs. value)

Dept. 1 has

- fewer males accepted than expected by model
- more females accepted than expected by model

If re-fit model (AD, DG) to $2 \times 2 \times 5$ table for Depts 2-6, $G^2 = 2.7$, $df = 5$, good fit.

Inference about Conditional Associations

EX. Model (AD, AG, DG)

$$\log \mu_{ijk} = \lambda + \lambda_i^G + \lambda_j^A + \lambda_k^D + \lambda_{ij}^{GA} + \lambda_{ik}^{GD} + \lambda_{jk}^{AD}$$

$H_0 : \lambda_{ij}^{GA} = 0$ (A cond. indep of G, given D)

Likelihood-ratio stat. $-2(L_0 - L_1)$
= deviance for (AD, DG) -deviance for (AD, AG, DG)
=21.7-20.3=1.5, with df=6-5=1 (P =.21)

H_0 plausible, but test "shaky" because model (AD, AG, DG) fits poorly.

Recall $\hat{\theta}_{AG(D)} = \exp\left(\hat{\lambda}_{11}^{AG}\right) = \exp(-.0999) = .90$

95%CI for $\theta_{AG(D)}$ is

$$\exp[-.0999 \pm 1.96(.0808)] = (.77, 1.06)$$

Plausible that $\theta_{AG(D)} = 1$.

There are equivalences between loglinear models and corresponding logit models that treat one of the variables as a response var., others as explanatory. (Sec. 6.5)

Note.

- Loglinear models extend to any no. of dimensions
- Loglinear models treat all variables symmetrically; Logistic regr. models treat Y as response and other var's as explanatory.
  Logistic regr. is the more natural approach when one has a single response var. (e.g. grad admissions) See output for logit analysis of data

Ex Text Sec. 6.4, 6.5

Auto accidents

$G =$ gender (F, M)

$L =$ location (urban, rural)

$S =$ seat belt use (no, yes)

$I =$ injury (no, yes)

I is natural response var.

Loglinear model (GLS, IG, IL, IS) fits quite well ($G^2 = 7.5, df = 4$)

Simpler to consider logit model with I as response.

$$\text{logit}\,[\hat{P}(I = \text{ yes })] = -3.34 + 0.54G + 0.76L + 0.82S$$

Controlling for other variables, estimated odds of injury are:

$e^{0.54} = 1.72$ times higher for females than males (CI: $(1.63, 1.82)$)

$e^{0.76} = 2.13$ times higher in rural than urban locations (CI: $(2.02, 2.25)$)

$e^{0.82} = 2.26$ times higher when not wearing seat belt (CI: $(2.14, 2.39)$)

Why ever use loglinear model for contingency table?

Info. about all associations, not merely effects of explanatory var's on response.

Ex. Auto accident data

Loglinear model (GI, GL, GS, IL, IS, LS) ($G^2 = 23.4, df = 5$)fits almost as well as (GLS, GI, IL, IS) ($G^2 = 7.5, df = 4$) in practical terms but n is huge(68,694)

Variables	$\hat{\lambda}$	Odds ratio ($\hat{\theta}$)		$1/\hat{\theta}$	
GL	-0.21	0.81	(fem. rur.)	1.23	
GS	-0.46	0.63	(fem. no)	1.58	
GI	-0.54	0.58	(fem. no)	1.72	e.g., for those
LS	-0.08	0.92	(rur. no)	1.09	
LI	-0.75	0.47	(rur. no)	2.13	
SI	-0.81	0.44	(no no)	2.26	

not wearing seat belts, the estimated odds of being in-

jured are 2.26 times the estimated odds of injury for those wearing seat belts, cont. for gender and location. (or. interchanges S and I in interp.)

Dissimilarity index

$$D = \sum |p_i - \hat{\pi}_i|/2$$

- $0 \leq D \leq 1$, with smaller values for better fit.

- $D =$ proportion of sample cases that must move to different cells for model to fit perfectly.

Ex. Loglinear model (GLS, IG, IL, IS) has $D = 0.003$.

Simpler model (GL, GS, LS, IG, IL, IS) has $G^2 = 23.4$ (df=5) for testing fit (P<0.001), but $D = 0.008$. (Good fit for practical purposes, and simpler to interpret GS,LS associations.)

For large n, effects can be "statistically significant" without being "practically significant."

Model can fail goodness-of-fit test but still be adequate for practical purposes.

Can be useful to describe closeness of sample cell proportions $\{p_i\}$ in a contingency table to the model fitted proportions $\{\hat{\pi}_i\}$.