

# The Final Lecture

Nov. 27

Tuesday - Stratified 2x2 tables

- Univ. admission data

→ 2 cat -  $Y$  = admission Yes/No  
6 cat  $X_1$  = Program applied to  
2 cat  $X_2$  = Sex of Applicant

→ 2x2x6 table

↔ - 6 2x2 tables

Simpson's paradox

- collapsed 2x2 table

- misleading pattern of association

→ stratified analysis

- Mantel-Haenszel estimate

OR ≈ 9  
Yes No

Test. →  $P = .25$

CI. .77 - 1.06



Logistic Regression

Stratified are essentially equivalent alternatives

- M.H. estimate may be

# Higher-Way Contingency Tables (ctd.)

*admission*

## Logistic Regression Analysis of the Admission Data

glm(formula = cbind(yes, no) ~ program + sex, family = binomial)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.58205	0.06899	8.436	<2e-16 ***
programB	-0.04340	0.10984	-0.395	0.693
programC	-1.26260	0.10663	-11.841	<2e-16 ***
programD	-1.29461	0.10582	-12.234	<2e-16 ***
programE	-1.73931	0.12611	-13.792	<2e-16 ***
programF	-3.30648	0.16998	-19.452	<2e-16 ***
sexF	0.09987	0.08085	1.235	0.217

*factors*

*- estimates on logit. scale*

*log(p/(1-p))*

*exponentiation*

*estimate odds ratios*

Null deviance: 877.056 on 11 degrees of freedom  
Residual deviance: 20.204 on 5 degrees of freedom

### TRANSFORMED COEFFICIENTS AND 95% C.I.'s

	Estimate	2.5 %	97.5 %
(Intercept)	1.78970606	1.56334599	2.04884127
programB	0.95753028	0.77207142	1.18753812
programC	0.28291804	0.22955913	0.34867975
programD	0.27400567	0.22268066	0.33716044
programE	0.17564230	0.13717694	0.22489361
programF	0.03664494	0.02626184	0.05113317
sexF	1.10502735	0.94309706	1.29476117

*relative to prog A*

### TESTING FOR INTERACTION

Analysis of Deviance Table  
Model: binomial, link: logit  
Response: cbind(yes, no)  
Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			11	877.06	
program	5	855.32	6	21.74	1.242e-182
sex	1	1.53	5	20.20	0.22
program:sex	5	20.20	0	-1.581e-13	1.144e-03

*OR .1.1 est. 1/OR for M<sub>2</sub> = 1.77-1.0*

more better if many  
"sparse" tables

ie if stratifying variable  
(eg - Program)  
has many categories

Log-linear models in  
Higher-Way tables

→ specialized applications

eg.  $2 \times 2 \times 2$  table

- subjects are suspected cases  
of deep vein thrombosis  
- blood clot

- 3 diagnostic tests

gold

standard

- invasive test  
Venography

- ultrasound ↑

- d-dimer - simple  
- quick & dirty blood test

Model 1 - complete independence

→ A/S & d-dimer

not related to venogr.

→ tests are not valid.

Model 2 = "interaction" between  
venography &  
u/s.

→ If sign. → u/s  
has some validity

Model 3 → association between  
d.dimer & venogr

- but no d.dimer × u/s

model conditional on  
venography - d.dimer &  
u/s are independent

→ independent tests

→ Combining information  
in tests may  
yield better test than  
either alone.

---

Review? -

# Categorical Data

→ Categorical response models

-  $Y$  - dependent, outcome  
→ categorical + binary

↳ multi-category

+ ordinal - ordered categories

- nominal

eg. hair colour

- no order.

→ more general perspective  
analysis at frequency or  
counts

→ counts arising from  
individuals

eg. Brazil trip data

- # trips by

→ Vineyard data a person

- # of logs of grapes  
in a row.

→ counts aggregated

→ reflect:  $\geq 1$  included

→ counts of events

- shark attacks
- leukemia in boys

→ counts arising from  
cross-tabulations

- contingency tables

Basic analytic frameworks

- single sample
- two samples
- $k$ -samples ( $k \geq 2$ )
- simple regression
- multiple regression  
/ linear model.

with assumptions or specifications  
sampling framework (pairing  
independence)  
→ appropriate  
methods

→ independent data.

→ Binary outcomes -  $Y \in \{ \text{yes}, \text{no} \}$

one-sample -  $\hat{p}$  s.e. ( $\hat{p}$ ) =  $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

→ C.I.'s for  $p$

large & small sample  
(exact) methods

Exact binomial CI.

two-sample  $= \hat{p}_1, \hat{p}_2$   $2 \times 2$

k-sample  $\rightarrow 2 \times k$

regression  $\rightarrow 306$   $\rightarrow 2 \times 2 \times 6$

$y$  - counts of events  
- Poisson regression

one sample estimate  $\mu = E(y)$

two sample  $\rightarrow$  focus on  
comparisons of rates

$\rightarrow$  rate =  $\frac{\# \text{ events}}{\text{denominator}}$

$\rightarrow$  number of persons  
 $\times$  length of observation

- person-time  $\rightarrow$   
- vehicle-mile

measures of exposure

rates =  $\lambda_1, \lambda_2$   
 $= \mu_1 / T_1, \mu_2 / T_2$

Cont for  $\lambda_i / \lambda_2$  - rate ratios  
 $\lambda_2$  - relative rates

estimates  $\lambda_i = RR$   
 $\lambda_1 = Y_1 / T_1$   $Y_{1,2}$   
 $\lambda_2 = Y_2 / T_2$  observed  
 com

95%  $\frac{\lambda_1}{\lambda_2} \times \exp \left\{ \pm 2 \alpha_{1/2} \sqrt{\frac{1}{Y_1} + \frac{1}{Y_2}} \right\}$

→ Poisson regression

observed counts  $Y_i$   
 $\mu_i = E(Y_i)$

$\mu_i = T_i \times \lambda_i$   
 ↳ person-years

$\log \mu_i = \log T_i + \log \lambda_i$   
 ↳ offset

$\Rightarrow \log \mu_i = \alpha + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$

→ estimate  $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_p$   
 $\hat{\lambda} = e^{\hat{\alpha}} + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_p X_{ip}$

→ model provides  
risk estimates

- I also risk comparisons

$e^{\beta_i \Delta X_i}$  = relative risk  
associated with  
change  $\Delta X_i$  - a c<sup>th</sup>  $X_i$ .

-  $X_i = 0/1 \rightarrow \Delta X_i = 1$

$e^{\hat{\beta}_i}$  → estimate of RR.

→ raw outputs of log-linear  
model → transformed to  
exp.

---

also - looked individual level  
counts

- # drops # logs graphs

not looking at risk

- same mechanics of  
poisson regression

- relative rate interpretation

- Poisson assumption more tentative

- event data, Poisson assumption mathematically justified by as approx. to Binomial

$$Y \sim \text{Bi}(n, p)$$

↓  
large 'n' some

$$Y \sim \text{Poisson}$$

- more likely Poisson assumption fails
  - over-dispersion models

---

Modeling !!  
→ art !!

- ↳ model building / selection
  - choosing X-variables
  - assessing assumption
  - diagnostic plots
  - AIC - model selection criteria

→ Hypothesis testing & estimation / prediction based on chosen model

Regression -  $\beta =$   
SE.

$$\text{test } z = \frac{\hat{\beta} - \beta_0}{\text{SE}}$$

$$\left( \hat{\beta} \pm z_{\alpha/2} \text{SE} \right)$$

↳ transform for interpretation

↳ multi-parameter inference

→ full & reduced model

- L.R.T.

---

## Contingency tables

- don't necessarily sum up out a response

→ Marriage - H → clean  
W → clean