Chapter 10: Conditional Logistic Regression

10.1 Introduction

- Usual maximum likelihood approach to estimation in logistic regression not appropriate if there is insufficient sample size, particularly if data highly stratified and small number of subjects in each stratum

- Highly stratified data often come from design with cluster sampling, e.g., fraternal twins, litter mates, right and left sides of body, two occasions for expression of opinion

Types of Stratified Data

- 1:1 the matched set consists of one case and one control from each stratum. Most common situation. (Section 10.2-10.6)

- 1:m the matched set consists of 1 case and m controls (usually m ranges between 2 and 5). (Section 10.7)

- n:m the matched set consists of n cases with m controls (usually both m and n are between 1 and 5)

Appropriate form of logistic regression for these types of data is called *conditional logistic regression*.

10.2 Paired Observations from a Highly Stratified Cohort Study

• Consider randomized clinical trial where $h = 1,2,\ldots, q$ centers randomly selected and, at each center, one randomly selected patient is placed on treatment and another on placebo. Interested in whether the patients improve.

•Since there are only 2 patients per center it is not possible to estimate a center effect without bias for all parameters. (Need at least 5 observations per category of each variable in model).

•Suppose $y_{hi} = 1$ if improvement occurs and $y_{hi} = 0$ otherwise ($i = 1,2$ for trt, placebo) and $x_{hi} = 1$ for treatment, $x_{hi} = 0$ for placebo, and $\boldsymbol{z}_{hi} = (z_{hi1}, z_{hi2}, \ldots, z_{hit})'$ represents the t explanatory variables.

- Usual logistic model for $\{y_{hi}\}$ can be written:

$$E\{y_{hi}=1\}=\pi_{hi}=\frac{\exp(\alpha_h+\beta x_{hi}+\gamma' z_{hi})}{1+\exp(\alpha_h+\beta x_{hi}+\gamma' z_{hi})}$$

Where α_h denotes the intercept for the *h*th center,

β is the treatment parameter,

$\boldsymbol{\gamma}' = (\gamma_1, \dots, \gamma_t)'$ is the parameter vector for the covariates $\boldsymbol{z}$.

- We can fit a model based on conditional probabilities that condition away the center effects, which results in a model that contains substantially fewer parameters. The α_h are known as *nuisance parameters*.

$$\Pr\{y_{h1}=1, y_{h2}=0 \mid y_{h1}=1, y_{h2}=0 \text{ or } y_{h1}=0, y_{h2}=1\}=$$

$$\frac{\Pr\{y_{h1}=1\}\Pr\{y_{h2}=0\}}{\Pr\{y_{h1}=1\}\Pr\{y_{h2}=0\}+\Pr\{y_{h1}=0)\Pr\{y_{h2}=1\}}$$

- Writing the probabilities in terms of the logistic model:

$$\Pr\{y_{h1}=1\}\Pr\{y_{h2}=0\}=\frac{\exp\{\alpha_h+\beta+\gamma'\mathrm{z}_{h1}\}}{1+\exp\{\alpha_h+\beta+\gamma'\mathrm{z}_{h1}\}}\times\frac{1}{1+\exp\{\alpha_h+\gamma'\mathrm{z}_{h2}\}}$$

and $\Pr\{y_{h1}=1\}\Pr\{y_{h2}=0\}+\Pr\{y_{h1}=0\}\Pr\{y_{h2}=1\}=$

$$\frac{\exp\{\alpha_h+\beta+\gamma'\mathrm{z}_{h1}\}}{1+\exp\{\alpha_h+\beta+\gamma'\mathrm{z}_{h1}\}}\times\frac{1}{1+\exp\{\alpha_h+\gamma'\mathrm{z}_{h2}\}}+\frac{1}{1+\exp\{\alpha_h+\beta+\gamma'\mathrm{z}_{h1}\}}\times\frac{\exp(\alpha_h+\gamma'\mathrm{z}_{h2}\}}{1+\exp(\alpha_h+\gamma'\mathrm{z}_{h2}\}}$$

Forming their ratio, and canceling like terms, the expression reduces to:

$$\frac{\exp\{\beta + \gamma'(z_{h1} - z_{h2})\}}{1 + \exp\{\beta + \gamma'(z_{h1} - z_{h2})\}}$$

Thus, by focusing on modeling a meaningful conditional probability, we develop a model with a reduced number of parameters that can be estimated without bias.

10.3 Clinical Trials Study Analysis

- In each of 79 clinics, one patient received new treatment for a skin condition, another placebo. Other variables collected: age, sex, initial grade for skin condition (ranged from 1 to 4 for mild to severe). Response was whether or not skin improved.

- Conditional logistic regression suitable (via the LOGISTIC procedure in SAS)

- Cross-tabulation of pairs by treatment and response:

Placebo Response	Treatment Response	
	No	**Yes**
No	**7**	**34**
Yes	**20**	**18**

- There are 20 discordant pairs of type No-Yes and 34 discordant pairs of type Yes-No. For asymptotic analysis, with 20 pairs of the one type, a conditional logistic model can support 20/5 ≈ 4 variables.

10.3.1 Example of matched pairs analysis using STRATA and CLASS statements in PROC LOGISTIC

- PROC LOGISTIC can be used to perform conditional logistic analyses in SAS version 9.3

- Advantage of PROC LOGISTIC in version 9.3 is direct operation on the actual observations (no need to create difference observations—see Appendix), use of a CLASS statement (no need to create indicator variables), and computation of odds ratios (no need to exponentiate parameter estimates by hand)

- Need to add STRATA statement to denote the conditioning variable

SAS version 9.3 code:

```
proc logistic data=trial;
     class sex(ref='f') treatment(ref='p') / param=ref;
     strata center;
     model improve(event='1') = initial age sex treatment
             sex*age sex*initial age*initial
             treatment*sex treatment*initial treatment*age /
             selection=forward include=4 details;
run;
```

Residual Score Statistics from PROC LOGISTIC version 9.3

Chi-Square	DF	Pr<ChiSq
4.7214	6	0.5800

- The residual chi-square test has p=0.5800, which does not support inclusion of the interaction terms in the model. The individual tests with one degree of freedom are displayed below:

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr>ChiSq
age*sex	1	0.6593	0.4168
initial*sex	1	0.1775	0.6736
initial*age	1	2.9195	0.0875
sex*treatment	1	0.2681	0.6046
initial*treatment	1	0.0121	0.9125
age*treatment	1	0.4336	0.5102

Model Fit Statistics from PROC LOGISTIC version 9.3

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	74.860	58.562
SC	74.860	70.813
-2 Log L	74.860	50.562

Global Fit Statistics from PROC LOGISTIC version 9.3

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.2976	4	<.0001
Score	19.8658	4	0.0005
Wald	13.0100	4	0.0112

Disagreement of p-values here implies need for exact analysis

Parameter Estimates from PROC LOGISTIC version 9.3

Analysis of Conditional Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
initial		1	1.0915	0.3351	10.6106	0.0011
age		1	0.0248	0.0224	1.2253	0.2683
sex	m	1	0.5312	0.5545	0.9176	0.3381
treatment	t	1	0.7025	0.3601	3.8053	0.0511

Odds Ratio Estimates from PROC LOGISTIC version 9.3

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
initial		2.979	1.545	5.745
age		1.025	0.981	1.071
sex	m vs f	1.701	0.574	5.043
treatment	t vs p	2.019	0.997	4.089

- PROC LOGISTIC version 9.3 provides odds ratios and corresponding confidence intervals
- Odds of improvement for those on treatment is $e^{0.7025} = 2.019$ times as high as the odds of improvement for those on placebo, adjusted for age and sex. 95% CI: (0.997, 4.089)
- Specifying `selection=forward` with `include=4` ensures that the initial skin grade (1 to 4), age, sex, and treatment main effects are included in the model. Here, none of the interaction terms were selected to be included in the model (score test p=0.58)
- Exact odds ratio estimates for treatment can be obtained with `exact` statement, but model must be re-run without `selection=forward`:

```
proc logistic data=trial;
     class sex(ref='f') treatment(ref='p') / param=ref;
     strata center;
     model improve(event='1') = initial age sex treatment;
     exact treatment /estimate=odds cltype=exact;
run;
```

Exact Odds Ratio Estimate for Treatment from PROC LOGISTIC v9.3

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value	Type
treatment t	1.943	0.950	4.281	0.0715	Exact

- Exact conditional analysis odds ratio estimate of 1.943 for treatment compared to placebo. 95% CI: (0.950, 4.281), p=0.0715

- Consider the model where the treatment is the only term:

```
proc logistic data=trial;
      class treatment(ref='p') / param=ref;
      strata center;
      model improve(event=‘1’) = treatment;
      exact treatment /estimate=odds cltype=exact;
run;
```

Maximum Likelihood Estimates from model only with Treatment Effect

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Treatment	1	0.5306	0.2818	3.5457	0.0597

- The odds ratio estimate is $e^{0.5306} = 1.70$

- Cross-tabulation of pairs by treatment and response:

Placebo Response	Treatment Response	
	No	**Yes**
No	**7**	**34**
Yes	**20**	**18**

- McNemar's test statistic is:

$$Q_M = \frac{(34-20)^2}{(34+20)} = 3.63$$

As sample size grows, Wald statistic for treatment and McNemar's test statistic become asymptotically equivalent

- Also note that $\frac{n_{12}}{n_{21}} = 1.7,$ which is the same as $e^{0.5306}$ which is the exact OR estimate in a treatment-only model

Let $h = 1, 2,\ldots,q$ index strata. Let $i = 1, 2$ index groups to be compared. Let $n_{hi} = 1$ be sample size for h,i. Let π_{hi} = Pr{response yes} for h,i. Let y_{hi} = 1 if response is yes, $y_{hi} = 0$ if response is no.

Likelihood: $\prod_{h=1}^{q}\prod_{i=1}^{2}\pi_{hi}^{y_{hi}}(1-\pi_{hi})^{1-y_{hi}}$

Logistic model: $\pi_{hi} = \dfrac{\exp(\mu+\xi_h+\beta_i)}{1+\exp(\mu+\xi_h+\beta_i)}$

$$\{\pi_{h1}/(1-\pi_{h1})\}/\{\pi_{h2}/(1-\pi_{h2})\} = \; e^{(\beta_1-\beta_2)}$$

Pr {Response = (yes, no) for group 1 and group 2 in stratum *h* given Response = [(yes, no) or (no, yes)]}

$$=\{\pi_{h1}(1-\pi_{h2})\}/\{\pi_{h1}(1-\pi_{h2}) + \; (1-\pi_{h1})\pi_{h2}\}$$

$$=\exp(\beta_1-\beta_2)\,/\{1+\exp(\beta_1-\beta_2)\}$$

$$\text{odds}\left\{\frac{\text{(yes, no)}}{\text{(no, yes)}}\right\} = \exp(\beta_1-\beta_2)$$

Exact Odds Ratio Estimate for Treatment from PROC LOGISTIC version 9.3

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value	Type
treatment t	1.700	0.951	3.117	0.0759	Exact

- Exact conditional analysis odds ratio estimate of 1.700 for treatment compared to placebo. 95% CI: (0.951, 3.117), p=0.0759

• Consider the exact analysis where the treatment and initial skin condition are included:

```
proc logistic data=trial exactonly;
      class treatment(ref='p') / param=ref;
      strata center;
      model improve(event=‘1’) = treatment initial;
      exact treatment initial / estimate=both;
run;
```

Exact Maximum Likelihood Estimates from model with Treatment Effect and Initial Skin Condition

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Two-sided p-value
Treatment	1	0.7034	0.3461	-0.005365	1.4836	0.0520
Initial	1	1.0542	0.3171	0.4625	1.8221	<0.0001

• The exact odds ratio estimate is $e^{0.7034} = 2.021$

Exact Odds Ratio Estimates from PROC LOGISTIC version 9.3

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		Two-sided p-Value
treatment t	2.021	0.995	4.409	0.0520
initial	2.870	1.588	6.185	<.0001

- Exact conditional analysis odds ratio estimate of 2.021 for treatment compared to placebo. 95% CI: (0.995, 4.409), p=0.0520

10.4 Crossover Design Studies

• In these designs, the study is divided into periods and patients receive a different treatment during each period. Thus, the patients act as their own controls. Interest lies in comparing treatments, adjusting for period and carryover effects.

10.4.1 Two-period Crossover Design

• Can be considered another example of paired data in the sense that there is a response for both Period 1 and Period 2.

Age	Sequence	Response Profiles				Total
		FF	FU	UF	UU	
Older	A:B	12	12	6	20	50
Older	B:P	8	5	6	31	50
Older	P:A	5	3	22	20	50
Younger	B:A	19	3	25	3	50
Younger	A:P	25	6	6	13	50
Younger	P:B	13	5	21	11	50

• Model the improvement for each patient in Period 1 vs. the probability of improvement in either period (but not both):

$$\frac{\Pr\{\text{Period1} = F\}\Pr\{\text{Period2} = U\}}{\Pr\{\text{Period1} = F\}\Pr\{\text{Period2} = U\} + \Pr\{\text{Period1} = U\}\Pr\{\text{Period2} = F\}}$$

• Analysis proceeds in the same manner as for the highly stratified paired data.

•Effects of interest are the period effect, effects for drugs A and B, and carryover effect for drugs A and B from Period 1 to Period 2. Using incremental effects parameterization.

- Note that there are 6 response functions, logits based on FU vs. UF, and thus 6 degrees of freedom with which to work. If we include the two effects for drugs A and B, the period effect, and the age × period effect, there are 2 d.f. left over. These can be used to explore the carryover or age × drug effects.

- The model employed includes carryover effects:

$$\Pr\{FU \mid FU \text{ or } UF\} = \frac{\exp\{\beta + \tau' z\}}{1 + \exp\{\beta + \tau' z\}}$$

where z consists of the difference between the two periods for period × age, Drug A, Drug B, Carry A and Carry B. The parameter β is the effect for period, τ_0 is the effect for period × age, τ_1 and τ_2 are the effects for Drug A and Drug B, and τ_3 and τ_4 are the effects for Carry A and Carry B.

• The model is specified through the implied structure for the difference between periods:

		Period 1	Period 2	(Period 1) – (Period 2)
Older	A:B	$\mu+\xi+\beta+\tau_0+\tau_1$	$\mu+\xi+\tau_2+\tau_3$	$\beta+\tau_0+\tau_1-\tau_2-\tau_3$
Older	B:P	$\mu+\xi+\beta+\tau_0+\tau_2$	$\mu+\xi+\tau_4$	$\beta+\tau_0+\tau_2-\tau_4$
Older	P:A	$\mu+\xi+\beta+\tau_0$	$\mu+\xi+\tau_1$	$\beta+\tau_0-\tau_1$
Younger	B:A	$\mu+\beta+\tau_2$	$\mu+\tau_1+\tau_4$	$\beta-\tau_1+\tau_2-\tau_4$
Younger	A:P	$\mu+\beta+\tau_1$	$\mu+\tau_3$	$\beta+\tau_1-\tau_3$
Younger	P:B	$\mu+\beta$	$\mu+\tau_2$	$\beta-\tau_2$

10.4.1.1 Two-Period Crossover Design -- Analysis Using the LOGISTIC Procedure in SAS version 9.3

Data can be specified in case-record format:

Obs	subject	period	age	seq	drug	response	carry
1	1	1	older	AB	A	F	P
2	1	2	older	AB	B	F	A
3	2	1	older	AB	A	F	P
4	2	2	older	AB	B	F	A
5	3	1	older	AB	A	F	P
6	3	2	older	AB	B	F	A
......							
369	185	1	young	BA	B	U	P
370	185	2	young	BA	A	F	B
371	186	1	young	BA	B	U	P
372	186	2	young	BA	A	F	B
373	187	1	young	BA	B	U	P
374	187	2	young	BA	A	F	B
. . . . .							

The syntax for the full model with carryover effects is

```
proc logistic data=cross2;
   class drug period age carry /param=ref;
   strata subject;
   model response = period drug period*age carry;
run;
```

Model Fit Statistics from PROC LOGISTIC version 9.3

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	166.355	129.579
SC	166.355	155.961
-2 Log L	166.355	117.579

Maximum Likelihood Estimates from PROC LOGISTIC v 9.3

Analysis of Conditional Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
period 1	1	-1.4370	0.7026	4.1832	0.0408
drug A	1	1.2467	0.6807	3.3547	0.0670
drug B	1	-0.00190	0.6412	0.0000	0.9976
period*age 1 older	1	0.6912	0.4654	2.2056	0.1375
carry A	1	-0.1903	1.1125	0.0293	0.8642
carry B	1	-0.5653	1.1556	0.2393	0.6247

Type 3 Analysis from PROC LOGISTIC v 9.3

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr>ChiSq
period	1	4.1832	0.0408
drug	2	4.5691	0.1018
period*age	1	2.2056	0.1375
carry	2	0.2450	0.8847

The 2 df Wald test of the carry-over effects has p=0.8847.

A reduced model without carry-over can be fit:

```
ods graphics on;
   proc logistic data=cross2;
     class drug period age / param=ref;
     strata subject;
     model response = period drug period*age;
     contrast 'A_B' drug 1 -1 /estimate=parm;
     oddsratio drug;
   run;
ods graphics off;
```

Model Fit Statistics from PROC LOGISTIC version 9.3

Model Fit Statistics

Criterion	Without Covariates	With Covariates
AIC	166.355	125.826
SC	166.355	143.413
-2 Log L	166.355	117.826

A likelihood ratio test (with 2 df) of the carryover effects may be conducted using the difference in -2 log L between the model with carryover and the model without:

-2 log L (full) =117.579
-2 log L (reduced) = 117.826

LR statistic = 117.826 – 117.579 = 0.247. For a chi-square distribution with df=2, this corresponds to p=0.8838. The Wald test had p=0.8847.

Maximum Likelihood Estimates from PROC LOGISTIC version 9.3

Analysis of Conditional Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Period 1	1	-1.1905	0.3308	12.9534	0.0003
drug A	1	1.3462	0.3289	16.7497	<.0001
drug B	1	0.2662	0.3233	0.6777	0.4104
period*age 1 older	1	0.7102	0.4576	2.4088	0.1207

Type 3 Analysis from PROC LOGISTIC version 9.3

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr>ChiSq
period	1	12.9534	0.0003
drug A	1	16.7497	<.0001
drug B	1	0.6777	0.4104
period*age	1	2.4088	0.1207

Contrast of Drug A vs. Drug B from PROC LOGISTIC version 9.3

Contrast Estimation and Testing Results

Contrast	Type	Est.	S.E.	Confidence Limits		Wald Chi-Square	Pr>ChiSq
A_B	PARM	1.080	0.327	0.440	1.721	10.9220	0.0010

The difference in the parameters for drug A and B is 1.080, which corresponds to an odds ratio of exp(1.080) = 2.945, with confidence limits [exp(0.440), exp(1.721)] = [1.552, 5.588]

Odds Ratio Estimates for Treatment Comparisons in PROC LOGISTIC version 9.3

Odds Ratio Estimates and Wald Confidence Intervals

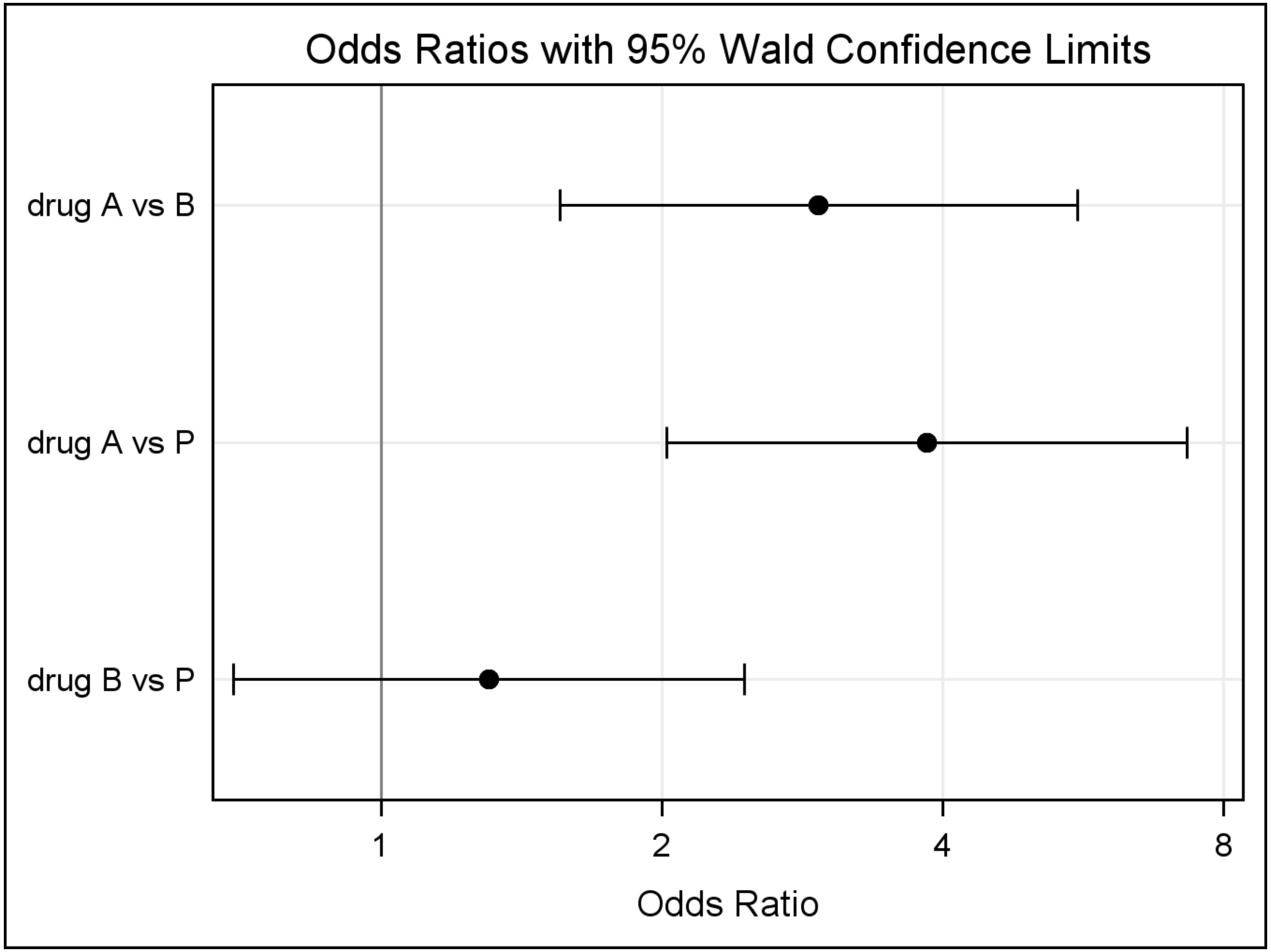
Label	Estimate	95% Confidence Limits	
drug A vs B	2.945	1.552	5.588
drug A vs P	3.843	2.017	7.322
drug B vs P	1.305	0.692	2.459

•Odds ratio estimate for comparison of Drug A to B is 2.945, 95% CI: (1.552, 5.588)

•Odds ratio estimate for comparison of Drug A to Placebo is 3.843, 95% CI: (2.017, 7.322)

•Odds ratio estimate for comparison of Drug B to Placebo is 1.305, 95% CI: (0.692, 2.459)

Graphical Display of Odds Ratios



Odds Ratios on Log 2 Scale

In crossover studies with r = 2 periods,

(y_{h1}, y_{h2}) has (0,0) as its only possible outcome when $(y_{h1} + y_{h2}) = 0$, and (y_{h1}, y_{h2}) has (1,1) as its only possible outcome when $(y_{h1} + y_{h2}) = 2$; and so these patterns are non-informative for the conditional likelihood for the estimation of β.

When $(y_{h1} + y_{h2}) = 1$, then (y_{h1}, y_{h2}) has (1,0) or (0,1) as its two possible outcomes, and their respective probabilities of occurrence are $\pi_{h1}(1 - \pi_{h2})$ and $(1 - \pi_{h1})\pi_{h2}$. The resulting contribution of such a patient to the conditional likelihood is

$$\frac{\Pr\{(y_{h1}, y_{h2}) = (1,0)\}}{\Pr\{(y_{h1} + y_{h2}) = 1\}} = \frac{\pi_{h1}(1-\pi_{h2})}{\pi_{h1}(1-\pi_{h2}) + (1-\pi_{h1})\pi_{h2}}$$

$$= \frac{\exp(x'_{h1}\beta)}{\exp(x'_{h1}\beta) + \exp(x'_{h2}\beta)}$$

$$= \frac{\exp\{(x_{h1} - x_{h2})'\beta\}}{1 + \exp\{(x_{h1} - x_{h2})'\beta\}}$$

Two period crossover study $\begin{cases} \text{Conditional Logistic Model} \\ \text{Gart Test} \end{cases}$

Treatment Seq	Response Evaluated at Periods (1,2)				No. of patients
	(F,F)	(F,U)	(U,F)	(U,U)	
A : P	20	16	5	9	50
P : A	16	6	18	10	50

F = Favorable, U = Unfavorable

Sequence	Period	Trmt.	Prop. Fav.	Model
A : P	1	A	36/50 = 0.72	τ
A : P	2	P	25/50 = 0.50	π
P : A	1	P	22/50 = 0.44	0
P : A	2	A	34/50 = 0.68	$\pi + \tau$

Conditional Logistic model $\pi_{hij} = e^{\alpha_h + \mu_{ij}} / (1 + e^{\alpha_h + \mu_{ij}})$

h : patients, i : sequence, j : period

$$\mu_{11} = \tau, \mu_{12} = \pi, \mu_{21} = 0, \mu_{22} = \pi + \tau$$

$$\Pr\{(\text{F,U}) \mid (\text{F,U}) \text{ or } (\text{U,F})\} = e^{\tau-\pi} / (1 + e^{\tau-\pi}) \text{ for A : P}$$

$$e^{-\tau-\pi} / (1 + e^{-\tau-\pi}) \text{ for P : A}$$

$$\frac{(\text{F,U})_{\text{A:P}} (\text{U,F})_{\text{P:A}}}{(\text{U,F})_{\text{A:P}} (\text{F,U})_{\text{P:A}}} \hat{=} e^{2\tau} \hat{=} \frac{16}{5} \times \frac{18}{6} \rightarrow e^{\tau} \hat{=} 3.1$$

Gart Test: H_0: $\tau = 0$ with Fisher's test p=0.001

Similarly, $\frac{16}{5} \times \frac{6}{18} \hat{=} e^{-2\pi} \hat{=} 1 \rightarrow e^{\pi} \hat{=} 1$

With $\pi = 0, (16 + 18) / (6 + 5) = 3.1 \hat{=} e^{\tau}$

10.4.2 Three-Period Crossover Study

• Exercise study in which subjects with chronic respiratory conditions were exposed to low, medium, and high air pollution while exercising on a stationary bike. Outcome: dichotomized as any respiratory distress (1,2, or 3) vs no distress (0). Baseline reading of no distress (0) or any distress (1).

Randomization Frequencies

Sequence	Frequencies	Percent
HLM	72	16.00
HML	78	17.33
LHM	72	16.00
LMH	72	16.00
MHL	60	13.33
MLH	96	21.33

• Conditional analysis of these data provides a way to detect within-subject effects (namely the pollution effect) and also investigates the period and carryover effects.

•For the three-period case, r = 3 and eight possible outcomes exist, two of which are non-informative $\left(\sum_{i=1}^{3} y_{hi} = 0, 3\right)$

When $\sum_{i=1}^{3} y_{hi} = 1, 2,$ there are three possible patterns for (y_{h1}, y_{h2}, y_{h3})

- The contributions to the conditional likelihood are:

$$\frac{\Pr\{y_{hi}=1, y_{hi'}=0 \text{ for all } i' \neq i\}}{\Pr\{y_{h1}+y_{h2}+y_{h3}=1\}} = \frac{\exp(x'_{hi}\beta)}{\sum_{i'=1}^{3}\exp(x'_{hi'}\beta)} \quad \text{for } i=1,2,3$$

for (1,0,0), (0,1,0), (0,0,1); and

$$\frac{\Pr\{y_{hi}=0, y_{hi'}=1 \text{ for all } i' \neq i\}}{\Pr\{y_{h1}+y_{h2}+y_{h3}=2\}} = \frac{\exp\left(\sum_{i'=1}^{3} x'_{hi'}\beta - x'_{hi}\beta\right)}{\sum_{i=1}^{3}\exp\left(\sum_{i'=1}^{3} x'_{hi'}\beta - x'_{hi}\beta\right)}$$

for (0,1,1), (1,0,1), and (1,1,0).

- Analysis first focuses on whether there is a carryover effect of exposure from an earlier period to a later period.
- Data coded as
    - Exposure: (L, M, H)
    - Period: (1,2,3)
    - Carry: (L, M, H)
    - Baseline: (Any distress at baseline=1, 0 otherwise),
    - Distress: (‘Any’, ‘None’)
- See page 321 for data manipulations.

10.4.2.1 Three-Period Crossover Design -- Analysis Using the LOGISTIC Procedure in SAS version 9.3

- PROC LOGISTIC (SAS version 9.3) code for obtaining results consistent with a dichotomous outcome of respiratory distress as ‘Any’ vs ‘None’:

```
proc logistic data=exercise descending;
    class period carry exposure /param=ref order=data;
    strata strata;
    model distress = exposure baseline period carry /include=2
                selection=forward details;
run;
```

- Residual Score Test of the period effects and carry-over effects has df=4 with p=0.9582. These terms are not included in the model, and estimates for the baseline and exposure variables are in output on the next slide.

Parameter Estimates from Model including Exposure and Baseline

Analysis of Conditional Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
exposure	h	1	2.2527	0.3983	31.9938	<.0001
exposure	m	1	0.6559	0.2547	6.6324	0.0100
Baseline		1	-0.4872	0.4457	1.1948	0.2744

•Model can be re-fit using include=1 option to evaluate whether the Baseline variable should enter the model:

```
proc logistic data=exercise descending;
     class exposure /param=ref order=data;
     strata strata;
     model distress = exposure baseline / selection=forward
               include=1 details;
run;
```

• Residual Score Test of Baseline has df=1 with p=0.2716. Baseline can be removed from the model and the model is re-fit:

```
proc logistic data=exercise descending;
     class exposure / param=ref order=data;
     strata strata;
     model distress = exposure;
     contrast ‘difference’ exposure 1 -1 / estimate=parm;
     oddsratio exposure;
run;
```

• The CONTRAST statement is a test of equivalence of the effects of high pollution and medium pollution. p<0.0001, indicating high pollution has a much stronger effect on response than medium pollution.

• The ODDSRATIO statement gives odds ratios for the exposure categories.

Odds Ratios for Model with only Exposure

Odds Ratio Estimates and Wald Confidence Intervals

Label	Estimate	95% Confidence Limits	
exposure high vs medium	4.968	2.250	10.970
exposure high vs low	9.617	4.403	21.006
exposure medium vs low	1.936	1.180	3.176

- Odds of any respiratory distress for high pollution exposure are 5 times as high as the odds for medium pollution. Odds of any distress for high pollution are about 10 times as high as the odds for low pollution. Odds of any distress for medium pollution are about twice the odds of any distress for low pollution.

- All confidence intervals exclude 1.0, indicating statistically significant effects for high vs. medium, high vs. low, and medium vs. low.

10.5 General Conditional Logistic Regression

- Consider the general model for stratified logistic regression:

$$\log\left\{\frac{\theta}{1-\theta}\right\} = \alpha_h + X\beta$$

- The α_h are stratum-specific parameters for each stratum ($h = 1, \ldots, q$). These are nuisance parameters and we eliminate them from the likelihood by conditioning on their sufficient statistic $T_0 = (T_{01}, \ldots, T_{0q})$ for which

$$T_{0h} = \sum_{i=1}^{n_h} y_{hi}$$

Where n_h is the number of observations from stratum h.

- Consider the model: $logit(\theta) = X_0\alpha + X\beta = X_A\beta_A$
- Partition the $(q + t) \times 1$ vector β_A into two components:

  α, the $q \times 1$ vector of stratum-specific intercepts

  β, the $t \times 1$ vector of parameters for variation within strata
- Partition X_A accordingly into X_0 and X.
- The sufficient statistics for α and β are $T_0 = X_0'y$ and

$T_1 = X'y$ where $y = (y_1', \ldots, y_q')'$ with $y_h' = (y_{h1}, \ldots, y_{hn_h})'$

• Conditional probability density function T_1 given $T_0 = t_0$

$$f_\beta(t_1 \mid t_0) = \frac{C(t_0, t_1) \exp(t_1'\beta)}{\sum_{u_1} C(t_0, u_1) \exp(u_1'\beta)}$$

where $C(t_0, u_1)$ are the number of y's such that $\{X_0'y = t_0, X_1'y = u_1\}$ for all possible values u_1 of T_1 when $T_0 = t_0$

• For this conditional likelihood function, apply an algorithm such as Newton-Raphson to obtain maximum likelihood estimates.

10.5.1 Analyzing Diagnostic Data

Time 1		Time 2		No. of Subjects
Standard	Test	Standard	Test	
Negative	Negative	Negative	Negative	509
Negative	Negative	Negative	Positive	4
Negative	Negative	Positive	Negative	17
Negative	Negative	Positive	Positive	3
Negative	Positive	Negative	Negative	13
Negative	Positive	Negative	Positive	8
Negative	Positive	Positive	Negative	0
Negative	Positive	Positive	Positive	8
Positive	Negative	Negative	Negative	14
Positive	Negative	Negative	Positive	1
Positive	Negative	Positive	Negative	17
Positive	Negative	Positive	Positive	9
Positive	Positive	Negative	Negative	7
Positive	Positive	Negative	Positive	4
Positive	Positive	Positive	Negative	9
Positive	Positive	Positive	Positive	170

• Two possible outcomes at 4 different combos of treatment and time ($r = 2^4 = 16$ response profiles).

•Can consider each subject to be a separate stratum, with 4 measurements in each stratum. Conditional logistic regression eliminates subject-to-subject variability.

•Effects of interest (time and treatment) are within-subject effects and can be handled by conditional logistic regression. If between-subject effects were of interest (such as age, sex), we'd need a different strategy.

• See pages 327-328 of text to input the data set as `diagnosis`

```
data diagnosis2; set diagnosis;
   drop std1 test1 std2 test2;
   subject=_n_;
   time=1; procedure='standard'; response=std1; output;
   time=1; procedure='test'; response=test1; output;
   time=2; procedure='standard'; response=std2; output;
   time=2; procedure='test'; response=test2; output;
run;

proc logistic data=diagnosis2;
   class time (ref=first) procedure (ref=first)/ param=ref;
   strata subject;
   model response(event='Neg') = time procedure time*procedure;
run;
```

Parameter Estimates for Full Model

Analysis of Conditional Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
time 2	1	-0.0625	0.2500	0.0625	0.8026
procedure test	1	0.3848	0.2544	2.2881	0.1304
time*procedure 2 test	1	0.4726	0.3630	1.6952	0.1929

Main effects model:

```
proc logistic data=diagnosis2;
   class time (ref=first) procedure (ref=first)/ param=ref;
   strata subject;
   model response(event='Neg') = time procedure time*procedure
        /selection=forward include=2 details;
run;
```

Score Statistic for test of interaction

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
1.7002	1	0.1923

Parameter Estimates for Main Effects Model

Analysis of Conditional Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
time	2	1	0.1627	0.1807	0.8114	0.3677
Procedure	test	1	0.6159	0.1836	11.2557	0.0008

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
time	2 vs 1	1.177	0.826	1.677
procedure	test vs standard	1.851	1.292	2.653

•Re-run without `selection=forward` and add `exact` statement:

```
proc logistic data=diagnosis2;
   class time (ref=first) procedure (ref=first)/ param=ref;
   strata subject;
   model response(event='Neg') = time procedure;
   exact procedure / estimate=odds cltype=exact;
run;
```

Exact Odds Ratio Estimate for Procedure

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value	Type
procedure test	1.849	1.274	2.703	0.0009	Exact

•Exact odds ratio estimate for Test procedure versus Standard procedure is 1.849, 95% CI (1.274, 2.703), p=0.0009

10.6 1:1 Conditional Logistic Regression

- Researchers studied women in a retirement community in the 1970s to determine if there was an association between the use of estrogen and the incidence of endometrial cancer.

- Each case was matched with a control who was within a year of the same age, had the same marital status, and was living in the same community at the time of the diagnosis of the case.

- Explanatory variables:

  GALL=1 if gallbladder disease history, 0 otherwise
  EST=1 if estrogen use, 0 otherwise
  HYPER=1 if hypertensive, 0 otherwise
  AGE = age in years
  NONEST=1 if non-estrogen drug use, 0 otherwise

- CASE = 1 if case, 0 if control

```
proc logistic data=match;
  strata id;
  model case(event='1') = gall est hyper age nonest
        /selection=forward details;
run;
```

Residual Score Statistic

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.2077	3	0.9763

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
hyper	1	0.0186	0.8915
age	1	0.1432	0.7051
nonest	1	0.0370	0.8474

Parameter Estimates

Analysis of Conditional Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
gall	1	1.6551	0.7980	4.3017	0.0381
est	1	2.7786	0.7605	13.3492	0.0003

- Odds ratio for endometrial cancer is $e^{2.7786} = 16.096$ for those taking estrogen vs. those not taking estrogen.

Odds Ratio Estimates

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
gall	5.234	1.095	25.006
est	16.096	3.626	71.457

•Re-run without `selection=forward` and add `exact` statement:

```
proc logistic data=match;
  strata id;
  model case(event='1') = gall est;
  exact gall est /estimate=both;
run;
```

Exact Odds Ratio Estimate for Estrogen Taking

Exact Odds Ratios

Parameter		Estimate	95% Confidence Limits		p-Value	Type
est	1	15.066	3.701	133.346	<.0001	Exact

•Exact odds ratio estimate for Estrogen users vs. Estrogen non-users is 15.066, 95% CI (3.701, 133.346), $p<0.0001$

10.7 1:m Conditional Logistic Regression

• Researchers in a midwestern county tracked flu cases requiring hospitalization in residents aged ≥ 65 during two-month period.

• Each case was matched with two controls according to sex and age (1 : 2 matched study). Researchers determined whether subjects had flu vaccine shots and whether they had lung disease.

• Researchers interested in whether vaccination had protective influence on odds of getting severe case of flu.

• OUTCOME = 1 if case, 0 if control
LUNG=1 if Lung Disease, 0 if not
VACCINE=1 if Vaccine, 0 if not

```
proc freq;
     tables outcome*lung outcome*vaccine / nocol nopct;
run;
```

Frequencies of Vaccine and Smoking by Cases and Controls

Table of outcome by lung

Outcome / Frequency / Row Pct	Lung: No Lung Disease	Lung Disease	Total
Case	87 58.00	63 42.00	150
Control	252 84.00	48 16.00	300
Total	339	111	450

Table of outcome by vaccine

Outcome / Frequency / Row Pct	Vaccine: No Vaccine	Vaccine	Total
Case	103 68.67	47 31.33	150
Control	183 61.00	117 39.00	300
Total	286	164	450

```
proc logistic data=matched;
  class lung vaccine;
  strata id;
  model outcome(event='1') = lung vaccine lung*vaccine
       /selection=forward include=2 details;
run;
```

Residual Score Statistic

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > ChiSq
lung*vaccine	0.0573	0.8107

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
0.0573	1	0.8107

Parameter Estimates

Analysis of Conditional Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
lung	1	1	1.3053	0.2348	30.8967	<.0001
vaccine	1	1	-0.4008	0.2233	3.2223	0.0726

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
lung	1 vs 0	3.689	2.328	5.845
vaccine	1 vs 0	0.670	0.432	1.038

- Odds ratio for getting a case of flu resulting in hospitalization is $e^{-0.4008} = 0.67$ for those with vaccine vs. those without vaccine. Study participants with vaccine reduced their odds of getting hospitalizable flu by 33% compared to their non-vaccinated counterparts.

•Re-run without `selection=forward` and add `exact` statement:

```
proc logistic data=matched;
  class lung vaccine;
  strata id;
  model outcome(event='1') = lung vaccine;
  exact vaccine /estimate=odds cltype=exact;
run;
```

Exact Odds Ratio Estimate for Vaccine

Exact Odds Ratios

Parameter		Estimate	95% Confidence Limits		p-Value	Type
vaccine	1	0.671	0.420	1.057	0.0886	Exact

•Exact odds ratio estimate for those getting vaccine vs. those not getting vaccine is 0.671, 95% CI (0.420, 1.057), p=0.0886

10.8 Exact Conditional Logistic Regression in the Stratified Setting

• In exact setting (used when data are sparse), same methodology is used. Only difference: in the unstratified case, you don't have stratification variables and so condition away only explanatory variables; in the stratified case, condition away both stratification and explanatory variables.

•Example: Cardiovascular study of 8 animals who received various drug treatments. Researchers arrested coronary flow → ischemia; recorded whether an adverse cardiovascular event occurred during 8-minute interval. Reperfused, and repeated for up to five measurements per animal.

- Because of sequence of treatments, not assumed to be a crossover study. Because of reperfusion, period and carryover effects considered ignorable.

- Drug effect assumed to be ordinal with equally spaced intervals

```
data cardio;
input animal treatment $ response $ @@;
if treatment = ‘S’ then delete;
else if treatment = ‘C’    then ordtreat = 1;
else if treatment = ‘DA’   then ordtreat = 2;
else if treatment = ‘D1’   then ordtreat = 3;
else if treatment = ‘D2’   then ordtreat = 4;
datalines;
 1    S   No     1   C     No     1   C      No     1   D2     Yes    1   D1    Yes
 2    S   No     2   D2    Yes    2   C      No     2   D1     Yes
 3    S   No     3   C     Yes    3   D1     Yes    3   DA     No     3   C     No
 4    S   No     4   C     No     4   D1     Yes    4   DA     No     4   C     No
 5    S   Yes    5   C     No     5   DA     No     5   D1     No     5   C     No
 6    S   No     6   C     No     6   D1     Yes    6   DA     No     6   C     No
 7    S   No     7   C     No     7   D1     Yes    7   DA     No     7   C     No
 8    S   Yes    8   C     Yes    8   D1     Yes
 ;
 proc logistic data=cardio descending exactonly;
         strata animal;
         model response =  ordtreat;
         exact  ordtreat / estimate = both;
 run;
```

Exact Tests

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-value --- Exact	Mid
Ordtreat	Score	10.4411	0.0009	0.0005
	Probability	0.000723	0.0009	0.0005

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		Two-sided p-Value
Ordtreat	6.974	1.620	198.976	0.0017

• Compare the score test for the exact stratified analysis to the score test for the asymptotic stratified analysis. To do this, specify the `selection=forward` option with `details`:

```
proc logistic data=cardio descending;
  strata animal;
  model response = ordtreat / selection=forward details
       slentry=0.05;
run;
```

Residual Score Test

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
10.4411	1	0.0012

Parameter Estimate and Odds Ratio with Wald *p*-value

Analysis of Conditional Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq
Ordtreat	1	1.9421	0.8932	4.7275	0.0297

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
ordtreat	6.974	1.211	40.159

•Note that Wald asymptotic *p*-value (0.0297) is greater than the exact *p*-value (0.0017), but the score *p*-value (0.0012) is smaller than the exact *p*-value.