Logistic regression in matched case-control studies

HL Chapter 7

Unmatched data

- Un-adjusted analysis
  - Probability of the outcome given the risk factor
- Example
  - Probability of lung cancer among smokers and non-smokers

Unmatched data – adjusting for age

- Age-adjusted analysis
  - Probability of the outcome given the risk factor assuming cases and controls are the same with respect to the confounder
- Example
  - Probability of lung cancer among smokers and non-smokers assuming cases and controls have equal ages

Matched data

- Controls are matched to cases based on one or more confounders (e.g., age)
- Matched analysis
  - Probability that the subject "identified" by the risk factor as the case really is the case
- This handout focuses on 1-1 matching (1 control selected for each case)

Example: Pairs 1-4 of 6 pairs Cases and controls matched on age

Pair ID	Lung cancer		Smoking		Age
1	1	Case	1	Smoker	60
1	0	Control	0	Non-smoker	60
2	1	Case	1	Smoker	68
2	0	Control	0	Non-smoker	68
3	1	Case	1	Smoker	59
3	0	Control	0	Non-smoker	59
4	1	Case	1	Smoker	72
4	0	Control	0	Non-smoker	72

Probability that the smoker is the case
= 4 pairs/6 pairs = 2/3 (pairs 1-4)

Example, cont: Pairs 5 & 6 of 6 pairs Cases and controls matched on age

Pair ID	Lung cancer		Smoking		Age
5	1	Case	0	Non-smoker	83
5	0	Control	1	Smoker	83
6	1	Case	0	Non-smoker	69
6	0	Control	1	Smoker	69

Probability that the non-smoker is the case
= 2 pairs/6 pairs = 1/3 (pairs 5-6)

Problem

- With a large data set and a large number of model covariates, the "calculate-by-hand" approach is not feasible
- We can use conditional logistic regression to get the same results
  - Step 1: Define conditional likelihood
  - Step 2: Run conditional logistic regression
  - Step 3: Plug coefficient resulting from step 2 into equation in step 1

Step 1: Conditional likelihood

- The probability that the subject "identified" by the covariate values as the case really is the case is referred to as the conditional likelihood and can be calculated as

$$l_k(\beta_1,\beta_2,...,\beta_p) = \frac{\exp(\beta_1 x_{1k,case}+...+\beta_p x_{pk,case})}{\exp(\beta_1 x_{1k,case}+...+\beta_p x_{pk,case})+\exp(\beta_1 x_{1k,control}+...+\beta_p x_{pk,control})}$$

p= # of model covariates, k= # of the case-control pair

Example: Conditional likelihood, pairs 1-4

- For pairs (i.e., strata) *k*=1-4
  - The smoker is the case ($smo_{Case\ k}=1$)
  - The non-smoker is the control ($smo_{Control\ k}=0$)

$$l_k(\hat{\beta}_{smo})\frac{\exp(\hat{\beta}_{smo}\times smo_{Case\ k})}{\exp(\hat{\beta}_{smo}\times smo_{Case\ k})+\exp(\hat{\beta}_{smo}\times smo_{Control\ k})}$$

$$=l_{smoker\ is\ case}(\hat{\beta}_{smo}) = \frac{\exp(\hat{\beta}_{smo}\times 1)}{\exp(\hat{\beta}_{smo}\times 1)+\exp(\hat{\beta}_{smo}\times 0)}$$

Example: Conditional likelihood, pairs 5 & 6

- For pairs (i.e., strata) *k*=5, 6
  - The non-smoker is the case ($smo_{Case\ k}=0$)
  - The smoker is the control ($smo_{Control\ k}=1$)

$$l_k(\hat{\beta}_{smo})\frac{\exp(\hat{\beta}_{smo}\times smo_{Case\ k})}{\exp(\hat{\beta}_{smo}\times smo_{Case\ k})+\exp(\hat{\beta}_{smo}\times smo_{Control\ k})}$$

$$=l_{non-smoker\ is\ case}(\hat{\beta}_{smo}) = \frac{\exp(\hat{\beta}_{smo}\times 0)}{\exp(\hat{\beta}_{smo}\times 0)+\exp(\hat{\beta}_{smo}\times 1)}$$

Step 2: Conditional logistic regression

- We use conditional logistic regression to estimate the model coefficient(s)
- In SAS, conditional logistic regression can be performed using proc logistic with the STRATA command
- Note: The matching variable should not be a model covariate except in interaction terms

Example

```
data test;  input pair lc smo age;
cards;
1 1 1 60
1 0 0 60
2 1 1 68
2 0 0 68
3 1 1 59
3 0 0 59
4 1 1 72
4 0 0 72
5 1 0 83
5 0 1 83
6 1 0 69
6 0 1 69
run;
```

```
proc logistic descending data=test;
  model lc=smo;
 strata pair;
run;
```

pair = name of the stratum ID variable in the data set

SAS results

Analysis of Conditional Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Squ	Pr > ChiSq
smo	1	0.6931	0.8660	0.6406	0.4235

Step 3: Conditional likelihood, pairs 1-4

$$l_{smoker\ is\ case}(\hat{\beta}_{smo}) = \frac{\exp(\hat{\beta}_{smo} \times 1)}{\exp(\hat{\beta}_{smo} \times 1) + \exp(\hat{\beta}_{smo} \times 0)}$$

$$= \frac{\exp(0.6931 \times 1)}{\exp(0.6931 \times 1) + \exp(0.6931 \times 0)} = 0.666 = 2/3$$

Agrees with hand calculation on slide 5

Step 3: Conditional likelihood, pairs 5 & 6

$$l_{non-smoker\ is\ case}(\hat{\beta}_{smo}) = \frac{\exp(\hat{\beta}_{smo} \times 0)}{\exp(\hat{\beta}_{smo} \times 0) + \exp(\hat{\beta}_{smo} \times 1)}$$

$$= \frac{\exp(0.6931 \times 0)}{\exp(0.6931 \times 0) + \exp(0.6931 \times 1)} = 0.333 = 1/3$$

Agrees with hand calculation on slide 6

That's nice, but what about an OR?

- From SAS:

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
smo	2.000	0.366	10.919

- Note: $e^{0.6931} = 2$
- Interpretation: The lung cancer risk is twice as high among smokers than among non-smokers

Definitions

- Discordant pairs
  - In the smoking & lung cancer example, the case and control of each pair have different values for smoking
  - These case-control pairs are discordant
- Concordant pairs
  - The value of smoking may be the same for the case and the control in a pair
  - These case-control pairs are concordant

Estimating model coefficients

- Concordant pairs cannot be used to estimate model coefficients
- To estimate the model coefficient(s) we need both types of discordant pairs
  - Type 1 pair: Smoker is case, non-smoker is control
  - Type 2 pair: Non-smoker is case, smoker is control
- If all discordant pairs are of the same type, the model cannot be fit

Estimating model coefficients

- Note: The complete absence of one type of discordant pairs in a matched analysis corresponds to a zero cell in an unmatched analysis

Example: 1-1 matched GLOW study

- The data set was created from the GLOW500 data set by randomly selecting an age matched control for each case
- Variables
  - PAIR (pair id)
  - FRACTURE (outcome variable)
  - AGE (matching variable)
  - HEIGHT, WEIGHT, BMI (continuous)
  - PRIORFRAC, PREMENO, MOMFRAC, ARMASSIST, SMOKE, RATERISK (categorical)

Frequency of discordant pairs

PRIORFRAC	Frequency	Percent
Non-exposed=case, exposed=control	16	13.45
Concordant	66	55.46
Exposed=case, non-exposed=control	37	31.09

ARMASSIST	Frequency	Percent
Non-exposed=case, exposed=control	17	14.29
Concordant	70	58.82
Exposed=case, non-exposed=control	32	26.89

Frequency of discordant pairs

MOMFRAC	Frequency	Percent
Non-exposed=case, exposed=control	12	10.08
Concordant	87	73.11
Exposed=case, non-exposed=control	20	16.81

PREMENO	Frequency	Percent
Non-exposed=case, exposed=control	7	5.88
Concordant	98	82.35
Exposed=case, non-exposed=control	14	11.76

Frequency of discordant pairs

SMOKE	Frequency	Percent
Non-exposed=case, exposed=control	7	5.88
Concordant	107	89.92
Exposed=case, non-exposed=control	5	4.20

Frequency of discordant pairs

- Both types of discordant pairs are present for all variables
- But…even though the data set has 119 pairs, estimation of the coefficients is based on fewer pairs

E.g.,

- The coefficient for PREMENO is based on only 21 pairs (7+14)
- The coefficient for SMOKE is based on only 12 pairs (7+5)

Descriptive statistics for the continuous variables

Fract	Pairs	Variable	Mean	SD	Min	Q1	Med	Q3	Max
No	119	Height	161.8	5.4	150.0	158.0	162.0	165.0	175.0
		Weight	71.1	16.8	39.9	59.0	67.1	81.6	115.7
		BMI	27.1	5.9	15.0	22.6	26.1	30.7	42.2
Yes	119	Height	159.8	6.9	134.0	155.0	160.0	164.0	178.0
		Weight	70.8	15.8	45.8	59.9	68.0	79.4	124.7
		BMI	27.7	5.9	17.0	23.0	26.4	31.1	44.0

- No obvious incorrect values

Extreme values for the continuous variables

		Lowest				
FRACT	Variable	1	2	3	4	5
No	HEIGHT	150	150	152	152	152
	WEIGHT	39.9	40.8	43.1	44.9	45.4
	BMI	15.0	17.1	17.4	18.4	18.5
Yes	HEIGHT	134	142	143	147	148
	WEIGHT	45.8	46.3	47.6	48.1	48.1
	BMI	17.1	18.4	18.5	20.0	20.1

- No obvious incorrect values

Extreme values for the continuous variables

		Highest				
FRACT	Variable	1	2	3	4	5
No	HEIGHT	173	173	173	175	175
	WEIGHT	104.3	105.2	110.7	112.0	115.7
	BMI	39.6	40.5	40.7	41.0	42.2
Yes	HEIGHT	173	173	173	175	178
	WEIGHT	111.1	111.6	113.4	117.0	124.7
	BMI	41.7	41.7	43.4	43.6	44.0

- No obvious incorrect values

Univariate coefficients,p-values, ORs and 95% CIs

Variable	Coeff	Std Err	p-value	Unit	OR	95% CI	
HEIGHT	-0.057	0.0238	0.0160	10	0.564	0.354	0.899
WEIGHT	-0.001	0.0084	0.8696	3	0.996	0.948	1.046
BMI	0.019	0.0229	0.4048	3	1.059	0.925	1.212
PRIORFRAC	0.838	0.2992	0.0051	1	2.312	1.286	4.157
PREMENO	0.693	0.4629	0.1343	1	2.000	0.807	4.955
MOMFRAC	0.511	0.3651	0.1618	1	1.667	0.815	3.409
ARMASSIST	0.633	0.3001	0.0351	1	1.882	1.045	3.390
SMOKE	-0.337	0.5855	0.5655	1	0.714	0.227	2.251
RATERISK2	0.552	0.2909	0.0578	1	1.737	0.982	3.071
RATERISK3	1.025	0.3669	0.0052	1	2.787	1.357	5.720

WEIGHT, BMI, SMOKE are non-significant at the 0.25 level

Univariate splines

- Graphs do not suggest extreme deviations from linearity (not shown)
- Re-check scale in multivariate model

Multivariate model 1

To match the text, include WEIGHT, BMI AND SMOKE in the first multivariate model.

Variable	Coeff	Std Err	p-value
HEIGHT	0.0633	0.1220	0.6042
WEIGHT	-0.1542	0.1310	0.2392
BMI	0.3865	0.3417	0.2580
PRIORFRAC	0.6935	0.3538	0.0500
PREMENO	0.2180	0.5523	0.6931
MOMFRAC	0.7254	0.4326	0.0936
ARMASSIST	0.8178	0.3824	0.0325
RATERISK2	0.1516	0.3412	0.6569
RATERISK3	0.5888	0.4256	0.1665

Too many variables measuring the same thing

Height, weight, BMI

Height, weight and BMI added *one at a time*
to the model with
PRIORFRAC, PREMENO, MOMFRAC, ARMASSIST, RATERISK2&3

	Wald p-value	Deviance
HEIGHT	0.0078	138.2
WEIGHT	0.2294	145.2
BMI	0.9696	146.9

- Only HEIGHT is statistically significant

Height, weight, BMI

Height, weight and BMI added *in sets of two*
to the model with
PRIORFRAC, PREMENO, MOMFRAC, ARMASSIST, RATERISK2&3

	Wald p-value	Deviance
HEIGHT	0.0138	138.4
WEIGHT	0.5785	
HEIGHT	0.0069	138.5
BMI	0.6528	
WEIGHT	0.0045	137.3
BMI	0.0098	

- Could choose HEIGHT only or WEIGHT and BMI
- The text chooses WEIGHT and BMI

Multivariate model 2

Variable	Coeff	Std Err	p-value	Unit	OR	95% CI	
BMI	0.2163	0.0837	0.0098	3	1.913	1.170	3.131
WEIGHT	-0.0888	0.0312	0.0045	3	0.766	0.638	0.921
PRIORFRAC	0.6897	0.3542	0.0515	1	1.993	0.995	3.991
PREMENO	0.2212	0.5529	0.6891	1	1.248	0.422	3.687
MOMFRAC	0.7477	0.4308	0.0826	1	2.112	0.908	4.914
ARMASSIST	0.8229	0.3810	0.0308	1	2.277	1.079	4.805
RATERISK 2	0.1302	0.3369	0.6992	1	1.139	0.589	2.205
RATERISK 3	0.5712	0.4240	0.1779	1	1.770	0.771	4.064

Multivariate model 3

Variable	Coeff	Std Err	P-value	Units	OR	95% CI	
BMI	0.2170	0.0833	0.0092	3	1.918	1.175	3.129
WEIGHT	-0.0894	0.0311	0.0040	1	0.765	0.637	0.918
PRIORFRAC	0.6974	0.3541	0.0489	1	2.009	1.003	4.020
MOMFRAC	0.7389	0.4283	0.0845	1	2.094	0.904	4.847
ARMASSIST	0.8518	0.3737	0.0226	1	2.344	1.127	4.875
RATERISK2	0.1489	0.3337	0.6554	1	1.161	0.603	2.232
RATERISK3	0.5972	0.4190	0.1541	1	1.817	0.799	4.130

No big changes in any OR
after removing PREMENO

Multivariate model 4

Variable	Coeff	Std Err	P-value	Units	OR	95%	CI
BMI	0.2224	0.0810	0.0060	3	1.949	1.211	3.138
WEIGHT	-0.0947	0.0299	0.0016	3	0.753	0.631	0.898
PRIORFRAC	0.8349	0.3396	0.0140	1	2.305	1.184	4.484
MOMFRAC	0.7266	0.4093	0.0759	1	2.068	0.927	4.612
ARMASSIST	0.8888	0.3666	0.0153	1	2.432	1.186	4.990

- Some change in OR of PRIORFRAC after removing RATERISK
- However, 95% CI is quite wide and the change is likely noise
- MOMFRAC is borderline non-significant; keep for now

Scale of weight – fp method

Dev_ linear	e_ fp1	Dev_ fp1	e1_ fp2	e2_ fp2	Dev_ fp2	p_lin_ fp1	p_lin_ fp2	p_fp1_ fp2
139.7	.	.	-2	-2	135.6	.	0.2496	.

- Best one-power = linear
- Best two-power = $\frac{1}{weight^2}$ and $\frac{1}{weight^2} \times \ln(weight)$
- Best two-power not significantly better than linear

Scale of BMI – fp method

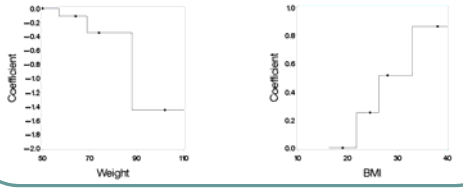
Dev_ linear	e_ fp1	Dev_ fp1	e1_ fp2	e2_ fp2	Dev_ fp2	p_lin_ fp1	p_lin_ fp2	p_fp1_ fp2
139.7	0	138.9	3	-2	136.6	0.3686	0.3752	0.3166

- Best one-power = $\ln(BMI)$
- Best two-power = BMI^3 and $\frac{1}{BMI^2}$

- Best one-power not significantly better than linear
- Best two-power not significantly better than linear or best one-power

Scale of weight and BMI Design variable plots

- Small sample size; therefore
  - When testing weight, keep BMI linear
  - When testing BMI, keep weight linear
- Linear OK for weight and BMI



Final main effects model

Variable	Coeff	Std Err	P-value	Units	OR	95%	CI
BMI	0.2224	0.0810	0.0060	3	1.949	1.211	3.138
WEIGHT	-0.0947	0.0299	0.0016	3	0.753	0.631	0.898
PRIORFRAC	0.8349	0.3396	0.0140	1	2.305	1.184	4.484
MOMFRAC	0.7266	0.4093	0.0759	1	2.068	0.927	4.612
ARMASSIST	0.8888	0.3666	0.0153	1	2.432	1.186	4.990

The text suggests that this is the best main effects model

Significance of interaction terms when added to the main effects model one at a time

Parameter	p-value
weight*bmi	0.5108
weight*priorfrac	0.2270
weight*momfrac	0.7843
weight*armassist	0.8075
bmi*priorfrac	0.2270
bmi*momfrac	0.7994
bmi*armassist	0.9325

Parameter	p-value
priorfrac*momfrac	0.6993
priorfrac*armassist	0.0829
momfrac*armassist	0.2427
weight*age	0.5492
bmi*age	0.5056
priorfrac*age	0.0905
momfrac*age	0.4790
armassist*age	0.8416

2 interaction terms are significant at the 0.1 level but are not used in the text

Matching variable, age, appears in interaction terms but it is NOT a main effect

Goodness-of-fit

- Cannot easily perform goodness of fit tests
- The distributional assumptions of goodness-of-fit tests that are currently available in statistical software packages do not hold in matched case-control studies

Outliers - Limitations

For conditional logistic regression

- Plots option is not available in SAS
- Cannot use covariate patterns
- Can only use leverage and (standardized) Pearson chi-square residual

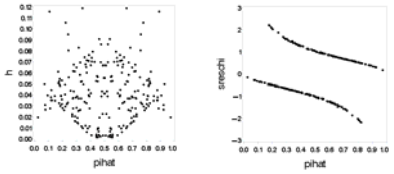
Outliers in SAS

```
proc logistic descending data=GLOW11M;
  model fracture=weight bmi priorfrac momfrac armassist;
  strata pair;
  output out=diag h=h stdreschi=sreschi p=pihat;
run;

axis1 label=(f=swiss h=3 'pihat') minor=none;
axis2 label=(f=swiss h=3 a=90 'h') minor=none;
axis3 label=(f=swiss h=3 a=90 'sreschi') minor=none;
```

Outliers in SAS

```
goptions FTEXT=swissb HTEXT=2.0
            HSIZE=6 in VSIZE=6 in;
symbol1 c=black v=dot i=none;

proc gplot data=diag;
  plot h*pihat/haxis=axis1 vaxis=axis2;
run; quit;
proc gplot data=diag;
  plot sreschi*pihat/haxis=axis1 vaxis=axis3;
run; quit;
```

Outlier plots



- h: Note that the right half is a mirror image of the left half
- sreschi: Note that the bottom curve is a mirror image of the top curve

Selecting outliers

- Show observations with h>0.11
- Show observations with sreschi>2.1 (or < -2.1)

- Cutpoints were selected based on plots
- Could choose different cutpoints to get more outliers

Explaining outliers: Outliers based on h

PAIR	FRACTURE	WEIGHT	BMI	PRIOR FRAC	MOM FRAC	ARM ASSIST	pihat	h	sres chi
49	0	61.7	23.22	0	0	0	0.66	0.12	-1.48
49	1	124.7	39.36	1	0	1	0.34	0.12	1.48
87	1	74.8	41.66	0	0	1	0.89	0.12	0.37
87	0	65.8	25.07	0	1	1	0.11	0.12	-0.37

Pairs 49 and 87

- 49: Case has an unusually high weight
- 87: Case has an unusually high BMI, especially given the weight

- Pairs 49 and 87 have little effect on goodness-of-fit (sreschi)

Explaining outliers: Outliers based on sreschi

PAIR	FRACTURE	WEIGHT	BMI	PRIORFRAC	MOMFRAC	ARMASSIST	pihat	h	sreschi
37	1	72.6	26.03	0	0	0	0.19	0.04	2.11
37	0	64.4	25.80	0	1	0	0.81	0.04	-2.11
38	1	72.6	26.67	0	0	0	0.18	0.03	2.20
38	0	52.2	21.18	1	0	0	0.82	0.03	-2.20
117	1	57.2	23.81	1	0	0	0.18	0.05	2.18
117	0	50.8	20.61	1	1	1	0.82	0.05	-2.18

Pair 37

- Case has MOMFRAC=0 while control has MOMFRAC=1
- Case weighs more than control (but the more you weigh the lower your risk)
- pihat for case=0.19 but pihat for control=0.81
- Control looks more like a case and vice versa

Pair 38

- Case has PRIORFRAC=0 while control has PRIORFRAC=1
- Case weighs more than control (but the more you weigh the lower your risk)
- pihat for case=0.18 but pihat for control=0.82
- Control looks more like a case and vice versa

Pair 117

- Case has MOMFRAC=0 and ARMASSIST=0 while control has MOMFRAC=1 and ARMASSIST=1
- Case weighs more than control (but the more you weigh the lower your risk)
- pihat for case=0.18 but pihat for control=0.82
- Control looks more like a case and vice versa

ORs before and after deleting outliers

Variable	All in	Delete 37	Delete 38	Delete 117	Delete all 3
Weight	0.910	0.902	0.904	0.906	0.892
BMI	1.249	1.275	1.260	1.257	1.298
PRIORFRAC	2.305	2.321	2.601	2.328	2.685
MOMFRAC	2.068	2.427	2.090	2.415	2.966
ARMASSIST	2.432	2.522	2.521	2.778	3.056

- Weight and BMI: No big changes
- PRIORFRAC, MOMFRAC, ARMASSIST: ORs tend to increase as outliers are deleted

p-values before and after deleting outliers

Variable	All	Delete 37	Delete 38	Delete 117	Delete all 3
Weight	0.0016	0.0009	0.0010	0.0012	0.0004
BMI	0.0060	0.0036	0.0052	0.0053	0.0026
PRIORFRAC	0.0140	0.0140	0.0067	0.0140	0.0065
MOMFRAC	0.0759	0.0378	0.0747	0.0384	0.0164
ARMASSIST	0.0153	0.0129	0.0133	0.0073	0.0048

- P-values tend to decrease as outliers are deleted
- However, outliers don't have unreasonable values and don't have huge effects on the ORs
- We'll keep the outliers

Final model

Variable	Coeff	Std Err	p-value	Unit	OR	95%	CI
Weight	-0.0947	0.0299	0.0016	3	0.753	0.631	0.898
BMI	0.2224	0.0810	0.0060	3	1.949	1.211	3.138
PRIORFRAC	0.8349	0.3396	0.0140		2.305	1.184	4.484
MOMFRAC	0.7266	0.4093	0.0759		2.068	0.927	4.612
ARMASSIST	0.8888	0.3666	0.0153		2.432	1.186	4.990

Final model interpretation

- For each 3kg increase in weight, the fracture risk decreases about 25%
- For each 3 unit increase in BMI, the fracture risk almost doubles
- Persons with a prior fracture have a 2.3 fold increased fracture risk compared to persons without a prior fracture

Final model interpretation

- Persons with a mother who had a hip fracture have a 2 fold increased fracture risk compared to persons without a mother who had a hip fracture
- Persons who need arms to stand up from chair have an almost 2.5 fold increased fracture risk compared to persons who don't need their arms

Final model interpretation

- The 95% CIs are wide
- The 95% CI for MOMFRAC includes 1
- The study's sample size was small
- The number of discordant pairs was small for the categorical variables