STAT 5500/6500 Conditional Logistic Regression for Matched Pairs

The data for the tutorial came from support.sas.com, The LOGISTIC Procedure: Conditional Logistic Regression for Matched Pairs Data :: SAS/STAT(R) 9.2 User's Guide, Second Edition http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect062.htm which is a subset of data taken from the Los Angeles Study of the Endometrial Cancer Data in Breslow and Day (1980). 63 cases of subjects with endometrial cancer (cancer = 1) were matched to a cancer free control subject (cancer = 0). The matching was done by age, marital status, non-hysterectomy status and living in the community at the time the case was diagnosed. The goal of the case-control analysis was to determine the relative risk of gall bladder disease while controlling for the effect of hypertension (high blood pressure.)

In matched case control studies each case is matched with a control subject based on variables that could affect the response but is not necessarily of interest to the research. Factors such as age, gender, race etc. are taken into consideration when matching. Because matching is subject specific each case-control pair potentially has a different probability of risk. Performing a logistic regression analysis on this would result in needing dummy variables for each pair! Doing so results in too many fixed effects to estimate with respect to the sample size and leads to biased estimates.

Logistic Regression Model (McKnight 427):

$$\operatorname{logit}(p) = \alpha + \underbrace{\alpha_2 x_{s2} + \cdots + \alpha_m x_{sm}}_{\text{Indicators for many small strata. Not of interest per se.}} + \underbrace{\beta_1 x_1 + \cdots + \beta_k x_k}_{\text{Variables of interest and any additional adjustment variables}}$$

Consider a generalized table of cases (i.e. cancer present) and controls observing a certain factor X (i.e. occurrence of gall bladder disease) of interest:

		Cases (Y = 1)	
Controls (Y = 0)	X = 1	X = 0	
X = 1	a	b	a + b
X = 0	c	d	c + d
	a + c	b + d	

This is the design that is used in McNemar's test. In fact, conditional logistic regression is considered an extension of McNemars test procedure as well as an extension of logistic regression.

Now, consider the partial tables obtained by reversing the roles of X and Y for each subject. The following tables show the possible case-control pair outcomes:

	a		b		c		D	
X	Case	Control	Case	Control	Case	Control	Case	Control
Yes	1	1	0	1	1	0	0	0
No	0	0	1	0	0	1	1	1

It is on these partial tables that conditional logistic regression is performed. For each table, the letter (a,b,c, or d) specifies the number of these that is represented by the paired data. For example, we would expect "a" number of tables where both the case and the control had gall bladder disease. In our example, we would have a total of 63 tables made up of combinations of such partial tables.

Conditional logistic regression stratifies on matching pairs where each stratum has its own intercept α_i. This allows each match to have its own risk of the event or outcome where a larger α_i indicates greater risk of event. Case matching can have more than 1 control and different numbers of controls but each match must have 1 case. Strata indicators are referred to as nuisance parameters because they are not of interest from a research standpoint but can result in confounding if not taken into consideration.

For subject i with our 2X2 tables, $logit[P(y_i = 1)] = \alpha_i + \beta x$ and, when matched-pairs response has k predictors, this extends to $logit[P(y_i = 1)] = \alpha_i + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

(Agresti 249-252)

The conditional likelihood is the product of probability functions over the j strata (or tables) (Bandyopadhyay 24):

$$L^c(\vec{y}) = \frac{\prod_{j=1}^{J} \binom{n_j}{y_j}}{\sum_{\vec{y}\in\Gamma} \prod_{j=1}^{J} \binom{n_j}{y_j}}$$

Taking stratification into account “conditions out” the stratum specific intercepts and so we obtain estimates of the slopes only which represent different effects of interest. We can see this with the likelihood contribution of the jth strata for a 1:1 pair matching (http://courses.washington.edu/b536/Archive/handouts/Lecture12.pdf 3):

$$\frac{e^{\alpha_j + X_{case}\,\beta}}{e^{\alpha_j + X_{case}\,\beta} + e^{\alpha_j + X_{control}\,\beta}} = \frac{e^{\alpha_j}}{e^{\alpha_j}} \quad \frac{e^{\,X_{case}\,\beta}}{e^{\,X_{case}\,\beta} + e^{\,X_{control}\,\beta}}$$

$$= \frac{e^{\,X_{case}\,\beta}}{e^{\,X_{case}\,\beta} + e^{\,X_{control}\,\beta}}$$

The conditional MLE, $\hat{\beta}$ is found by maximizing $L^c(\beta)$ (Bandyopadhyay 24). $\hat{\beta}$ will be approximately unbiased and approximately normal if sample size (meaning the number # of individual strata) is large.

In our example, the data was read into SAS where matching was indicated by the variable ID and the STRATA statements was included in the logistic regression procedure. The log odds estimate for gall bladder disease is .9704 with a p-value = .0675. For hypertension, the log odds estimate is .3481 with a p-value of .3558. These values need to be exponentiated for interpretation. After controlling for hypertension, the odds that a subject with gall bladder disease (x=1) is a cancer case is equal 2.64 times the odds that a subject without gall bladder disease (x=0) has cancer. Or, the odds a subject with gall bladder disease is a cancer case is 164% more than the odds of a subject without gall bladder disease. Neither the coefficient for gall bladder disease or hypertension is significant at α = .05 though gall bladder disease does provide some evidence (< .10.)

If the number strata is small, an exact test is available in SAS. Running this on the sample data we get a log odds estimate for gall bladder disease equals .9530 with a p-value = .0969. For hypertension, the log odds estimate is .3425 with a p-value of .4622. Exponentiating on the gall bladder disease coefficient we get an odds ratio of 2.593. These are all similar to the results in the asymptotic method.

Since hypertension was non-significant, it might be worth removing it from the model. Doing so gives an odds ratio estimate of 2.6, again not appreciably different from when hypertension was included in the model which is not surprising since it was not significant in the original fitting of the model.

Other applications of conditional logistic regression include matched prospective studies where individuals are matched by age, gender, race, etc. but two different treatments are given and response to treatment is analyzed. Also, cross-over trials where the same patient is given two different drugs and response to treatment is analyzed. Here the strata consists of two binary measurements on same subject. In genetics, gene environment interactions can be analyzed by observing whether alleles of interest is passed from parents to child with the ability to account for environmental factors. The case here is the allele getting passed and the control is the unpassed allele and the strata's are made of individual families.

The upside to this method is that it gives better precision in the slope estimates, it works well with sparse (within strata) data and it also has a lot of flexibility for matching and assessing effects of interest. Interactions between variables can be examined including those used in the matching. With regards to the matching, cases can be matched to multiple controls and different number of controls from case to case. On the down side, matching is complicated, data must be prepared case by case for SAS and the method does not handle large numbers (within strata) well.

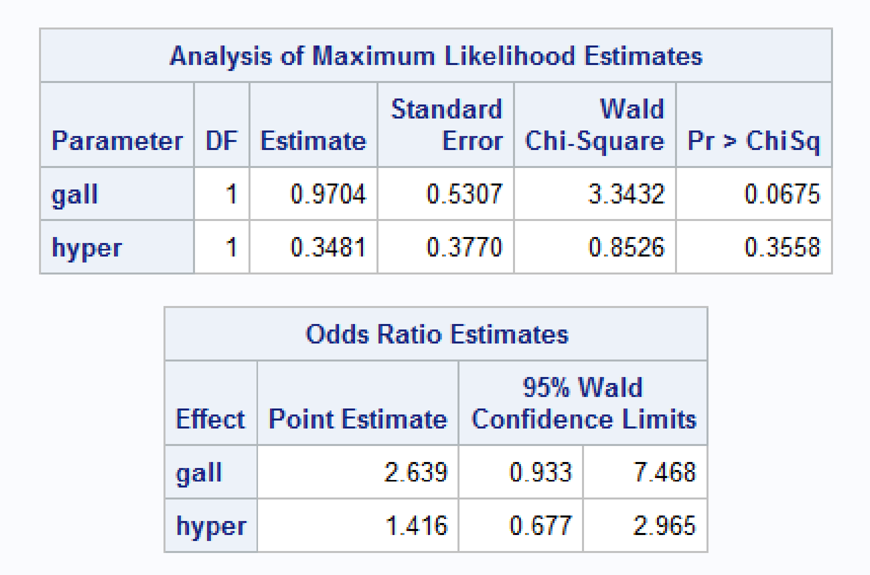
Appendix:

SAS code used and partial outputs

```
  data Data1; input ID cancer gall hyper @@;
   cards;
   1  1 0  0  1  0  0  0   2  1 0  0  2  0  0  0
   3  1 0  1  3  0  0  1   4  1 0  0  4  0  1  0
   5  1 1  0  5  0  0  1   6  1 0  1  6  0  0  0
   7  1 1  0  7  0  0  0   8  1 1  1  8  0  0  1
   9  1 0  0  9  0  0  0  10  1 0  0 10  0  0  0
  11  1 1  0 11  0  0  0  12  1 0  0 12  0  0  1
  13  1 1  0 13  0  0  1  14  1 1  0 14  0  1  0
  15  1 1  0 15  0  0  1  16  1 0  1 16  0  0  0
  17  1 0  0 17  0  1  1  18  1 0  0 18  0  1  1
  19  1 0  0 19  0  0  1  20  1 0  1 20  0  0  0
  21  1 0  0 21  0  1  1  22  1 0  1 22  0  0  1
  23  1 0  1 23  0  0  0  24  1 0  0 24  0  0  0
  25  1 0  0 25  0  0  0  26  1 0  0 26  0  0  1
  27  1 1  0 27  0  0  1  28  1 0  0 28  0  0  1
  29  1 1  0 29  0  0  0  30  1 0  1 30  0  0  0
  31  1 0  1 31  0  0  0  32  1 0  1 32  0  0  0
  33  1 0  1 33  0  0  0  34  1 0  0 34  0  0  0
  35  1 1  1 35  0  1  1  36  1 0  0 36  0  0  1
  37  1 0  1 37  0  0  0  38  1 0  1 38  0  0  1
  39  1 0  1 39  0  0  1  40  1 0  1 40  0  0  0
  41  1 0  0 41  0  0  0  42  1 0  1 42  0  1  0
  43  1 0  0 43  0  0  1  44  1 0  0 44  0  0  0
  45  1 1  0 45  0  0  0  46  1 0  0 46  0  0  0
  47  1 1  1 47  0  0  0  48  1 0  1 48  0  0  0
  49  1 0  0 49  0  0  0  50  1 0  1 50  0  0  1
  51  1 0  0 51  0  0  0  52  1 0  1 52  0  0  1
  53  1 0  1 53  0  0  0  54  1 0  1 54  0  0  0
  55  1 1  0 55  0  0  0  56  1 0  0 56  0  0  0
  57  1 1  1 57  0  1  0  58  1 0  0 58  0  0  0
  59  1 0  0 59  0  0  0  60  1 1  1 60  0  0  0
  61  1 1  0 61  0  1  0  62  1 0  1 62  0  0  0
  63  1 1  0 63  0  0  0
  ;
  run;

/*conditional logistic regression code defining event of interest*/
  proc logistic data = Data1;
  strata ID;
  model cancer(event = '1') = gall hyper;
run;
```

```
/*alternate way to code to get same result*/
  proc logistic data = Data1 descending;
  strata ID;
  model cancer = gall hyper;
run;
```



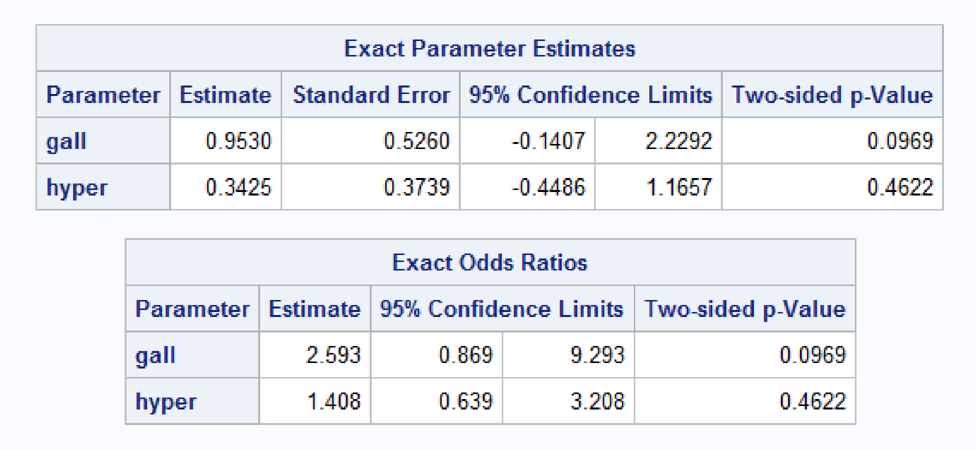
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
gall	1	0.9704	0.5307	3.3432	0.0675
hyper	1	0.3481	0.3770	0.8526	0.3558

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
gall	2.639	0.933	7.468
hyper	1.416	0.677	2.965

```
/*code for exact test*/;
proc logistic data=Data1 exactonly;
    strata ID;
    model cancer(event='1')=gall hyper;
    exact gall hyper / estimate=both;
run;
```



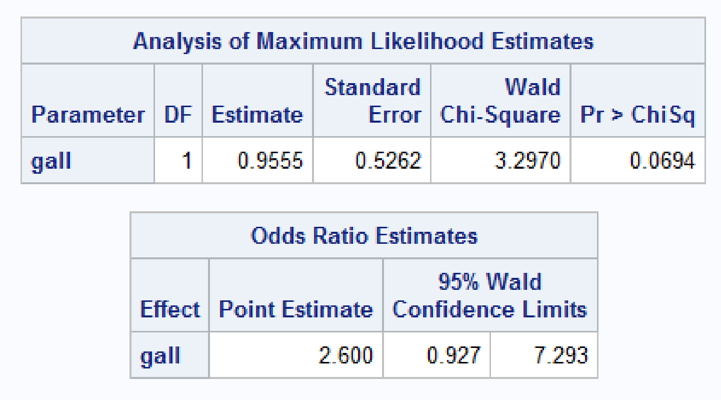
Exact Parameter Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Two-sided p-Value
gall	0.9530	0.5260	-0.1407	2.2292	0.0969
hyper	0.3425	0.3739	-0.4486	1.1657	0.4622

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		Two-sided p-Value
gall	2.593	0.869	9.293	0.0969
hyper	1.408	0.639	3.208	0.4622

```
/*conditional logistic regression code gallbladder only*/
  proc logistic data = Data1;
  strata ID;
  model cancer(event = '1') = gall;
run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
gall	1	0.9555	0.5262	3.2970	0.0694

Odds Ratio Estimates			
		95% Wald	
Effect	Point Estimate	Confidence Limits	
gall	2.600	0.927	7.293

Data taken from:

The LOGISTIC Procedure: Conditional Logistic Regression for Matched Pairs Data :: SAS/STAT(R) 9.2 User's Guide, Second Edition http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect062.htm

Reference Material:

The LOGISTIC Procedure: Conditional Logistic Regression for Matched Pairs Data :: SAS/STAT(R) 9.2 User's Guide, Second Edition http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect062.htm

Bandyopadhyay, Dipankar. N.p.. Web. 9 Dec 2013. <http://www.biostat.umn.edu/~dipankar/bmtry711.11/lecture_19.pdf>.

McKnight, Barbara.*http://courses.washington.edu/b536/Archive/2009/private/notes/13-conditional.ppt.pdf*. N.p.. Web. 9 Dec 2013.

Agresti, Alan. *An Introduction to Categorical Data Analysis*. 2nd Editions. Hoboken, New Jersey, U.S.A.: John Wiley & Sons, Inc., 249-252. Print.

"http://courses.washington.edu/b536/Archive/handouts/Lecture12.pdf." . N.p.. Web. 9 Dec 2013.