Lecture 15 (Part 2): Logistic Regression & Common Odds Ratio, (With Simulations)

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Mantel-Haenszel Estimator of Common Odds Ratio

- Mantel and Haenszel also proposed an estimator of the common odds ratio
- For table $W = j$, the observed odds ratio is

$$\widehat{OR}_j^{XY.W} = \frac{y_{j11}y_{j22}}{y_{j21}y_{j12}}$$

- If there is a common OR across tables, we could estimate the common OR with a 'weighted estimator':

$$\widehat{OR}_{MH} = \frac{\sum_{j=1}^{J} w_j \widehat{OR}_j^{XY.W}}{\sum_{j=1}^{J} w_j},$$

for some 'weights' w_j. (Actually, any weight will give you an asymptotically unbiased estimate).

- Mantel-Haenszel chose weights

$$w_j = \frac{y_{j21} y_{j22}}{y_{j++}}$$

when $OR_j^{XY \cdot W} = 1$, giving

$$\widehat{OR}_{MH} = \frac{\sum_{j=1}^{J} y_{j11} y_{j22} / y_{j++}}{\sum_{j=1}^{J} y_{j21} y_{j12} / y_{j++}}$$

- A good (consistent) estimate the variance of $\log[\widehat{OR}_{MH}]$ is (Robbins, et. al, 1985), based on a Taylor series expansion,

$$\widehat{Var}[\log \widehat{OR}_{MH}] =$$

$$\frac{\sum_{j=1}^{J} P_j R_j}{2[\sum_{j=1}^{J} R_j]^2} + \frac{\sum_{j=1}^{J} P_j S_j + Q_j R_j}{2[\sum_{j=1}^{J} R_j][\sum_{j=1}^{J} S_j]} + \frac{\sum_{j=1}^{J} Q_j S_j}{2[\sum_{j=1}^{J} S_j]^2},$$

where

$$P_j = (Y_{j11} + Y_{j22})/Y_{j++}$$

$$Q_j = (Y_{j12} + Y_{j21})/Y_{j++}$$

$$R_j = \frac{Y_{j11} Y_{j22}}{Y_{j++}}$$

$$S_j = \frac{Y_{j12} Y_{j21}}{Y_{j++}}$$

which is given in SAS.

Notes about M-H estimate

- 1. This estimate is easy to calculate (non-iterative), although its variance estimate is a little more complicated.
- 2. Asymptotically normal and unbiased with large strata (strata sample size y_{j++} large).
- 3. When each y_{j++} is large, the Mantel-Haensel estimate is not as efficient as the MLE, but close to MLE for logistic regression, which is iterative. When each y_{j++} is small, the MLE from logistic model could have a lot of bias.
- 4. Just like the Mantel-Haenzel statistic, unlike the logistic MLE, this estimator actually works well when the strata sample sizes are small (y_{j++} small), as long as the number of strata J is fairly large. (When doing large sample approximations, something must be getting large, either y_{j++} or J, or both).

Example - Age, Vaccine, Paralysis Data

- We showed earlier that the logistic regression estimate of the 'common odds ratio' between VACCINE (X) and PARALYSIS (Y) controlling for AGE (W) is

$$\exp(\hat{\boldsymbol{\beta}}) = \exp(1.2830) = 3.607,$$

  with a 95% confidence interval,

  `[1.791,7.266]`

  which does not contain 1. Thus, controlling for 'age', individuals who take the vaccine have 3.6 times the odds of not getting POLIO than individuals who do not take the vaccine.

- The Mantel-Haenzel Estimator of the common Odds Ratio is

$$\widehat{OR}^{MH} = 3.591$$

  with a 95% confidence interval of

  `[1.781,7.241]`

- Thus, individuals who take the vaccine have about 3.6 times the odds of not getting POLIO than individuals who do not take the vaccine.

Confounding

There are four useful diagnostics for potential confounding of the effect of a predictor (e.g., treatment) of interest:

1. The potential confounder must be associated with the outcome
2. The potential confounder must be associated with the predictor of interest (Note: Randomization in a clinical trial minimizes this)
3. Adjustment for the potential counfounder must affect the magnitude of the coefficient estimate for the predictor of interest
4. The potential confounder must make sense in terms of the hypothetical causal framework

Confounding in Logistic Regression

- Here, we are interested in using logistic regression to see if W confounds the relationship between X and Y.
- For simplicity, suppose we have 3 dichotomous variables
- In the logistic regression model,

$$w = \begin{cases} 1 \\ 0 \end{cases} . \qquad x = \begin{cases} 1 \\ 0 \end{cases} . \qquad Y = \begin{cases} 1 \\ 0 \end{cases} .$$

- The logistic regression model of interest is

$$\text{logit}\{P[Y = 1|w, x]\} = \beta_0 + \alpha w + \beta x.$$

The conditional odds ratio between Y and X given W is

$$\exp(\beta) = OR^{XY.W}.$$

- The marginal odds ratio between Y and X can be obtained from logistic regression model

$$\text{logit}\{P[Y = 1|x_]\} = \beta_0^* + \beta^* x,$$

  and is

$$\exp(\beta^*) = OR^{XY}.$$

- If there is no confounding, then

$$\beta = \beta^*$$

- Basically, you can fit both models, and, if

$$\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\beta}}^*,$$

  then you see that there is no confounding.

More Formal Check of Confounding of W

- However, to be more formal about checking for confounding, one would check to see if
- 1. W and Y are conditionally independent given X,
  or
- 2. W and X are conditionally independent given Y.
- To check these two conditions, you could fit a logistic model in which you make W the response, and X and Y covariates;

$$\text{logit}\{P[W = 1|x, Y]\} = \alpha_0 + \tau x + \alpha y,$$

- In this model, α is the conditional log-odds ratio between W and Y given X, and is identical to α in the logistic model with Y as the response and W and x as the covariates,

$$\alpha = \log[OR^{WY.X}]$$

- Also, τ is the conditional log-odds ratio between W and X given Y

$$\tau = \log[OR^{WX.Y}].$$

- Thus, if there is no confounding, the test for one of these two conditional OR's equalling 0 would not be rejected, i.e., you would either not reject $\alpha = 0$, or you would not reject $\tau = 0$.

Alternative Procedure

- However, if it was up to me, if you really want to see if there is confounding, I would just fit the two models:

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

and

$$\text{logit}\{P[Y=1|x]\} = \beta_0^* + \beta^* x,$$

and see if

$$\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\beta}}^*$$

- Rule of thumb in Epidemiology is that

$$\left| \frac{\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*}{\hat{\boldsymbol{\beta}}^*} \right| \leq 20\%?$$

- If there were many other covariates in the model, this is probably what you would do.

If $J > 2$, then you would fit the two models

$$\text{logit}\{P[Y = 1|W = j, X = x]\} = \beta_0 + \alpha_j + \beta x$$

and

$$\text{logit}\{P[Y = 1|X = x]\} = \beta_0^* + \beta^* x,$$

and see if

$$\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\beta}}^*$$

Notes about Models

- In journal papers, the analysis with the model

$$\text{logit}\{P[Y=1|x]\} = \beta_0^* + \beta^* x,$$

  is often called **univariate** (or unadjusted) analysis (the univariate covariate with the response)

- The analysis with the model

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

  is often called a **multivariate** analysis (more than one covariate with the response).

- Strictly speaking,

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

  is a **multiple** logistic regression analysis.

- In general, you state the results from a multiple regression as **adjusted** ORs.

Bias vs. Variance Trade off

- In the presence of confounding, the model not controlling for the confounding variable produces a biased estimate of key predictor (“treatment”)
- Introducing additional parameters to control for the confounding reduces this bias
- However, including too my parameters (i.e., “overfitting” the data) may cause the model to be too closely tied to the data at hand (poor predictive ability)
- There is a trade off between too few parameters (and some bias) and poor prediction (in terms of large standard errors for estimated parameters)
- Lets examine these notations using simulations

- Suppose you fit the two models, and there is no confounding,
- Then, in the models

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

and

$$\text{logit}\{P[Y=1|x]\} = \beta_0^* + \beta^* x,$$

we have

$$\beta = \beta^*$$

- Suppose, even though there is no confounding, W is an important predictor of Y, and should be in the model.
- Even though $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^*$ are both asymptotically unbiased (since they are both estimating the same β), you can show that

$$Var(\hat{\boldsymbol{\beta}}) \quad \leq \quad Var(\hat{\boldsymbol{\beta}}^*)$$

$$\text{FULLER} \quad \leq \quad \text{REDUCED}$$

Quasi-proof

- Heuristically, this is true because W is explaining some of the variability in Y that is not explained by X alone,
- and thus, since more variability is being explained, the variance of the estimates from the fuller model (with W) will be smaller.

Suppose $\alpha = 0$.

- Now, suppose that, in real life, you have overspecified the model, i.e., $\alpha = 0$, so that W and Y are conditionally independent given X, i.e., the true model is

$$\text{logit}\{P[Y = 1|w, x]\} = \beta_0^* + \beta^* x$$

- However, suppose you estimate (β_0, α, β) in the model

$$\text{logit}\{P[Y = 1|w, x]\} = \beta_0 + \alpha w + \beta x$$

  you are estimating β from an 'overspecified' model in which we are (unnecessarily) estimating α, which is 0.

- In this case, $\hat{\boldsymbol{\beta}}$ from the overspecified model will still be asymptotically unbiased, however estimating a parameter α that is 0 actually adds more error to the model, and

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &\geq Var(\hat{\boldsymbol{\beta}}^*) \\ \text{FULLER} &\geq \text{REDUCED} \end{aligned}$$

Simulation Framework

- Prior to writing SAS code, lets start with the basic logistic model

$$P[Y_i = 1|X\beta] = \frac{exp(X_i\beta)}{1 + exp(X_i\beta)}$$

- In a simulation, the vector β is known and assumed to be constant
- Simulation will vary the values in the vector X_i
- Given the linear term, $X\beta$, we would generate a random Bernoulli trial (binomial w/ n=1) with success probability defined above
- Generate many random data sets, analyze each and summarize
- It's "simple"

```
dm "output" clear; dm "log" clear;
options nocenter nostimer nonotes pageno=1;

/* The true model is defined to be
pi_i = alpha + beta_1 tx + beta_2 confound + error
where TX = 1 if treatment, 0 control
      confound = 1 if confounder is present, 0 else
*/
%macro generate_random(seed,alpha,txeffect, x1, coneffect, x2);
    tx = &x1;
    confound = &x2;
    txeffect = &txeffect;
    coneffect = &coneffect;
    xbeta =&alpha + &txeffect * tx + &coneffect * confound;
    pi_i = exp(xbeta)/(1+exp(xbeta));
    y = ranbin(&seed,1,pi_i);
%mend generate_random;

%macro SIMDATA(loops,seed,subjects,
               alpha,cutvaluetx, cutvalueplacebo,
               txeffect,coneffect);
data tempsim;
run; /*erase working file*/
%do i = 1 %to &loops;
data sim2;
 num_with_confound_placebo = floor(&cutvalueplacebo * &subjects);
 num_without_placebo = &subjects - num_with_confound_placebo;
 num_with_confound_tx = floor(&cutvaluetx * &subjects);
 num_without_tx = &subjects - num_with_confound_tx;

  /* generate placebo  - with confounder*/
   do j = 1 to num_with_confound_placebo;
     %generate_random(&seed,&alpha,&txeffect,0,&coneffect,1);
    output;
   end;

   /* generate placebo  - without confounder*/
   do j = 1 to num_without_placebo;
     %generate_random(&seed,&alpha,&txeffect,0,&coneffect,0);
    output;
   end;
  /* generate tx  - with confounder*/
   do j = 1 to num_with_confound_tx;
     %generate_random(&seed,&alpha,&txeffect,1,&coneffect,1);
    output;
   end;
/* generate tx  - without confounder*/
   do j = 1 to num_without_tx;
   %generate_random(&seed,&alpha,&txeffect,1,&coneffect,0);
```

```
    output;
   end;
run;

ods listing close;
ods results off;
ods output ParameterEstimates=tmyParm_twofact ;
proc logistic descending data=sim2;
 model y = tx confound;
run;


ods output ParameterEstimates=tmyParm_txfact ;
proc logistic descending data=sim2;
 model y = tx ;
run;

ods listing;
ods results on;

data t2myparm_twofact;
 set tmyparm_twofact;
 model = 1;
 simnum = &i;
run;

data t2myparm_txfact;
 set tmyparm_txfact;
 model = 2;
 simnum = &i;
run;


data tempsim;
 set tempsim t2myparm_twofact t2myparm_txfact ;
 if simnum = . then delete;
run;
%end; /* end of the do loop for simulation iterations */
data simresults;
 set tempsim;
 if ProbChiSq ge 0.05 then decision = 0;
  else decision = 1;
  Label decision = "Reject H0 = 1, fail to reject H0 = 0";
 if variable = "tx" then do;
   if (estimate - 1.96*stderr ) < &txeffect and
   (estimate + 1.96*stderr) > &txeffect then Covered = 1;
     else covered = 0;
    end;
    if variable = "confound" then do;
```

```
   if (estimate - 1.96*stderr ) < &coneffect and
      (estimate + 1.96*stderr) > &coneffect then Covered = 1;
     else covered = 0;
    end;

run;

title2 "Summary of Rejection Rate of Null Hypothesis (Power)";
proc means data=simresults n mean ;
 class model variable;
  var decision Covered estimate stderr;
run;



%mend SIMDATA;
options nomprint;
/*%SIMDATA(loops, = number of simulation replicates
            seed, = random simulation seed - values other than 0 non-random
            subjects, = sample size PER arm
            alpha, = intercept
            cutvaluetx,  = % of subjects in treatment arm with confounding variable
            cutvalueplacebo, = % of subjects in placebo arm with confounding variable
            txeffect, = beta for treatment (assumes dummy coding {1, 0} at time of ran
                        number generate
            coneffect = beta for confounding effect ({1,0} coding)
             );*/

title "No confound effect: 1:1 confound allocation, OR_TX = 3 (N=200)";
%SIMDATA(200,0,200,-1.504,.5, .5,log(3),0);
```

This will run 200 simulation replicates with a random initialization seed for a sample size of 200. 50% of the sample will be in treatment 1 (active) and 50% will be in the control. The confounding variable will not be associated with outcome (beta =0) and is not associated with the treatment (equally distributed across treatment)

Simulation 1 Results

model	Variable	N Obs	Variable	Mean
(Overspecified)				
1	Intercept	200	decision	1.0000000
			Covered	.
			Estimate	-1.5251040
			StdErr	0.2177174
	confound	200	decision	0.0450000
			Covered	0.9550000
			Estimate	0.0098527
			StdErr	0.2283039
	tx	200	decision	0.9900000
			Covered	0.9400000
			Estimate	1.1042132
			StdErr	0.2352918

model	Variable	N Obs	Variable	Mean
(correctly specified)				
2	Intercept	200	decision	1.0000000
			Covered	.
			Estimate	-1.5161111
			StdErr	0.1848402
	tx	200	decision	0.9900000
			Covered	0.9400000
			Estimate	1.1014665
			StdErr	0.2349531

Summary results for over specified model

```
StdErr for TX overspecified          0.2352918
StdErr for TX correctly specified    0.2349531
```

Note that the overspecified standard error is slightly larger

The true $\beta_{tx} = log(3) = 1.0986$ and 94% of the repeated samples contained this value. We also rejected the null hypothesis 99% of the time (i.e., Power = 99%)

Same simulation as before, larger N

```
title "No confound effect: 1:1 confound allocation, OR_TX = 3 (N=1000)";
%SIMDATA(200,0,1000,-1.504,.5, .5,log(3),0);
```

model	Variable	Obs	Variable	Mean
(incorrect)				
1	Intercept	200	decision	1.0000000
			Covered	.
			Estimate	-1.5013574
			StdErr	0.0964341
	confound	200	decision	0.0500000
			Covered	0.9500000
			Estimate	-0.0026302
			StdErr	0.1014900
	tx	200	decision	1.0000000
			Covered	0.9700000
			Estimate	1.0982165
			StdErr	0.1044118

```
       model    Variable      Obs    Variable      Mean
-------------------------------------------------------
(correct)
           2    Intercept     200    decision  1.0000000
                                     Covered   .
                                     Estimate -1.5018095
                                     StdErr    0.0819981

                tx            200    decision  1.0000000
                                     Covered   0.9700000
                                     Estimate  1.0976297
                                     StdErr    0.1043794
-------------------------------------------------------
```

Bias

```
title "Small confound effect: 1:3 confound allocation, OR_TX = 3 (N=200)";
%SIMDATA(200,0,200,-1.504,.25, .75,log(3),log(.50));
```

Here: we use an unequal allocation of placebo, tx w/ and w/o confounding variables

- 200 Placebo subjects, 75% of which have the confounding variable
- 200 Active subjects, 25% of which have the confounding variable
- That is, the confounding variable is associated with treatment
- $\beta_{confounding} = \log(.50)$ so it is associated with Y

model	Variable	N Obs	Variable	Mean
(correct model)				
1	Intercept	200	decision	1.0000000
			Covered	.
			Estimate	-1.5100889
			StdErr	0.2882333
	confound	200	decision	0.6900000
			Covered	0.9600000
			Estimate	-0.6989296
			StdErr	0.2882740
	tx	200	decision	0.9700000
			Covered	0.9550000
			Estimate	1.0951893
			StdErr	0.2948837

model	Variable	N Obs	Variable	Mean
(incorrect model)				
2	Intercept	200	decision	1.0000000
			Covered	.
			Estimate	-1.9916979
			StdErr	0.2187634
	tx	200	decision	1.0000000
			Covered	0.8000000
			Estimate	1.4197298
			StdErr	0.2641194

Summary

Correct (Full) Model
Estimate 1.0951893

Incorrect model
Estimate 1.4197298

Note: The incorrect model is biased. When controlled for the confounding variable, the Log(OR) for TX is unbiased (recall, 1.0986 is the true value)

No confounding, But has effect

```
title "No confound effect: 1:1 confound allocation, OR_TX = 3 (N=200)";
%SIMDATA(200,0,200,-1.504,.50, .50,log(3),log(.50));
```

model	Variable	N Obs	Variable	Mean
(correct model)				
1	Intercept	200	decision	1.0000000
			Covered	.
			Estimate	-1.5297037
			StdErr	0.2282264
	confound	200	decision	0.8400000
			Covered	0.9800000
			Estimate	-0.7095019
			StdErr	0.2507242
	tx	200	decision	1.0000000
			Covered	0.9600000
			Estimate	1.1465769
			StdErr	0.2589997

```
                                  N
       model    Variable        Obs    Variable      Mean
-------------------------------------------------------------
(incorrect model)
           2    Intercept       200    decision    1.0000000
                                       Covered       .
                                       Estimate   -1.8349762
                                       StdErr      0.2063221

                tx              200    decision    1.0000000
                                       Covered     0.9700000
                                       Estimate    1.1209348
                                       StdErr      0.2558027
-------------------------------------------------------------
```

Note that the mean stderrs are not as predicted

```
StdErr  fuller    0.2589997 <-- expected to be smaller
StdErr  reduced   0.2558027
```

However, the coverage probability suggest the fuller model is suggesting smaller standard errors (smaller intervals = less coverage probability)

```
Based on 200 simulation replicates
Covered fuller   0.9600000 <-- smaller as expected
Covered reduced  0.9700000

Based on 1000 simulation replicates
Covered fuller   0.9320000 <-- same trend
Covered reduced  0.9370000
```