

An Illustration of Rate Difference Estimation with SAS in Logistic Regression

Yun Guo, Victor Wu, Xinyu Li, Xinyu Xu and Crystal Cheng

Covance Pharmaceutical Research and Development (Beijing) Co., Ltd.

ABSTRACT

Rate difference is widely used for the binary endpoints in non-inferiority studies. General statistical methods for unstratified and stratified tests are Z test and CMH type tests respectively. They, however, are not suitable for adjusting covariates, which sometimes are essential in clinical trials. Logistic regression along with Delta method could be used when covariate adjustments are needed. We used an example to show how to program in a SAS environment to get the rate difference along with the confidence interval.

INTRODUCTION

Rate difference (RD) is widely used for the binary endpoints in non-inferiority studies. Several statistical methods are available for estimation of RD along with its confidence interval.

As always the case in statistics, the easiest one is Z test that can be found in most elementary statistical textbooks. It's proved to be useful when the sample size is fairly large with non-sparse event rates and no stratification factors are deemed important. Some extensions of Z tests are reviewed in Newcombe (1998) for both independent group studies and matched ones.

When stratification is essential, for example in some studies where the difference in separate gender is to be emphasized, Cochran–Mantel–Haenszel (CMH) type tests (Agresti 2002) can be used instead to adjust the stratification and they are more efficient than Z test. Surely CMH type tests have some good properties, for example it's asymptotic even under very sparse data situation (applications of CMH type tests in such situation could be found in those studies with pre-post design and subjects are treated as stratification). It however, has limited use when the 'stratification' variable is continuous, i.e. covariates are included (thorough out the paper, we use terminologies from experimental design where 'factor' is a categorical predictor and 'covariate' refers to a continuous predictor). Some may argue that by categorization of covariates, we could still apply CMH type tests. This is true when certain natural cut-offs of covariate are available. However it should be noted that fitting CMH type test with categorization of covariates is in the expense of potential information loss and thus may induce a wider confidence interval. In addition, in some early phase studies, the investigators are more interested in the parametric relationships between covariates and binary response. In such case, a CMH test with covariates categorized could only provide limited information for further reference in the later studies since it poses no specific parametric form on the categorized covariate.

Logistic regression (Agresti 2002) is a good alternative and can still preserve the large sample property of CHM type statistics. As a model based strategy, the logistic regression is capable of displaying the relationship between predictors and response. It however, needs further understanding on the regression function and sometimes additional programming efforts are required. In the following sections, we will give a brief description of logistic regression and Delta method, which could be used for estimation of confidence interval. To make things clear and easy, example as well as SAS codes are presented in a step by step manner.

LOGISTIC REGRESSION

For binary (success/failure) response randomly sampled from independent subjects, logistic regression model is a natural consideration. Consider the following case from a hypothetical clinical trial: in a study, we have one stratification factor s (readers could think of examples like site, gender, HBV e-antigen existence or not, etc.), a covariate x (for example age, baseline systolic blood pressure, etc.) and a treatment indicator t (drug A, placebo, different doses, etc.), the logistic regression model assumes that (for the sake of simplicity, we don't use matrix notation as in most textbooks and write the predictors out explicitly.)

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i + \gamma s_i + \tau t_i, \quad i = 1, 2, \dots, n \quad (1)$$

, where $\alpha, \beta, \gamma, \tau$ are the model coefficients attached to different predictors (to be estimated of course) and subscript i indexes individual subject and n is the total sample size for the study.

Note that predictors enter into logistic model linearly on the Logit scale, i.e., linear in the link function $\log\left(\frac{p_i}{1-p_i}\right)$. This nice property makes estimation of Odds Ratio (OR) and its confidence interval rather straightforward. To get the OR for a particular predictor, simply by subtracting among different levels of the interested predictor, the effects of other predictors are cancelled out. However, it's not the case for RD estimate as shown next.

Let us assume that we are facing the model (1) with treatment encoded 1 for the test and 0 for the control in a trial. A little algebra can bring us the rates for both groups for 'certain' pre-specified level of x and s (we will mention later in this section how to 'find' the 'certain' level for the covariates as well as factors):

$$\hat{p}_{\tau=1} = \frac{1}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x + \hat{\gamma}s + \hat{\tau})]} \quad (2)$$

for the test and

$$\hat{p}_{\tau=0} = \frac{1}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x + \hat{\gamma}s)]} \quad (3)$$

for the control. Here the hat ($\hat{\cdot}$) sign means estimates.

Clearly a simple subtracting can't get rid of the x , s or even the intercept α (as logistic regression is linear on the logit form, not on the 'rate' form). This is not critical if only the point estimate of RD is our focus (which is rare in clinical trial) since we could always get it by subtracting (2) by (3). Rather, the difficulty is imposed on the variance estimation, i.e. $\text{Var}(\hat{p}_{\tau=1} - \hat{p}_{\tau=0})$, which involves intercept α , covariates x as well as factors s with non-linear forms, not to mention the complex estimated covariance matrix (not shown here but could be found in our example).

DELTA METHOD

To get the confidence intervals, a good solution is the Delta method (Agresti 2002) which is a linear approximation technique. Loosely speaking (for strict definitions, please refer to Agresti 2002), it states that if a random variable $Y \xrightarrow{D} N(\theta, \sigma^2)$ (theoretically should be the limiting distribution of a random variable but we purposefully ignore

this here for simplicity) then the functional form $g(Y) \xrightarrow{D} N(g(\theta), (\frac{\partial g(\theta)}{\partial \theta})^2 \sigma^2)$, where $\xrightarrow{D}$ denotes converge in

distribution, N denotes a normal distribution with two parameters in the brackets and $\frac{\partial g(\theta)}{\partial \theta}$ denotes the first

derivate of functional form $g(\theta)$ with respect to θ . Here we give an example. If a random variable Y follows $N(2, 5^2)$, then Y^2 converges to a normal distribution with mean $4 = (2^2)$ and variance $400 = (2 \cdot 2)^2 \cdot 25$.

Note that in the above definition we are using Delta method in a univariate setting. The multivariate counterpart is in the same logic. The only difference is that multivariate normal distribution is involved instead of one dimensional normal distribution and first derivative is with respect to a vector instead of a real value. For example in our model (1), the distribution of the estimated coefficients is assumed to converge in multivariate normal distribution (under some regularity conditions and readers could consult any books on generalized linear regression) with estimated mean

vector $(\hat{\alpha} \ \hat{\beta} \ \hat{\gamma} \ \hat{\tau})$ and a 4 by 4 estimated covariance matrix $\hat{V}$.

The following two derivatives with respect to estimated coefficient vector are very useful in logistic regression:

$$\frac{\partial \hat{p}_{\tau=1}}{\partial (\hat{\alpha} \ \hat{\beta} \ \hat{\gamma} \ \hat{\tau})} = \hat{p}_{\tau=1} (1 - \hat{p}_{\tau=1}) (1 \ x \ s \ 1) \quad (4)$$

, and

$$\frac{\partial \hat{p}_{\tau=0}}{\partial (\hat{\alpha} \ \hat{\beta} \ \hat{\gamma} \ \hat{\tau})} = \hat{p}_{\tau=0} (1 - \hat{p}_{\tau=0}) (1 \ x \ s \ 0) \quad (5)$$

, where $(1 \ x \ s \ 1)$ and $(1 \ x \ s \ 0)$ are the 1 by 4 row vectors for predictors (intercept in the first column and

treatment in the last column). We postpone the variance estimation of $\hat{RD}$ until the end of this section.

As we mentioned earlier, how to 'fix' the 'certain' level of covariates and factors should also be clear before we could get $\hat{RD}$. We take the marginal approach termed in Wilcosky and Chambless (1985) as this seems very plausible in clinical trial. Interested readers could consult their paper for more details. The marginal approach states that:

$$\hat{p}_{\tau=1} = \sum_i \frac{1/n}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x_i + \hat{\gamma}s_i + \hat{\tau})]} \quad (6)$$

for the test and

$$\hat{p}_{\tau=0} = \sum_i \frac{1/n}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x_i + \hat{\gamma}s_i)]} \quad (7)$$

for the control.

Note summation over i is for all the subjects included in the analysis, regardless of which treatment group they belong to (This may seem a little ridiculous for some readers at the first glimpse. Think of the word 'marginal' in probability world, it means integrate out other random variables to get the marginal probability density for the targeted random variables. This is what Wilcosky and Chambless were doing with their marginal approach and is very plausible as the covariate/factor other than treatment is adjusted to the same level in the sense of 'adjustment'). This implies that in programming, we have to first treat all the subjects as from test group and then treat them as from control group, no matter what real treatment they were assigned to. As could be seen in our programming example below, we duplicated two data sets in the dataset preparation step, exactly realizing the definition of marginal approach.

Now by subtracting (6) by (7) we get $\hat{RD}$

$$\hat{RD} = \hat{p}_{\tau=1} - \hat{p}_{\tau=0} \quad (8)$$

, and by Delta method with the following step by step arguments we get the variance of $\hat{RD}$:

$Var(\hat{RD}) =$

$$\begin{aligned} & Var\left(\sum_i \frac{1/n}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x_i + \hat{\gamma}s_i + \hat{\tau})]} - \sum_i \frac{1/n}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x_i + \hat{\gamma}s_i)]}\right) = \\ & \frac{1}{n^2} \sum_i Var\left(\frac{1}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x_i + \hat{\gamma}s_i + \hat{\tau})]} - \frac{1}{1 + \exp[-(\hat{\alpha} + \hat{\beta}x_i + \hat{\gamma}s_i)]}\right) = \\ & \frac{1}{n^2} \sum_i [p_{1i}(1-p_{1i})A_i'V_{A_i}p_{1i}(1-p_{1i}) + p_{0i}(1-p_{0i})B_i'V_{B_i}p_{0i}(1-p_{0i}) - 2p_{1i}(1-p_{1i})A_i'V_{B_i}p_{0i}(1-p_{0i})] \end{aligned} \quad (9)$$

, where A_i and B_i are predictor row vectors (including in the first column a '1' for intercept) with element of treatment equal 1 or 0 respectively for subject i ; the subscript 1 or 0 indexes treatment group.

Define $d_{r1} = \sum_i \hat{p}_{1i}(1-\hat{p}_{1i})A_i/n$ and $der_0 = \sum_i \hat{p}_{0i}(1-\hat{p}_{0i})B_i/n$, we have

$$Var(\hat{RD}) = der_1' \hat{V}_{der_1} der_1 + der_0' \hat{V}_{der_0} der_0 - 2der_1' \hat{V}_{der_0} der_0 \quad (10)$$

, where $-2der_1' \hat{V}_{der_0} der_0$ is covariance of $\hat{p}_{\tau=1}$ and $\hat{p}_{\tau=0}$.

Practically, programming could either be based on (9) or (10), but the latter seems easier if 'proc iml' (SAS Institute Inc. 2010) could be implemented from our own experience.

EXAMPLE AND SAS CODES

The example is taken from SAS example in 'proc logistic' (SAS Institute Inc. 2010). We have some modifications before we proceed to estimate the rate difference in order to make the example to represent model (1): only treatment='A' or 'P' is included and variable duration is not included in further analysis. The response is 'pain' with two levels; the stratification factor is 'gender' with 0 for female and 1 for male; 'age' performs the role as a covariate which is continuous; treatment is encoded as '1' for 'A' and '0' for the control 'P'. For the complete description of the dataset, please refer to the SAS documentation. Some comments are attached in the program to help understand programming logic.

DATASET PREPARATION

For dataset 'temp' we replicate the observations twice, one for treatment='1' and another for treatment '0'. This is in

correspondence with marginal approach. The responses in the replicated parts are called missing (this is very important and saves a lot of programming) in order to exclude the replicated observations from model fitting in 'proc logistic'.

Data Neuralgia;

```
input Treatment $ Sex $ Age Duration Pain $ @@;
  if sex='F' then gender=0;
  else gender=1;
  drop sex;
  if treatment in ('A','P');
  if treatment='A' then treat=1;
  else treat=0;
  drop treatment;
  rename treat=treatment;
```

datalines;

```
P F 68 1 No B M 74 16 No P F 67 30 No
P M 66 26 Yes B F 67 28 No B F 77 16 No
A F 71 12 No B F 72 50 No B F 76 9 Yes
A M 71 17 Yes A F 63 27 No A F 69 18 Yes
B F 66 12 No A M 62 42 No P F 64 1 Yes
A F 64 17 No P M 74 4 No A F 72 25 No
P M 70 1 Yes B M 66 19 No B M 59 29 No
A F 64 30 No A M 70 28 No A M 69 1 No
B F 78 1 No P M 83 1 Yes B F 69 42 No
B M 75 30 Yes P M 77 29 Yes P F 79 20 Yes
A M 70 12 No A F 69 12 No B F 65 14 No
B M 70 1 No B M 67 23 No A M 76 25 Yes
P M 78 12 Yes B M 77 1 Yes B F 69 24 No
P M 66 4 Yes P F 65 29 No P M 60 26 Yes
A M 78 15 Yes B M 75 21 Yes A F 67 11 No
P F 72 27 No P F 70 13 Yes A M 75 6 Yes
B F 65 7 No P F 68 27 Yes P M 68 11 Yes
P M 67 17 Yes B M 70 22 No A M 65 15 No
P F 67 1 Yes A M 67 10 No P F 72 11 Yes
A F 74 1 No B M 80 21 Yes A F 69 3 No
```

```
;
```

run;

***creation of 'temp' dataset to get replicates;

data temp;

```
set Neuralgia Neuralgia (in=a) Neuralgia (in=b);
if a or b then call missing(pain);
```

```
if a then do;***replicates for the test treatment;
```

```
  treat=1;***treat is an indicator of p1 or p0;
```

```
  treatment=1;
```

```
end;
```

```
if b then do;*** replicates for the control treatment;
```

```
  treat=0;
```

```
  treatment=0;
```

```
end;
```

run;

INVOKE LOGISTIC REGRESSION

The following codes are for the estimation of predicted rates (outputted in the dataset 'pred_p') and covariance matrix (outputted in the dataset 'est' as shown in figure 1 below) through the 'proc logistic'. The response variable as well as predictors is reflected in the 'model' statement and a 'class' is used for the categorical factors. Attention should be paid on the 'class' statement in which the 'param=ref' option is generally needed as the default method to get dummy variable is 'effect coding' in 'proc logistic'. Note that here we are using 'proc logistic' to get the estimates. Other procedures can also be used, for example 'proc genmod'.

```
proc logistic data=temp covout outest=est(where=(_type_='PARMS'));
  class treatment (ref='0')/param=ref;
  model pain (event='Yes')=gender age treatment;
  output out=pred_p p=p;
run;
```

	LINK	TYPE	STATUS	NAME	Intercept	gender	Age	treatment1
1	LOGIT	COV	0 Converged	Intercept	47.888383	-0.592253	-0.69894	1.882514
2	LOGIT	COV	0 Converged	gender	-0.5922527	0.828839	0.0051708	-0.37576
3	LOGIT	COV	0 Converged	Age	-0.6989399	0.0051708	0.010286	-0.03094
4	LOGIT	COV	0 Converged	treatment1	1.88251358	-0.375763	-0.030935	0.92005

Figure 1. Estimated covariance matrix from logistic regression (dataset=est)

GET $\bar{RD}$

The following codes are used for $\bar{RD}$. As previously shown in (7), (8) and (9), the rate differences are merely weighted average of individual rate difference according to the marginal approach. The 'proc ttest' thus can be used (we are not running t test for significance tests!!!). The results are in Figure 2. It shows that $\bar{RD} = -0.483$.

```
***Note: no t test intended here, just for difference of rates;
proc ttest data=pred_p;
  class treat;
  var p;
  where treat^=.;
run;
```

The SAS System										
22:46 Wednesday, March 31, 2012										
The TTEST Procedure										
Statistics										
Variable	treat	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	
p	0	40	0.6832	0.7441	0.805	0.156	0.1904	0.2445	0.0301	
p	1	40	0.1884	0.2611	0.3338	0.1862	0.2273	0.2918	0.0359	
p	Diff (1-2)		0.3897	0.483	0.5764	0.1813	0.2097	0.2486	0.0469	

Figure 2. 'proc ttest' output

GET VARIANCE ESTIMATE FOR $\bar{RD}$

The datasets 'der1' and 'der0' are for the test and control respectively. The variable 'der_sub' stores the derivatives that are used in the Delta method. Since 'intercept' acts as a 'covariate', we assign '1' to it. To get the variance estimator, 'proc iml' is used. The reader not familiar with this procedure may consult the relevant SAS documentation. In 'proc iml', the 'use' and 'read' statements load the derivatives and estimated covariance matrix. The Delta method is implemented in the assign statement 'var=...'. Finally, we have the estimated variance saved in the dataset 'var'

and the confidence interval can be constructed by the usual way of $\bar{RD} \pm Z_{\alpha/2} \sqrt{Var(\bar{RD})}$ with a suitable pre-specified alpha level.

```
data der1 der0;
  set pred_p;
  der_sub=p*(1-p);***derivative for individual;
  intercept=1;
  if treat=1 then output der1;
  if treat=0 then output der0;
run;

proc iml;
  use der1;
  read all var{der_sub} into der1;
  read all var{intercept gender age treatment} into A;
```

```

    ***note the order of variables should be the same as in the model statement in proc logistic;
    use der0;
    read all var{der_sub} into der0;
    read all var{intercept gender age treatment} into B;
    close der1 der0;

    use est;
    read all var{intercept gender age treatment1} into cov;
    close est;
    ***application of Delta method;
    der_1=t(der1);
    der_0=t(der0);
    n=nrow(der1);
    var=j(1,1,0);
    ***apply formula (10) here;
    var=der_1*A*cov*t(der_1*A)+der_0*B*cov*t(der_0*B)-2*der_1*A*cov*t(der_0*B);
    var=var/n/n;
    create var from var;***as could seen in dataset=var, estimated variance=0.0148575375;
    append from var;
    quit;

```

CONCLUSION

In this paper, we illustrated the estimation of RD and its confidence interval by logistic regression and Delta method. A small amount of SAS codes are provided along with a step by step explanation.

Ge et al (2011) had their macro published and due to some reasons, the SAS codes they provided may not be executable. We don't mean to copy their codes in this paper. Rather, we have it clearly explained in a manner that statistical programmers could understand. Since it's the basic idea that matters when programming, our codes are not written in a standard Macro to complex the situation. It could easily be extended to include more factors/covariates and thus is very practical for the SAS users. For the sake of simplicity, we only include one stratification factor 'gender' with two levels and a covariate 'age'. As we commented in our codes, the readers should be aware that since gender is either 0 (male) or 1 (female), no dummy variables were created. This is not always true. For example, in some multi-center trials where site is a stratification factor, the dummy variables are needed. If only a small number of levels exist, some easy if-then statements (or other similar statements depends on the practice of the programmer) suffice. When the number of levels is big, a 'proc transreg' which is optimally used for the 'spline' or other transformation regression may help. Following codes are an illustration.

```

proc transreg data=xx design;
  model class(a b /zero='1' '1');
  output out=xxxx;
run;

```

For estimation of confidence interval, we employed a Delta method. Note that Delta method is not the only method to get the confidence interval. It has limitations too, i.e. only valid within large sample scenarios. Other methods are also available. For example Monte Carlo simulation based on the estimated coefficients vector and covariance matrix, can be used if computing time is not a concern. Interested readers could consult any books on simulation. The question of which method to use is merely a subject matter. Since it's not the routine output from SAS, our recommendation is to use one method as the main analysis method and perhaps the other one as a validation method.

REFERENCE

1. Agresti A. 2002. Categorical Data Analysis, 2nd edition. New York: Wiley.
2. Ge M, Durham LK., Meyer RD., Xie W. and Thomas N. 2011. Covariate-adjusted Difference in Proportions from Clinical Trials Using Logistic Regression and Weighted Risk Differences. Drug information J 4:481-493.
3. Newcombe RG. 1998. Interval Estimation for the Difference Between Independent Proportions: Comparison of Eleven Methods. Statist. Med 17: 873-890.
4. Newcombe RG. 1998. Improved Confidence Intervals for the Difference Between Binomial Proportions Based on Paired Data. Statist. Med 17: 2635-2650.

5. SAS Institute Inc. 2010. SAS/IML® 9.22 User's Guide. Cary, NC: SAS Institute Inc.

6. SAS Institute Inc. 2010. SAS/STAT® 9.22 User's Guide. Cary, NC: SAS Institute Inc.

7. Wilcosky TC, Chambliss LE. 1985. A Comparison of Direct Adjustment and Regression Adjustment of Epidemiologic Measures. J Chron Dis 38: 849-856.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Derek Li, Yun Guo, Victor Wu
Enterprise: Covance Pharmaceutical Research and Development (Beijing) Co., Ltd.
Address: Units1501-1508, Tower A, Gemdale Plaza, No. 91, JianGuo Road
City, State ZIP: Beijing, China 100022
Fax: (86 -10) 5732 3001/3002
E-mail: Yun.guo@covance.com
Victor.wu@covance.com
Derek.li@covance.com
Xinyu.xu@covance.com
Crystal.Cheng@covance.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.