Proper Estimation of Relative Risk Using PROC GENMOD in Population Studies

Kechen Zhao, University of Southern California, Los Angeles, California

ABSTRACT

Relative risk (RR) is usually the parameter of interest in cohort studies. Binary outcomes in cohort studies are commonly analyzed by applying a logistic model to obtain odds ratio (OR) for comparing risks among groups with different sets of characteristics. However, OR always overestimates RR, sometimes dramatically for more common outcomes. Nevertheless, recent medical literature has frequently included uncritical applications of logistic regression to cohort studies. The purpose of this paper is to demonstrate the correct application of a *modified* Poisson regression method to directly estimate relative risk from a cohort data set, which has quickly gain popularity in medical and public health research. Simulated population data is used to illustrate statistical methods with PROC GENMOD in SAS® 9.3. For the purpose of method comparison, OR estimation with a logistic regression, which is less desirable for assessment of risk in a cohort study with more common outcomes, will also be demonstrated here. In addition, SAS® ODS and Macro facilities relating to modeling building and reporting results will also be introduced in this paper.

INTRODUCTION

Epidemiological and clinical research is largely grounded on the assessment of risk. Typically, when the outcome variable of interest is dichotomous, a popular tool in accessing the risk of exposure or the protective effect from a treatment is a logistic regression model, which directly yields an estimated odds ratio. Although logistic regression may be correctly applied to case-control studies, in cohort studies we are often interested in estimating a relative risk (or, rate ratio), not the odds ratio. In studies of common outcomes, the estimated odds ratio can substantially overestimate the relative risk. Regardless the difference between an odds ratio and a relative risk, authors and consumers of medical report often interpret the odds ratio as a relative risk, leading to its potential exaggeration. Extensive discussion in much of the literature has reached a consensus that relative risk is always preferred over the odds ratio for most cohort studies. Nevertheless, recent medical literature has frequently included uncritical applications of logistic regression to cohort studies.

To estimate the relative risk directly, binomial regression and Poisson regression are usually recommended. However, as is commonly known, neither works very well. Binomial regression models may suffer convergence problems and fail to provide a valid estimate of relative risk. On the other hand, although ordinary Poisson regression models can provide a valid point estimate of relative risk, they tend to provide a wider confidence interval on a relative risk, leading to conservative results. The purpose of this paper is to demonstrate how to estimate relative risk by using a *modified* Poisson regression model with a robust error variance (or so-called sandwich error variance).

THE SIMULATED POPULATION DATA

A simulated (or hypothetical) data set was created to illustrate two methods of estimating relative risk using SAS®. The outcome generated is called disease, to indicate if the simulated study participants develop a hypothetical infectious disease within a span of one year. Assume all participants are disease-free at the starting point of the study. Suppose we want to know if developing such disease is associated with a gene which regulates the activity of immune response, and that we screened everyone for the risk allele of this gene at the entry of the study (gene = 0 if they carry none, = 1 if they carry one copy, = 2 if they carry two copies). We also screened everyone for smoking status (=1 if smoking, =0 if never), we note their genders (=1 if male, =0 if female). All values (N=500000) were assigned using a random number generator in SAS®. The following SAS® code generates six independent cohort data sets with an incidence rate fixed at 1%, 5%, 10%, 20%, 50% and 80% respectively:

```
libname wuss13 'C:\Users\User\Desktop\Kechen\wuss2013';

%let N= 500000; /* Number of observations in the cohort */
%let MAF1 = 0.15; /* Population frequency of the risk allele*/
%let intercept1 = -6.4; /* model intercept for generating 1% incidence
rate */
%let intercept2 = -4.465; /* model intercept for generating 5% incidence
rate */
```

```
%let intercept3 = -3.52; /* model intercept for generating 10% incidence
rate */
%let intercept4 = -2.53; /* model intercept for generating 20% incidence
rate */
%let intercept5 = -0.884; /* model intercept for generating 50% incidence
rate */
%let intercept6 = 0.66; /* model intercept for generating 80% incidence
rate */

%MACRO gendata;
%LOCAL j;
%DO j = 1 %TO 6;
data wuss13.cohort&j;
array freq1{&N} _temporary_;
array S{&N} _temporary_;
array G1{&N} _temporary_;
array E1{&N} _temporary_;
array A{&N} _temporary_;
call streaminit(1); /* Set seeds */

do i = 1 to &N;
  /*Simulate Gene variable based on Hardy-Weinberg equilibrium model */
  freq1{i} = rand("Uniform");
  if (freq1{i} le &MAF1*&MAF1) then G1{i} = 2;
  else if (freq1{i} gt (&MAF1*&MAF1) AND freq1{i} le
(2*&MAF1*&MAF1+&MAF1*&MAF1)) then G1{i} = 1;
  else if (freq1{i} gt (2*&MAF1*&MAF1+&MAF1*&MAF1) AND freq1{i} le 1) then
G1{i} = 0;

  /*Simulate Smoking variable based on age distribution*/
  A{i} = rand("Weibull", 2, 40);
  if A{i} lt 15 then E1{i} = rand("Bernoulli", 0.05);
  else if A{i} ge 15 then E1{i} = rand("Bernoulli", 0.35);

  /*Simulate Gender variabl*/
  S{i} = rand("Bernoulli", 0.57);
  end;

/*Simulate the binary outcome variable based on the logistic regression
model*/
do i = 1 to &N;
  gene = G1{i};
  smoking = E1{i};
  gender = S{i};

  id + 1;
  /* Specify regression coefficients */
  eta = &&intercept&j + 2.1*gene + 1.21*smoking + 0.69*gender;
  mu = exp(eta) / (1 + exp(eta));
  disease = rand("Bernoulli", mu);
  output;
  end;
run;

%END;
%MEND;
```

```
/*Invoke the Macro to simulate the data set*/
%gendata;
```

Table 1: Summary of predictor variables and outcome variables

Variable Name	Description	Categories
Disease	Outcome variable	0 = not having the disease 1 = having the disease
Gene	Covariate 1	0 = having zero copy of "bad" version 1 = having one copy 2 = having two copies
Smoking	Covariate 2	0 = Never smoking 1 = Smoking
Gender	Covariate 3	0 = female 1 = male

Table 1 presents the summary of the 3 predictors and the outcome variable. We then label these variables. There are two main items that can be labeled, variables and values. These labels will appear in the output of statistical procedures and reports that you may produce from SAS®. They are also displayed by some of the SAS/GRAPH procedures. For simplicity and instructive purposes, we only demonstrate how to assign labels to the variable '**smoking**'. Same labeling process is applied to all other predictors and outcome variable in a single PROC FORMAT step followed with a single DATA step.

First we create the label format with PROC FORMAT using a value statement. The program below creates the **smoking** variable format, **smokingf**.

```
Proc format;
VALUE smoking
0 = 'never'
1 = 'smoking';
run;
```

Now the format **smokingf** has been created, they must be linked to the variable **smoking**. This is accomplished by including a FORMAT statement in a DATA step. In the program below the format statement is used in a DATA step. The following program also assign label to **smoking**.

```
Data wuss13.cohort3;
set wuss13.cohort3;
label smoking = "Smoking status";
format smoking smokingf.;
run;
```

MODELING PROCEDURE: STUDYING ASSOCIATION VERSUS PREDICTION

Statistician George E. P. Box once famously wrote that "essentially, all models are wrong, but some are useful" in his book on response surface methodology with Norman R. Draper. Rarely is there only one best statistical model that adequately fits a set of data. Instead, researchers usually choose a few models that well summarize the information about the data. The choice between models that adequately fit the data is based on various criteria, one of which is the research question. If our focus is to measure associations between predictors and the outcome of interest, computing relative risks is one way to measure such associations. Unlike predictive models where parsimony is desired, regression models for studying associations often keep several factors that may not explain much of the variance in the outcome. Yet, these variables are included in the model to control for confounding effect. Other selection criteria may include the existence of influential observation and other factors related to model fit.

THE SIMPLE REGRESSION MODEL

We begin our analysis by building simple regression models with SAS®. In this type of regression, we only include one predictor and one outcome variable in the model. There are two main reasons we build simple regression models here. First, we want to access the direction and rough size of relationships between predictors and outcome variable. Second, we want to calculate an incidence rate of an outcome within a particular group (sub-cohort) from a *modified* Poisson regression model. Here, we will build a *modified* Poisson regression model and compare it to three other different models using a same pair of predictor and outcome:

1. Logistic regression model
2. Log-binomial model
3. Poisson regression model

For simplicity and instructive purposes, we only demonstrate simple model building process with just smoking variable **smoking** as the predictor and **disease** as the outcome variable. Same simple model building process is applied to other predictors and outcome using an iterative method with SAS® Macro, as will be shown later. The simple regression analysis is based on the simulated population data with a disease prevalence rate fixed at 10% (wuss13.cohort3).

LOGISTIC REGRESSION MODEL

We could use either PROC LOGISTIC or PROC GENMOD to calculate the odds ratio (OR) with a logistic regression model. Since PROC LOGISTIC will provide OR estimates directly in the output, it will be used to calculate the OR (and it gives the same results as PROC GENMOD). Here is the logistic regression with just smoking variable **smoking** as the predictor and **disease** as the outcome variable:

```
Proc logistic data=wuss13.cohort3;
class smoking (ref="never" param=ref);
model disease (event = "1") = smoking;
run;
```

Output 1: Odds ratio estimate with a logistic regression model

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.5898	0.00668	150083.669	<.0001
smoking	**smoking**	1	0.9922	0.00952	10857.0125	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
smoking smoking vs never	2.697	2.647	2.748

From **Output 1**, the odds developing disease is 2.697 higher in those who smoke as compared to those never smoke (95%CI: 2.65-2.78), and was statistically significant ($p < 0.0001$).

LOG-BINOMIAL REGRESSION MODEL

With a very few modifications of the statements used above for the logistic regression, a log-binomial model can be run with PROC GENMOD to get relative risk instead of the odds ratio. The PROC GENMOD statement with the DESCENDING option causes the levels of the response variable to be sorted from highest to lowest instead of lowest to highest. Invoking the DESCENDING option causes the model to refer to a constructed binary variable Y that equals 1 when a sample is a case and 0 otherwise. The CLASS statement with the REF option specifies the reference level you desire. The MODEL statement with the DIST option specifies the probability distribution of the data and the LINK option specifies a user-defined link function. In a log-binomial model, we assign binomial distribution as the probability distribution and logarithm as the linking function. The EXP option on the ESTIMATE statement gives us the estimated relative risk for those who smoke versus those who never smoke.

```
proc genmod data = wuss13.cohort3 DESCENDING;
class smoking (ref="never" param=ref);
model disease = smoking/ dist = binomial link = log;
estimate 'beta' smoking 1 /exp;
run;
```

Output 2: Relative risk estimate with a log-binomial regression model

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.6622	0.0062	-2.6743	-2.6500	183279	<.0001
smoking	smoking	1	0.8802	0.0084	0.8638	0.8967	10993.8	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Contrast Estimate Results										
Label	**Mean Estimate**	**Mean Confidence Limits**		**L'Beta Estimate**	**Standard Error**	**Alpha**	**L'Beta Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
beta	2.4114	2.3721	2.4515	0.8802	0.0084	0.05	0.8638	0.8967	10994	<.0001
Exp(beta)				2.4114	0.0202	0.05	2.3721	2.4515		

From **Output 2**, the risk developing disease is 2.41 higher in those who smoke as compared to those never smoke (95%CI: 2.3721-2.4515), and was statistically significant ($p < 0.0001$).

Although a simple log-binomial regression model usually converges, a multiple log-binomial regression model rarely converges. To see this, the following program builds a log-binomial model by including just 2 predictors:

```
proc genmod data = wuss13.cohort3 DESCENDING;
class smoking (ref="never"  param=ref);
model disease = gene smoking/ dist = binomial link = log;
run;
```

Log 1: Convergence information with a multiple log-binomial regression model

```
4040  ods rtf close;
4041  proc genmod data = wuss13.cohort3 DESCENDING;
4042  class smoking (ref="never"  param=ref);
4043  model disease = gene smoking/ dist = binomial link = log;
4044  run;

NOTE: PROC GENMOD is modeling the probability that disease='1'.
WARNING: The specified model did not converge.
```

Log 1 shows that the multiple log-binomial model fails to converge and thus the model fails to provide valid relative risk estimate. This is why Poisson regression and *modified* Poisson regression approaches are presented here.

POISSON REGRESSION MODEL AND MODIFIED POISSON REGRESSION MODEL

Although ordinary Poisson regression models can provide a valid point estimate of relative risk, they tend to provide a wider confidence interval on a relative risk due to the misspecification of the outcome distribution, leading to conservative results. Researchers suggest using a *modified* Poisson approach to estimate the relative risk and confidence intervals by using robust error variances. To see how this works, we will provide a Poisson regression model and compare it to a *modified* Poisson regression model.

With a very few modifications of the statements used above for the log-binomial regression, here is how it is done to build a Poisson regression model in SAS®:

```
proc genmod data = wuss13.cohort3 DESCENDING;
class smoking (ref="never"  param=ref);
model disease = smoking/ dist = poisson link = log;
estimate 'beta' smoking 1 /exp;
run;
```

Output 3: Relative risk estimate with a Poisson regression model

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.6622	0.0064	-2.6748	-2.6495	170486	<.0001
smoking	smoking	1	0.8802	0.0089	0.8627	0.8977	9707.44	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Contrast Estimate Results										
		Mean					L'Beta			
Label	Mean Estimate	Confidence Limits		L'Beta Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
beta	2.4114	2.3696	2.4540	0.8802	0.0089	0.05	0.8627	0.8977	9707.4	<.0001
Exp(beta)				2.4114	0.0215	0.05	2.3696	2.4540		

From **Output 3**, the risk developing disease is 2.41 higher in those who smoke as compared to those never smoke (95%CI: 2.3696-2.4540), and was statistically significant ($p < 0.0001$).

The robust error variance can be estimated by using the REPEATED statement and subject identifier (here **id**), even if there is only one observation per subject. Notice that the TYPE = UNSTR option tells PROC GENMOD to use an unstructured correlation matrix. , here is how it is done to build a *modified* Poisson regression model in SAS®:

```
proc genmod data = wuss13.cohort3 DESCENDING;
class id smoking (ref="never"  param=ref);
model disease = smoking/ dist = poisson link = log;
repeated subject = id / type = unstr;
estimate 'beta' smoking 1 /exp;
run;
```

Output 4: Relative risk estimate with a *modified* Poisson regression model

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-2.6622	0.0062	-2.6743	-2.6500	-428.11	<.0001
smoking	smoking	0.8802	0.0084	0.8638	0.8967	104.85	<.0001

Contrast Estimate Results										
		Mean					L'Beta			
Label	Mean Estimate	Confidence Limits		L'Beta Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
beta	2.4114	2.3721	2.4515	0.8802	0.0084	0.05	0.8638	0.8967	10994	<.0001
Exp(beta)				2.4114	0.0202	0.05	2.3721	2.4515		

From **Output 4**, the risk developing disease is 2.41 higher in those who smoke as compared to those never smoke (95%CI: 2.3721-2.4515), and was statistically significant ($p < 0.0001$).

CALCULATING THE INCIDENCE RATE FROM A MODIFIED POISSON REGRESSION MODEL

A Poisson model can be written as,

log(Expected incidences) = log(Total person-year) + intercept + $\boldsymbol{x}^t\boldsymbol{\beta}$,

where $\boldsymbol{x}$ represents the vector of predictors and $\boldsymbol{\beta}$ represents the corresponding vector of coefficients. Since each individual was followed exactly one year in our study, log(Total person-year) equals log(1), which is zero. The above model can be reduced to

log(Expected incidences) = Intercept + $\boldsymbol{x}^t\boldsymbol{\beta}$,

which can be further transformed into,

Expected incidences = exp(Intercept + $\boldsymbol{x}^t\boldsymbol{\beta}$).

A Poisson model with **smoking** as the predictor and **disease** as the outcome can be written as,

Expected incidences of disease = exp(Intercept + β*smoking),

where "**smoking** = 1" if smoking and "**smoking** = 0" if never smoking. Therefore, the number of disease cases for those who never smoke can be calculated using the following equation:

number of cases who never smoke = exp(Intercept + β*0) = exp(Intercept).

Similarly, the number of disease cases who smoke can be calculated using the following equation:

number of cases who smoke = exp(Intercept + β*1)
= exp(Intercept)*exp(β).

We can obtain incidence rates automatically in SAS® using the following program:

```
/* Output result table into "myout" data set */
proc genmod data=wuss13.cohort3;
class id smoking (ref="never"  param=ref);
model disease = smoking / dist   = poisson
                          link   = log;
repeated subject = id / type = unstr;
```

```
ods output GEEEmpPEst = myout;
run;

/* Exponentiate parameter estimates */
data myout_exp;
retain Parm Level1 Estimate expest;
set myout;
expest = exp(Estimate);
run;

/* Keep the exponentiated Intercept estimate as the baseline incidence
rate */
data _null_;
set myout_exp;
if Parm = "Intercept";
call symputx('baserate', expest);
run;

/* Calculate group-specific incidence rate from the baseline rate and the
relative risk */
data myout_IR (drop=expest);
retain Parm Level1 Estimate expest IncidenceRatePer1000 RelativeRisk;
set myout_exp;
if Parm = "Intercept" then RelativeRisk = .;
else RelativeRisk = expest;
if Parm = "Intercept" then IncidenceRatePer1000 = &baserate * 1000;
else IncidenceRatePer1000 = &baserate * expest * 1000;
run;
```

Output 5: Calculating incidence rate with a *modified* Poisson regression model

Obs	Parm	Level1	Estimate	IncidenceRate Per1000	Relative Risk	Stderr	LowerCL	UpperCL	Z	ProbZ
1	Intercept		-2.6622	69.798	.	0.0062	-2.6743	-2.6500	-428.11	<.0001
2	smoking	smoking	0.8802	168.314	2.41145	0.0084	0.8638	0.8967	104.85	<.0001

Output 5 shows that the incidence rate for disease is about 70 per 1000 in those who never smoke and about 168 per 1000 in those who smoke. The relative risk (or incidence rate ratio) for developing disease is 2.41 higher in those who smoke as compared to those never smoke (95%CI: 2.3721-2.4515), and was statistically significant ($p < 0.0001$).

The incidence rates and rate ratio can be easily verified by traditional a two-by-two table:

```
proc freq data = wuss13.cohort3;
table disease*smoking /  nocum nopercent nocol norow missing;
run;
```

Output 6: Two by two table to calculate incidence rate

Table of disease by smoking			
disease	smoking(Smoking status)		
Frequency	never	smoking	Total
0	320597	129200	449797
1	24056	26147	50203
Total	344653	155347	500000

Incidence Rate per 1000 (smoking) = 24056/344653*1000 = 69.79774
Incidence Rate per 1000 (never smoking) = 26147/155347*1000 = 168.3135
Relative Risk (smoking Vs. never smoking) = (26147/155347)/(24056/344653) = 2.411447

The incidence rates and relative risk we obtained from the modified Poisson regression is consistent with that from the two-by-two table.

BUILDING SIMPLE REGRESSION MODELS ITERATIVELY USING SAS® MACRO

You can build a simple regression model and store the parameter estimates for all selected predictors in a single step. The following SAS® Macro, %uniReg(), can execute model building and storing results repetitively for each of 3 predictors. **Figure 1** and **Figure 2** shows list of tables generated by %uniReg() and an example from those tables.

```
/*
   %uniReg() takes three arguements
   "varlist" takes all the predictors to be included
   "reflist" takes values for reference groups
   "num" takes number of predictors to be included
*/
%macro uniReg(varlist, reflist, num);
         %local i;
   %do i =1 %to &num;
         %let var&i = %scan(&varlist, &i);
         %let ref&i = %scan(&reflist, &i);
   %end;

   %do j=1 %to &num;
      proc genmod data = wuss13.cohort3;
      class id &&var&j(ref="&&ref&j" param=ref);
      model disease = &&var&j / dist = poisson link = log;
       repeated subject = id / type = unstr;
       ods output GEEEmpPEst = myout;
       run;

       data myout_exp;
       retain Parm Level1 Estimate expest;
       set myout;
       expest = exp(Estimate);
       run;

       data _null_;
       set myout_exp;
```

```
        if Parm = "Intercept";
        call symputx('baserate', expest);
        run;

        data myout_IR_&j (drop=expest);
        retain Parm Level1 Estimate expest IncidenceRatePer1000
RelativeRisk;
        set myout_exp;
        if Parm = "Intercept" then RelativeRisk = .;
        else RelativeRisk = expest;
        if Parm = "Intercept" then IncidenceRatePer1000 = &baserate * 1000;
        else IncidenceRatePer1000 = &baserate * expest * 1000;
        run;
   %end;
%mend;

/* Invoke uniReg() */
%uniReg(gene smoke gender, 0 0 0, 3 );
```

Figure 1: List of tables containing the simple modeling results generated by %uniReg()

Figure 2: Table containing the simple modeling results for predictor variable gender

	Parameter	Level1	Estimate	IncidenceRatePer1000	RelativeRisk	Empirical Standard Error Estimates	95% Lower Confidence Limit	95% Upper Confidence Limit	Z	Pr > |Z|
1	Intercept		-2.6191	72.8665747	.	0.0077	-2.6342	-2.6041	-340.64	<.0001
2	gender	1	0.5090	121.22020857	1.6635914212	0.0092	0.4910	0.5270	55.35	<.0001

THE MULTIPLE REGRESSION MODEL

Finally, we build a multiple *modified* Poisson regression model that includes all 3 predictors for each of the outcome variables **disease**. A multiple model allows us to summarize the information the about data jointly while adjusting for possible confounding factors. The following program does the job:

```
proc genmod data = wuss13.cohort3;
class id
      gene (ref="0"  param=ref)
      smoking (ref="0"  param=ref)
     gender (ref="0"  param=ref);
model disease = gene smoking gender
                  /dist = poisson link = log;
repeated subject = id/ type = unstr;
ods output GEEEmpPEst = myout;
run;

/* Exponentiate parameter estimates */
data myout_exp;
```

```
retain Parm Level1 Estimate expest;
set myout;
expest = exp(Estimate);
run;

/* Keep the exponentiated Intercept estimate as the baseline incidence
rate */
data _null_;
set myout_exp;
if Parm = "Intercept";
call symputx('baserate', expest);
run;

/* Calculate Incidence rate and relative risk */
data myout_IR_dis (drop=expest);
retain Parm Level1 Estimate expest RelativeRisk;
set myout_exp;
if Parm = "Intercept" then RelativeRisk = .;
else RelativeRisk = expest;
run;
```

Figure 3: Parameter estimates with a multiple regression for outcome disease

	Parameter	Level1	Estimate	RelativeRisk	Empirical Standard Error Estimates	95% Lower Confidence Limit	95% Upper Confidence Limit	Z	Pr > |Z|
1	Intercept		-3.3183	.	0.0085	-3.3350	-3.3015	-388.82	<.0001
2	gene	1	1.6045	4.9753228582	0.0100	1.5849	1.6241	160.45	<.0001
3	gene	2	2.3955	10.974144799	0.0081	2.3797	2.4113	297.31	<.0001
4	smoking	1	0.8768	2.4032045316	0.0077	0.8616	0.8920	113.24	<.0001
5	gender	1	0.5065	1.6595119168	0.0084	0.4901	0.5229	60.56	<.0001

RESULTS AND DISCUSSION

The *modified* Poisson regression model is used to compute the relative estimates for the predictors. There are total 3 predictors included in the model (**Table 1**). The reference/baseline groups are set to those with the lowest risk at our knowledge. The outcome variable, "disease", is modeled by *modified* Poisson regression. Relative risk estimates and their statistical significance are shown above in **Figure 3**.

From **Figure 3**, the relative risk for developing disease is 4.98 higher in those with one copy of the risk allele, and 10.97 higher in those with two copy of the risk allele as compared to those who do not carry the risk allele after adjusting for all other covariates. The relative risk for developing disease is 2.40 higher in those who smoke as compared to those who never smoked after adjusting for all other covariates. The relative risk for developing disease is 1.66 higher in males as compared to females after adjusting for all other covariates.

Note that, from **Table 2** below, an odds ratio computed from a logistic regression model always overestimates the relative risks computed from the other three models. The effect of overestimation increases as the disease prevalence rate increases. That is, odds ratio can dramatically overestimate relative risk for common outcomes (outcomes with higher prevalence rate). Also note that Poisson and modified Poisson regression generate same the point estimate of relative risk as that from log-binomial regression. However, Poisson regression results in wider confidence interval as compared to log-binomial regression due to the misspecification of outcome distribution. The modified Poisson regression is able to correctify the confidence interval by estimating a robust variance error. The resulting confidence interval from the modified Poisson regression is identical to that from the log-binomial regression.

Table 2: Summary of relative risk and 95% CI in different simulated population data set

	Disease Prevalence Rate					
	1%		5%		10%	
Model Type	Relative Risk	95% CI	Relative Risk	95% CI	Relative Risk	95% CI
Logistic regression	2.764	2.617-2.919	2.643	2.576-2.711	2.697	2.647-2.748
Log-binomial regression	2.7307	2.5871-2.8822	2.5026	2.4430-2.5636	2.4114	2.3721-2.4515
Poisson regression	2.7307	2.5862-2.8832	2.5026	2.4412-2.5655	2.4114	2.3696-2.4540
Modified Poisson regression	2.7307	2.5871-2.8822	2.5026	2.4430-2.5636	2.4114	2.3721-2.4515

Table 2 (continued): Summary of relative risk and 95% CI in different simulated population data set

	Disease Prevalence Rate					
	20%		50%		80%	
Model Type	Relative Risk	95% CI	Relative Risk	95% CI	Relative Risk	95% CI
Logistic regression	2.85	2.810-2.891	3.099	3.059-3.138	3.281	3.220-3.344
Log-binomial regression	2.2491	2.2249-2.2736	1.6552	1.6466-1.6638	1.2085	1.2056-1.2116
Poisson regression	2.2491	2.2215-2.2771	1.6552	1.6421-1.6683	1.2085	1.2007-1.2164
Modified Poisson regression	2.2491	2.2249-2.2736	1.6552	1.6466-1.6638	1.2085	1.2056-1.2116

CONCLUSIONS

This paper has demonstrated the correct application of a *modified* Poisson regression method to directly estimate relative risk from a cohort data set, which has quickly gain popularity in medical and public health research. Simulated population data is used to illustrate statistical methods with PROC GENMOD in SAS® 9.2. OR estimation with a logistic regression, which is less desirable for assessment of risk in a cohort study with more common outcomes, has also been demonstrated here. In addition, SAS® ODS and Macro facilities relating to modeling building and reporting results has been introduced in this paper.

REFERENCES

1. Li-Hao Chu, Fagen Xie (2012). Using SAS to Calculate Incidence and Prevalence Rates in a Dynamic Population, WUSS paper, proceedings 12-103.
2. Russ Lavery (2010), An Animated Guide: An Introduction To Poisson Regression, NESUG paper, sa04.
3. Steve Selvin, Statistical Analysis of Epidemiological Data, Oxford press.
4. Lawrence L. Kupper (2008), Applied Regression Analysis and Other Multivariable Methods, Thomson.
5. Arthur Li (2013), Handbook of SAS DATA Step Programming, CRC press.
6. McNutt LA, Wu C, Xue X, Hafner JP (2003). Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes, Am J Epidemiol 2003; 157(10):940-3.
7. Zou G (2004). A Modified Poisson Regression Approach to Prospective Studies with Binary Data. Am J Epidemiol 2004; 159(7):702-6.
8. Sander Greenland (2004). Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies, American Journal of Epidemiology 2004;160:301-305.
9. Karla Lindquist, How can I estimate relative risk in SAS using proc genmod for common outcomes in cohort studies?.from http://www.ats.ucla.edu/stat/sas/notes2/ (accessed November 24, 2007).
10. Rick Wiklin (2013), Simulating Data with SAS, SAS Press.

ACKNOWLEDGMENTS

The author would like to thank Emily Putnam-Hornstein for her support, Arthur Li for his insightful comments and review.

CONTACT INFORMATION

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

Name: Kechen Zhao
Enterprise: University of Southern California,
Department of Preventive Medicine,
Division of Biostatistics.
Address: 2001 N. Soto Street, Los Angeles, CA 90032
Mailbox #112
Phone: 510-584-1950
E-mail: zhao_kechen@hotmail.com