Using Log-linear and Logistic Regression for Inferences on Adjusted Estimates of Relative Risk in Randomized Comparative Trials

William D. Johnson[1], William H. Replogle[2] and Hongmei Han[1]

1. Pennington Biomedical Research Center, 6400 Perkins Rd, Baton Rouge, LA 70808 USA
2. University of Mississippi Medical Center, 2500 North State Street, Jackson, MS 39216 USA

ABSTRACT

Randomized comparative trials are often used to assess the relative merits of two or more interventions aimed at having beneficial effects on the incidence of categorical outcomes. In simple applications chi-square tests can be used to analyze contrasts among proportions of incident events or risk ratios (relative risks). However, assessment of intervention differences may be obscured by outcome variations attributable to covariates. There are advantages to using logistic regression analysis to assess intervention effects in terms of odds ratios (*OR*) adjusted for covariates. The limitation of using *OR* rather than relative risk (*RR*) estimates in making statistical inferences about incidence rates is well documented. Subject-specific estimates of probabilities for a specified covariate profile are readily obtained by logistic and log-linear regression models. Functions of the marginal probabilities provide estimates of incident risk and *RR* for each intervention. We illustrate novel applications of the inferential methods in this paper.

Key Words: Binomial, Generalized linear model**,** Incidence, Large sample inference, Odds ratio

1. INTRODUCTION

In studying the occurrence of dichotomous events, a fundamental concept is the probability a person randomly selected from a well-defined population has a characteristic of interest. In randomized controlled trials, it is often of interest to compare the probability an event occurs (i.e., incidence) among subjects who are in an active treatment group to the probability it occurs in a control (e.g., placebo) group. Thus, if *A* is a dichotomous outcome and incidence in the comparative groups is expressed, respectively, as *P*(*A*|*Active*) and *P*(*A*|*Control*), statistical inferences are often made in terms of the risk difference: *Riskdiff* = *P*(*A*|*Active*) − *P*(*A*|*Control*); in terms of summary statistics such as the relative risk: *RR* = *P*(*A*|*Active*)÷*P*(*A*|*Control*); or in terms of the odds ratio: *OR* = *Odds*(*A*|*Active*) ÷ *Odds*(*A*|*Control*). For example, if = *P*(*A*|*Active*) = 0.085 and P(*A*|*Control*) = 0.025 then *Riskdiff* = 0.085 − 0.025 = 0.060, *RR* = (0.085 ÷ 0.025) = 3.40, $Odds(A|Active) = 0.085 \div 0.915 = 0.093$, $Odds\ (A|Control) = 0.025 \div 0.975 = 0.026$, and *OR* = $0.093 \div 0.026 = 3.58$. Various versions of the chi-square test are widely used to test hypotheses relevant to the research that give rise to use of these summary statistics. In investigations where the outcome is a dichotomous random variable, as in those where the outcome is continuous, concomitant variables that are related to the outcome variable may be used as covariates in an appropriate statistical model to reduce residual variability and enable more sensitive statistical testing significance of differences between treatments. For dichotomous outcomes, statistical modeling is often performed employing logistic regression analysis where the logarithm of the odds of the outcome is modeled in terms of intervention effects coupled with one or more covariates that are potential effect modifiers. The results are typically reported as the covariate-adjusted odds ratio defined as the odds of a favorable outcome given a trial participant received the active treatment relative to the odds of a favorable outcome given the participant received the intervention control. Logistic regression analysis had its origin in the development of analytic models that are suitable for analyzing dichotomous outcomes in retrospective case-control studies. In these studies, a "random" sample of persons with a disease (cases) was investigated to estimate the probability of having a characteristic (exposure) given a person is a case. Similarly a "random" sample of persons who do not have the disease was investigated to estimate the probability of having previously acquired the characteristic given a person is from the control population.

It was shown that if the disease is rare in the general population, the odds ratio provides a robust estimate of the relative risk of disease given a randomly selected person has the characteristic. Widespread historical use of logistic regression models in case-control studies led to development of user friendly statistical software that subsequently became widely employed for the analysis of dichotomous outcomes in prospective follow-up studies. Although relevant software has improved substantially in the last decade, there continues to be a preponderance of literature reporting research results in terms of odds ratios estimated from logistic regression analysis. However, the precise interpretation of an odds ratio is not always transparent. For example, if a patient with a health impairment is told the probability of improvement is 2.8 times greater if he/she takes a specified medication, he/she is likely to understand better than if he/she is told the odds for improvement are 3.3 greater with the treatment compared to placebo (i.e., OR = 3.3).

In this paper, we give three examples to illustrate use of prevalence ratios and probability ratios (relative risk) analogously use of least square means in the analysis of continuous random variables. In this context, the focus of the analysis is on inferences pertaining to estimating (i.) prevalence of a specified attribute relevant to cross-sectional probability samples from well-defined populations and (ii.) probability (incidence) of a specified event relevant to comparing interventions in randomized controlled trials. We employ the SAS procedure PROC GLIMMIX to illustrate the analytical details.

2. GENERALIZED LINEAR MODEL

In the generalized linear model, the dependent variable **Y** is assumed to have a distribution that is a member of the exponential family (e.g., the normal, binomial and Poisson distributions). It is further assumed that the expected value of **Y,** denoted E(**Y**) = **μ,** is linked to the independent **X** through

$$\mathbf{E}(\mathbf{Y}) = \mu = \boldsymbol{g}^{-\mathbf{1}}(\mathbf{X}\boldsymbol{\beta})$$

where g is the link function (e.g., log, logistic), **β** is a vector of unknown parameters. This model enables using a unified approach to the analysis of a wide class of applied statistical problems. The examples in this paper assume the distribution of Y is binomial and either the logistic or log link is appropriate.

3. ILLUSTRATIVE APPLICATIONS

The first two illustrative examples used herein are based on data from Stokes, Davis and Koch (2012). Example 3 is an extension of an example of logistic regression for random intercepts given in SAS 9.22 User's Guide.

Example 1.1 Inferences on Odds Ratio

PROC GLIMMIX performs estimation and statistical inference for generalized linear mixed models. This procedure can be used to calculate probabilities, odds ratios, and relative risks. The following example was based on a study of coronary artery disease and was used here to demonstrate how to fit a generalized linear mixed model for binomial data and estimate odds ratios with the GLIMMIX procedure.

The study population consisted of people who visited a clinic on a walk-in basis and required a catheterization. Investigators were interested in determining whether electrocardiogram (ECG) measurement was associated with disease status.

The following DATA step creates the data set for the analysis.

```
Data CAD;
input sex ecg cad count;
datalines;
0 0 0  11
0 0 1   4
0 1 0  10
0 1 1   8
1 0 0   9
1 0 1   9
1 1 0   6
1 1 1  21;
```

The variable CAD is the response variable indicating the presence or absence of disease. Variables SEX and ECG (ST segment depression) are explanatory variables. First consider the SAS code below:

```
PROC GLIMMIX data = CAD order=data;
  CLASS sex ecg;
  MODEL cad(event = '1')  = sex ecg / dist = binomial link = logit solution;
  LSMEANS  sex ecg / ilink oddsratio cl;
  estimate ‘OR sex' sex 1 -1 / exp cl;
  estimate ‘OR ecg' ecg 1 -1 / exp cl;
RUN;
```

The PROC GLIMMIX statement invokes the procedure. CLASS defines SEX and ECG as classification variables. MODEL defines the model, DIST = Binomial indicates CAD is assumed to have a binomial distribution, LINK = LOGIT specifies the logistic model and SOLUTION requests a listing of estimates of fixed effects parameters. The LSMEANS statement requests the least squares means of the fixed effects on the logit scale. The CL option requests confidence limits for least square means. The ILINK option adds estimates, standard errors, and confidence limits on the mean scale. The results are displayed in Table 1 and Table 2.

Table 1: Summary of model estimates - least squares means

Parameter	**Estimate**	**Standard Error**	**Pr > ChiSq**
Intercept	-1.1747	0.4854	0.0155
Sex	1.2770	0.4980	0.0103
ECG	1.0545	0.4980	0.0342

Table 2: Summary of model estimates - odds ratio

Effect	**Odds Ratio**	**95% Wald Confidence Limits**	
M v F	3.586	1.351	9.516
Abn v Norm	2.871	1.082	7.618

Both variables SEX and ECG are significant ($p < 0.05$). The model equation can be written as follows:

$ln(odds) = \alpha + \beta_1(sex) + \beta_2(ecg)$

= -1.1747 + 1.2770 SEX + 1.0545 ECG

The odds ratio for males compared to females is the ratio of the predicted odds of CAD for males versus females, as shown below. This odds ratio is significant indicating the odds for having CAD is higher for males. Similarly, the significant odds ratio for abnormal ECG versus normal ECG indicated the odds of having CAD is higher for an abnormal ECG reading.

$OR(CAD|M:F) = exp(1.2770) = 3.59$

$OR(CAD|Abn:Norm) = exp(1.0545) = 2.87$

Example 1.2 Inferences on Prevalence Ratio

Prevalence Ratio (PR) can be requested in the GLIMMIX procedure by simply changing the link "logit" to "log" to specify he log-linear model, as shown below. The results are displayed in Table 3 – 5.

```
PROC GLIMMIX data = CAD order=data;
  CLASS sex ecg;
  MODEL cad(event = '1')  = sex ecg / dist = binomial
            link = log solution;
  LSMEANS  sex ecg / ilink oddsratio cl;
  estimate 'PR sex' sex 1 -1 / exp cl;
  estimate 'PR ecg' ecg 1 -1 / exp cl;
RUN;
```

The variables SEX is significant ($p < .05$) and the variable ECG is marginally significant ($p = 0.05$). The model equation can be written as follows:

$ln(Prev) = \alpha + \beta_1(sex) + \beta_2(ecg)$

= -1.2830 + 0.5774 SEX + 0.4565 ECG

Table 3: Summary of model estimates - least squares means

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-1.2830	0.2777	75	-4.62	<.0001
Sex	0.5774	0.2464	75	2.34	0.0218
ECG	0.4565	0.2306	75	1.98	0.0514

The prevalence of CAD is clearly higher in males than in females and within sex groups it's higher for those with abnormal ECG (Table 4).

Table 4: Prevalence of CAD by SEX and ECG status

Sex	ECG	n	Obs	Pred	LCB	UCB
F	Normal	15	0.267	0.236	0.107	0.444
F	Abnormal	18	0.444	0.470	0.281	0.668
M	Normal	18	0.500	0.526	0.328	0.715
M	Abnormal	27	0.778	0.761	0.590	0.875

A preliminary model indicated SEX by ECG interaction was not significant and therefore the main effect model provides a good summary of the data (Table 5). The main effect prevalence of CAD for males is 0.621 and for females it is 0.348.

Table 5: Prevalence of CAD by SEX

Sex	Beta	SE	DF	t-test	p	Lower	Upper	Mean	SEM	LCB	UCB
M	-0.477	0.123	75	-3.87	0.0002	-0.723	-0.232	0.621	0.077	0.485	0.793
F	-1.055	0.230	75	-4.58	<.0001	-1.513	-0.596	0.348	0.080	0.220	0.551

For males vs. females, the prevalence ratio for CAD is 1.781. The 95% confidence intervals for the prevalence ratio are 1.090 to 2.911 (Table 6).

Table 6: Prevalence Ratio (PR) for CAD: males vs. females

Beta(diff)	SE	DF	t-test	p	Alpha	Lower	Upper	PR	LCB	UCB
0.577	0.246	75	2.34	0.022	0.05	0.086	1.068	1.781	1.090	2.911

Example 2.1 Relationship between Prevalence of CAD and Age

Age was added to the previous dataset to investigate the overall relationship between prevalence of CAD and age. As a first step, AGE was tested for significance in predicting the prevalence of CAD.

```
PROC GLIMMIX data = CAD;
    MODEL cad(event = '1')  = age/
            dist = bin link = log solution oddsratio;
RUN;
```

As shown in Table 7, AGE is a significant predictor of the prevalence of CAD ($p < 0.05$). The prevalence is increasingly higher in the older age groups (Table 8 and Figure 1).

Table 7: Summary of model estimates

Effect	Estimate	Std Error	DF	t Value	Pr > |t|
Intercept	-2.7328	0.7506	76	-3.64	0.0005
AGE	0.04338	0.01430	76	3.03	0.0033

$$Predicted\ Prevalence = exp(-2.7328 + 0.04338\ age)$$

Table 8: Estimated Prevalence of CAD

Age (yr)	Est Prev CAD
30	0.238
40	0.369
50	0.569
60	0.878

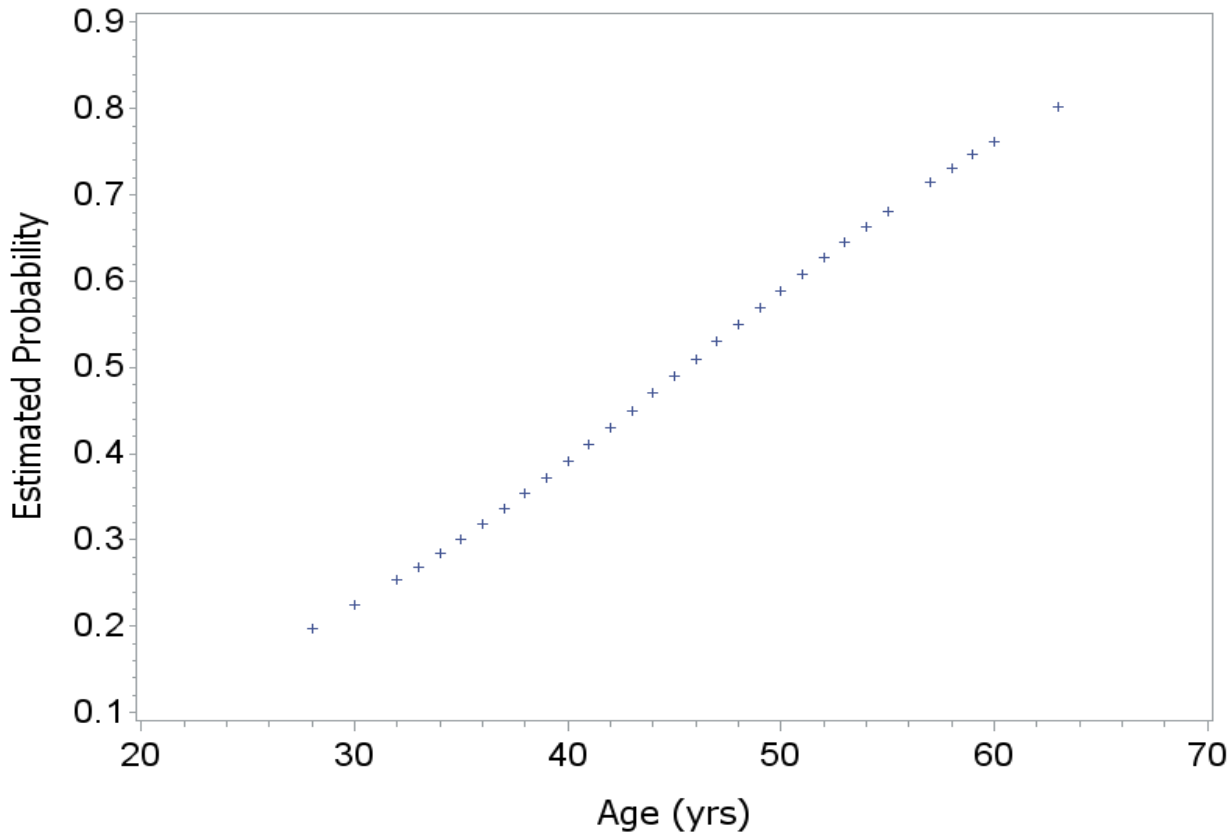

Figure 1: The relationship between estimated prevalence (probability) of CAD and age (yrs).

Example 2.2 Prevalence Ratio with Age as a Covariate

Adding AGE as a covariate in the previous model, SEX was significant in the unadjusted model and ECG was borderline so results were virtually the same after adjusting for AGE. However, some useful insights into the role of AGE are provided by the results summarized in Table 9. Specifically, prevalence of CAD increases incrementally with AGE and AGE is a significant predictor of CAD irrespective of sex and ECG status; prevalence of CAD is significantly higher in males than females irrespective of ECG status or age (PR=1.727); and prevalence of CAD is significantly higher in persons with ECG status '2' versus those with ECG status '1' irrespective of sex or age (PR=1.766).

```
PROC GLIMMIX data = CAD order=data;
  CLASS sex ecg;
  MODEL cad(event = '1')  = sex ecg age/
    dist = binomial link = log solution ;
  LSMEANS  sex ecg / ilink diff oddsratio cl ;
  estimate 'Sex' sex 1 -1 / exp cl;
  estimate 'ECG 2 vs 1' ecg 1 -1 0 / exp cl;
  estimate 'ECG 2 vs 0' ecg 1 0 -1 / exp cl;
  estimate 'ECG 1 vs 0' ecg 0 1 -1 / exp cl;
  estimate 'Age 1 yr' age 1 -1 / exp cl;
  estimate 'Age 10 yrs' age  1 -1 / exp cl  ;
RUN;
```

Table 9: Prevalence Ratios and Model Estimates

	Mean Estimate	Mean Confidence		L'Beta Estimate	Standard Error	Alpha	L'Beta Confidence		Chi-Square	Pr > ChiSq
Sex	1.727	1.090	2.734	0.546	0.235	0.05	0.086	1.006	5.42	0.020
Exp(Sex)				1.727	0.405	0.05	1.090	2.734		
ECG 2 vs 1	1.107	0.758	1.616	0.102	0.193	0.05	-0.277	0.480	0.28	0.599
Exp(ECG 2 vs 1)				1.107	0.214	0.05	0.758	1.616		
ECG 2 vs 0	1.766	1.042	2.995	0.569	0.269	0.05	0.041	1.097	4.46	0.035
Exp(ECG 2 vs 0)				1.766	0.476	0.05	1.042	2.995		
ECG 1 vs 0	1.596	0.970	2.625	0.467	0.254	0.05	-0.031	0.965	3.39	0.066
Exp(ECG 1 vs 0)				1.596	0.405	0.05	0.970	2.625		
Age 1 yr	1.034	1.011	1.058	0.034	0.011	0.05	0.011	0.056	8.76	0.003
Exp(Age 1 yr)				1.034	0.012	0.05	1.011	1.058		
Age 10 yrs	1.400	1.120	1.749	0.336	0.114	0.05	0.114	0.559	8.76	0.003
Exp(Age 10 yrs)				1.400	0.159	0.05	1.120	1.749		

Example 3. Relative Risk in Multicenter Randomized Trials

In this example, researchers investigated the performance of two medical procedures in a multicenter study. They randomly selected 15 centers for inclusion. Patients were randomly assigned to one of the two procedures to compare the occurrence of their side effects.

The following DATA step creates the data set for the analysis.

```
data multicenter;
input center group$  n SideEffect @@;
datalines;
1 A 32 14   1 B 33 18   2 A 30 4    2 B 28 8
3 A 23 14   3 B 24 9    4 A 22 7    4 B 22 10
5 A 20 6    5 B 21 12   6 A 19 1    6 B 20 3
7 A 17 2    7 B 17 6    8 A 16 7    8 B 15 9
9 A 13 1    9 B 14 5    10 A 13 3   10 B 13 1
11 A 11 1   11 B 12 2   12 A 10 1   12 B 9 0
13 A 9 2    13 B 9 6    14 A 8 1    14 B 8 1
15 A 7 1    15 B 8 0
;
```

The variable group identifies the two procedures; n is the number of patients who received procedure A or B each center, and SideEffect gives the number of patients who reported side effects. The random variable SideEffect/n is assumed to have a binomial distribution. The random option specifies a random intercepts model. The SAS code is listed below:

```
proc glimmix data=multicenter;
    class center group;
    model sideeffect/n = group / dist = binomial link =  log solution;
    random intercept / subject=center;
    lsmeans group /ilink diff cl;
    estimate 'RR' group 1 -1 / exp cl;
RUN;
```

The probability of side effects was significantly lower in group A (0.221 vs. 0.298) (Table 10) with relative risk 0.741, p = 0.034 (Table 11).

Table 10: Estimates of probabilities of side effects by procedures.

Group	Beta	SE	DF	t-test	p	Lower	Upper	Mean	SEM	LCI	UCI
A	-1.509	0.180	14	-8.37	<.0001	-1.896	-1.122	0.221	0.040	0.150	0.326
B	-1.210	0.170	14	-7.12	<.0001	-1.574	-0.845	0.298	0.051	0.207	0.429

Table 11: Relative Risk (RR) of side effects in procedure A vs. procedure B.

Beta(diff)	SE	DF	t-test	p	Alpha	Lower	Upper	RR	LCI	UCI
-0.300	0.128	14	-2.35	0.034	0.05	-0.573	-0.026	0.741	0.564	0.974

4. CONCLUDING REMARKS

The practice of using odds ratios as the primary summary statistics in reporting results from comparative studies where the outcome is dichotomous appears to have been largely driven in the past by lack of convenient software to easily conduct relevant statistical analyses. The software has improved in recent years but odds ratios continue to be reported frequently in the scientific literature. Nevertheless, reporting prevalence or incidence of events in studies with dichotomous outcomes has great heuristic appeal and it is inevitable that the analytical focus will shift to prevalence and incidence. The somewhat rudimentary examples presented here are intended to demonstrate ease of use in rudimentary applications rather than in a range of complex study designs.

REFERENCES

Stokes M. E., Davis, C. S. and Koch, G. G. *Categorical Data Analysis Using the SAS® System*, Cary, NC: SAS Institute Inc., 2012. 580 pp.

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

ACKNOWLEDGEMENT

Supported by 1 U54 GM104940 from the National Institute of General Medical Sciences of the National Institutes of Health which funds the Louisiana Clinical and Translational Science Center

Submitting author: William D. Johnson, Ph.D., Department of Biostatistics, Pennington Biomedical Research Institute, Louisiana State University, 6400 Perkins Road, Baton Rouge, LA 70808-0102, USA
Phone (225) 763-2932, E-mail: william.johnson@pbrc.edu