Research Methods

Statistical Issues in Estimation of Adjusted Risk Ratio in Prospective Studies

Leila Janani PhD[1,2], Mohammad Ali Mansournia MD MPH PhD[1], Keramat Nourijeylani PhD[1], Mahmood Mahmoodi MPH PhD[1], Kazem Mohammad PhD•[1]

Abstract

Background: Binary outcomes are common in prospective studies such as randomized controlled trials and cohort studies. Logistic regression is the most popular regression model for binary outcomes. Logistic regression yields an odds ratio that approximates the risk ratio when the risk of outcome is low. A consensus has been reached in an extensive argument in much of the literature that the risk ratio is preferred over the odds ratio for prospective studies. To obtain a model-based estimate of risk ratios, log-binomial regression has been recommended. However, this model may fail to converge and many methods have been provided as an alternative in these situations.

Methods: In this paper, we discuss the methods to obtain adjusted risk ratios in settings with independent and clustered data and we will review the results of comparisons between these methods based on simulation studies, especially a large simulation study which was conducted by the authors. We use hypothetical examples to show how log-Poisson regression with modified standard errors can be used to estimate risk ratio in practice using popular statistical software.

Conclusion: The potential misinterpretation of odds ratios should be considered by researchers, especially when the risk of the outcome is high. When researchers want to estimate the effect of exposure or intervention by controlling potential covariates, the misinterpretation of odds ratios can be avoided using regression models that can estimate risk ratios instead of logistic regression. The log-Poisson regression with modified standard errors can be considered to estimate risk ratios in both independent and clustered data settings.

Keywords: Binary outcome, log-binomial regression, prospective study, risk ratio, simulation

Cite this article as: Janani L, Mansournia MA, Nourijeylani K, Mahmoodi M, Mohammad K. Statistical Issues in Estimation of Adjusted Risk Ratio in Prospective Studies.*Arch Iran Med*. 2015; 18(10): 713 – 719.

Background

Binary outcomes are common in prospective studies such as randomized controlled trials and cohort studies. Logistic regression, which estimates odds ratios, is the most popular regression model for binary outcomes.[1] If binary outcomes are represented by Yi for subject i (I =1, 2,..., n) where Yi =1 (if the subject experiences the outcome of interest) and Yi =0 otherwise, then logistic regression can be written as:

$$Log\ it\ (\pi_i) = Log\ (\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

Where $\pi_i = E[Y_i]$ is the probability of experiencing the outcome of interest for subject i, and X_{1i}, ..., X_{ki} are predictor variables. Based on this model, the effect of each covariate on the outcome can be expressed as an odds ratio by $\exp(\beta_i)$ As a measure of effect odds ratio, calculated as the ratio of two odds, is difficult to interpret[2–4] and it is not collapsible[5,6]; this means that the odds ratio can be identical in each stratum defined by the levels of a covariate which is not a confounder, but the overall odds ratio can differ from the stratum-specific odds ratios. For example, Mansournia and Greenland[7] presented a numerical example for non-collapsibility of odds ratio. Consider Table 1 which shows the association between a binary outcome (D) and an exposure (E) stratified on a binary covariate (C). This example illustrates a situation where exposure and covariate are independent and the covariate cannot be a cofounder. The stratum-specific odds ratios comparing exposed to unexposed subjects are 6, but the overall odds ratio ignoring C is equal to 3.45.

In contrast to the odds ratio, the 'risk ratio', also called the 'relative risk', is considered simple to interpret[2–4,8] and the risk ratio is also collapsible.[5,6]

In case–control studies, the odds ratio is an appropriate effect measure. Depending on the sampling method, the odds ratio can be sometimes interpreted as a risk ratio or rate ratio in case-control studies.[9–11] When the risk of the outcome is low, the difference between the odds ratio and risk ratio is negligible[12]; however, in spite of repeated emphasis on the importance of the low risk assumption, consumers of medical reports often interpret the odds ratio as a risk ratio even in studies with common outcomes. Knol,*et al.*[13] performeda survey of published cohort studies (n = 75) and randomized controlled trials (n = 288) and reported that about one-third of cohort studies calculated odds ratios adjusted for baseline variables using logistic regression, and 40% of these presented odds ratios differed from the risk ratio more than 20%. Only about 5% of randomized controlled trials reported adjusted odds ratios using logistic regression; however, about two-thirds of these pre-

Authors' affiliation: [1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.[2]Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran.

•**Corresponding author and reprints:**Kazem Mohammad PhD, Professor of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. P.O. Box: 14155-6446, Tehran, Iran. Tel: +98-21-88951396; Fax: +98-21-88989127; E-mail: mohamadk@tums.ac.ir

Accepted for publication: 18 June 2015

Table 1. An example of odds ratio non-collapsibility without confounding.

D	C=1		C=0		Collapsed	
	E		E		E	
	1	0	1	0	1	0
1	180	60	80	10	260	70
0	20	40	120	90	140	130
Total	**200**	**100**	**200**	**100**	**400**	**200**
	OR = 6		**OR = 6**		**OR = 3.45**	

OR = indicates odds ratio.

sented odds ratios differed more than 20% from the risk ratio.

A consensus has been reached in an extensive argument in much of the literature that the risk ratio is preferred over the odds ratio for prospective studies.[3,12,14,15]

To obtain model-based estimate of risk ratios directly, log-binomial regression has been recommended.[16] Log-binomial regression model is similar to logistic regression model, except that it assumes a log link instead of a logit link, hence providing risk ratios instead of odds ratios. It can be presented as,

$$\log(\pi_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \quad (2)$$

Based on this model, the effect of each covariate on the outcome can be expressed as a risk ratio by $\exp(\beta_i)$.

This model has to satisfy certain restrictions to ensure that the probability of outcome lies between 0 and 1. Since π_i is a probability which must lie between zero and one, the left-hand side of model (2) is constrained to be less than or equal to zero, while the right-hand side is unconstrained. The model may fail to converge as a result of these restrictions. When the maximum likelihood estimate lies on the boundary of the parameter space, then convergence will not occur and it will fail to provide an estimate of the risk ratio.[17,18]

The mentioned difficulties with odds ratios can also occur in the presence of clustered data. The usual assumption of independent observations required for ordinary regression models is violated in the presence of clustering. Clustered data are common in prospective studies, and may occur as a result of repeated measurements on the same subject over time (i.e., longitudinal data), or measurements taken at the same time on sub-units within the primary unit (e.g., patients within clinics).

When clustering is present in the data, it should be taken into account in statistical analysis. One common approach to account for clustering in statistical analysis is using generalized estimating equations (GEE). GEEs use a marginal or population-averaged approach and control clustering implicitly through the use of a working correlation structure.[19] The other approach which can be considered in practice is generalized linear mixed models (GLMM). Parameters obtained from the two approaches have different interpretations and the choice between them is often important. However, there are some special conditions where the parameters of two approaches coincide. If we assume the log link function, such as log-binomial model, the parameters coincide apart from the intercept. For the logit link, the parameters generally differ.[20]

Due to the potential for convergence problems with log-binomial regression, many alternative methods have been introduced for estimating risk ratios.

In this paper, we discuss methods to obtain adjusted risk ratios in settings with independent and clustered data and we will review the results of comparisons between these methods based on simulation studies, especially a large simulation study which was conducted by the authors. At last, we will conclude with practical recommendations on these methods and their applications in both independent and clustered data settings.

We have to mention that in this paper, we consider prospective study settings where the scientific goal is to estimate the effect of an exposure or intervention on a common binary outcome, with adjustment for additional covariates. When prediction is the scientific goal of a study, models in which the constraints on the probabilities are automatically satisfied (e.g., logistic regression) should be considered.[21]

Several alternatives to log-binomial regression for estimating adjusted risk ratio in independent data settings

The Mantel-Haenszel risk ratio method is straightforward and gives a weighted risk ratio over strata of covariates.[22,23] However, this method can only be used to control categorical but not continuous covariates.[17] Only if continuous covariates are categorized, we can use this method for estimating the adjusted risk ratio by controlling these types of variables.

Some of the alternative methods for estimating risk ratios focus on improving the convergence of the log binomial model by constraining the estimation process. Wacholder[16] described a constrained iterative estimation procedure and Yu and Wang[24] suggested using the nonlinear programming procedure PROC NLP, available in SAS statistical software, to constrain the estimation process.

Deddens, *et al*.[18] proposed a method which involves analyzing a modified dataset for obtaining a solution close to the maximum likelihood estimate. The new dataset contains C-1 copies of the original data and 1 copy of the original dataset with the outcomes reversed (Y set to 1 − Y). Risk ratio can be estimated using the log-binomial regression model for the modified dataset. Authors suggested using C = 1000 in practice and called this the "COPY 1000" method.

Several authors have proposed a conversion formula for calculating risk ratios from odds ratios.[3,25,26] For example, the Zhang and Yu[25] method is a simple formula that calculates the risk ratio based on the odds ratio and the incidence of the outcome in the unexposed group.

Schouten, *et al*.[27] provided the doubling of-cases method, which concerns changing the original dataset in such a way that logis-

tic regression yields risk ratio instead of an odds ratio. They suggested duplicating each observation that has Y = 1, setting Y = 0 for the duplicate. The probability of success in the original dataset will be equal to the odds of success in the modified dataset and so a logistic regression model fitting to the new dataset results in risk ratio instead of an odds ratio. The robust standard errors are needed to account for the within-subject correlation resulted from the duplicated observations.

Some authors have named this method "expanded logistic regression".[28]

Lee[2] suggested that Cox's proportional hazards model can be used to estimate risk ratios if the risk period is held constant.

McNutt, *et al*.[15] proposed estimating risk ratios using a log-Poisson regression model (i.e., using a log link function and a Poisson distribution for response). Log-Poisson regression is expected to overestimate the standard errors of the parameter estimates. Zou[29] suggested using robust standard error and termed this approach 'modified Poisson regression'.

Lumley, *et al*.[1] suggested using a log link function and a normal distribution for response. Again, standard errors should be corrected using a robust 'sandwich' variance estimator or applying jackknife or bootstrapping methods.

A Bayesian approach using a Markov-chain Monte-Carlo method, with a focus on applying the linear inequality constraint and the estimation of risk ratio from a log-binomial model has been proposed by Chu and Cole.[30]

Alternatives to log-binomial regression for estimating adjusted risk ratio in clustered data settings

Each of the methods for estimating risk ratios based on independent data setting which was described in the previous section could be extended to account for clustering. For example, Zou[31] extended log-Poisson regression to studies with correlated binary outcomes occurring in longitudinal or cluster randomized studies.

Simulation studies

The relative performance of some of the alternative methods has been examined in a number of simulation studies. In particular, log-binomial regression has been compared to constrained log-binomial regression,[32] the COPY method,[17,24,28] expanded logistic regression,[1,28,33,34] log-Poisson regression[1,28,29,32,35] and log-normal regression.[1,28]

Overall, convergence rates can be improved by employing one of the alternative methods to log-binomial regression for risk ratio estimation.

Log-Poisson regression was introduced as a useful method for estimating risk ratio in practice.[1,28,33] The study by Yelland, *et al*.[28] was a large simulation study and 10 different methods were compared based on various statistical characteristics. In conclusion, they recommended that log-Poisson regression can be a useful tool for providing an adjusted estimate of risk ratio. A comparison between the methods for estimating risk ratios in the clustered data is limited. Only two studies[36,37] have been reported. Santos, *et al*.[36] conducted a simulation study to compare log-Poisson regression and other methods based on fitting a logistic regression model and using different methods of standardization to calculate the risk ratio. Based on their study results, the log-Poisson method was inferior to the logistic regression approach in terms of the coverage probability of the Wald 95% confidence interval but this may have occurred because the data were simulated under a logistic model which assumes a constant odds ratio, rather than a constant risk ratio. Yelland, *et al*.[37] conducted a large simulation study and used log-binomial model to simulate the data. They also used GEEs to account for clustering and found that the log-Poisson approach performed well in this setting and can be practically considered.

We also performed a large simulation study to assess the performance of six different methods for estimating the risk ratio in independent and clustered data settings. The results of our study will be published in details in the future. Based on the results of our study, Log-Poisson regression can also be considered in practice. This method is simple to run and popular statistical softwares can be used to estimate the risk ratio aplying this model.

Illustrative examples

In this section, we use a hypothetical example to show how log-Poisson regression can be used to estimate the risk ratio. In our large simulation study, we considered a two-group parallel RCT design comparing a new treatment group to a control group. Half of the subjects (or clusters) were assigned to the treatment group, while the other half were assigned to the control group and a single binary or a single continuous baseline covariate was considered. We defined different scenarios and simulated 1000 datasets for each scenario. For this example, we selected only one dataset for four different scenarios. For independent data setting, we fitted log-binomial regression, log-Poisson regression and log-Poisson regression with robust standard errors. For clustered data setting, log-binomial and log-Poisson regressions were fitted and GEE approach with independent working correlation structure was used to account for clustering. The purpose of the analysis was to estimate the risk ratio of success comparing the treatment group with the control group conditional on the baseline covariates. We used R (version 3.2) to implement the analysis.

Table 2 shows the results of log-binomial regression, log-Poisson regression and modified log-Poisson regression for two scenarios in the settings of independent data. Table 3 shows the results of log-binomial regression and log-Poisson regression for the clustered data setting. When adjustment was made on a continuous covariate, log-binomial regression was not converged in either independent and clustered data settings. This finding is consistent with the results of our large simulation study in which convergence problems mostly occurred when a continuous covariate was adjusted for in the analysis. As seen in Table 2, robust variance estimation is useful to avoid overestimating the standard errors of parameter estimates by log-Poisson regression.

Recommendations for researchers

Case-control studies are usually used when outcomes are rare in the population from which study subjects are sampled. Outcome risks and odds often cannot be estimated directly from case-control data, because the sampling proportions of cases and controls-may be unknown. However, the odds ratio is the appropriate measure of effect in these studies and the adjusted odds ratio can be estimated using logistic regression model.

In cohort studies and randomized controlled trials (RCTs) with binary outcomes, the risk ratio is the preferred measure of effect. If the risk of the outcome is low, the difference between the risk

Table 2. The estimated slope and risk ratio using log-binomial regression, log-Poisson regression and modified log- Poisson regression for the independent data setting.

Type of scenario	Log-binomial regression			Log-Poisson regression			Modified log-Poisson regression		
	$\hat{\beta}$	$\widehat{RR}=\exp(\hat{\beta})$	SE($\hat{\beta}$)	$\hat{\beta}$	$\widehat{RR}=\exp(\hat{\beta})$	SE($\hat{\beta}$)	$\hat{\beta}$	$\widehat{RR}=\exp(\hat{\beta})$	SE($\hat{\beta}$)
200 individuals, a binary covariate with a prevalence of 0.5, risk ratio for treatment = 2	0.704	2.022	0.295	0.701	2.016	0.327	0.701	2.016	0.296
200 individuals, a normal covariate with mean = 0.5 and variance = 0.25, risk ratio for treatment = 2	Not converged	—	—	0.727	2.069	0.338	0.727	2.069	0.302

Table 3. The estimated slope and risk ratio using log-binomial regression and modified log-Poisson regression for the clustered data setting.

Type of scenario	Log-binomial regression			Modified log-Poisson regression		
	$\hat{\beta}$	$\widehat{RR}=\exp(\hat{\beta})$	SE($\hat{\beta}$)	$\hat{\beta}$	$\widehat{RR}=\exp(\hat{\beta})$	SE($\hat{\beta}$)
50 clusters of size 10, binary cluster covariate with prevalence of 0.5, risk ratio for treatment = 2	0.711	2.036	0.186	0.707	2.028	0.187
500 clusters of size 10, normal cluster covariate with mean = 0.5 and V = 0.25, risk ratio for treatment = 2	Not converged	—	—	0.631	1.880	0.178

ratio and the odd ratio will be negligible and so the adjusted odds ratio estimated using logistic regression can well approximate the adjusted risk ratio.

However, if the risk of the outcome is high, which occurs frequently in RCTs, the adjusted risk ratio should be estimated using log-binomial regression model. However, this model has convergence problems and thus, based on the simulation studies especially our simulation study, log-Poisson regression is recommended instead to estimate the adjusted risk ratio. The software implementation of log-binomial and log-Poisson regression models has been described in the Appendix.

Conclusion

In this paper, we have discussed the problems of using odds ratios as an approximation of risk ratios in prospective studies. The potential misinterpretation of odds ratios should be considered by researchers, especially when the risk of the outcome is high. When researchers want to estimate the effect of exposure or intervention, the misinterpretation of odds ratios can be avoided by using regression models which estimate adjusted risk ratios instead of using the logistic regression. Risk ratios can be estimated using log-binomial regression but this model may fail to converge. When this occurs, the log-Poisson regression can be considered to estimate risk ratios in both independent and clustered data settings.

The mentioned regression models for estimating risk ratios are most useful where the scientific goal is to estimate the effect of an exposure or intervention on a common binary outcome. When prediction is the scientific goal of a study, models in which constraints on the probabilities are automatically satisfied (e.g., logistic regression) should be considered

References

1. Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms.*UW Biostatistics Working Paper Series*. Working Paper 293. 2006.
2. Lee J. Odds ratio or relative risk for cross-sectional data? *Int J Epidemiol*. 1994;23(1):201-3.
3. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol*. 1994;47(8):881-9.
4. Walter SD. Choice of effect measure for epidemiological data. *J Clin Epidemiol*. 2000;53(9):931-9.
5. Miettinen OS, Cook EF. Confounding: essence and detection. *Am J Epidemiol*. 1981;114(4):593-603.
6. Cummings P. The relative merits of risk ratios and odds ratios. *Arch Pediatr Adolesc Med*. 2009;163(5):438-45.
7. Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. *Epidemiology*. 2015;26(4):466-72.
8. Schechtman E. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use? *Value Health*. 2002;5(5):431-6.
9. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol*. 1982;116(3):547-53.
10. Greenland S, Thomas DC, Morgenstern H. The rare-disease assumption revisited. A critique of" estimators of relative risk for case-control studies". *Am J Epidemiol*. 1986;124(6):869-83.
11. Knol MJ, Vandenbroucke JP, Scott P, Egger M. What do case-control studies estimate? Survey of methods and assumptions in published case-control research. *Am J Epidemiol*. 2008;168(9):1073-81.
12. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125(5):761-8.
13. Knol MJ, Duijnhoven RG, Grobbee DE, Moons K, Groenwold R. Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials. *PLoS One*. 2011;6(6):e21248.
14. Nurminen M. To use or not to use the odds ratio in epidemiologic analyses? *Eur J Epidemiol*. 1995;11(4):365-71.
15. McNutt LA, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. *Am J Epidemiol*. 2003;157(10):940-3.
16. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986;123(1):174-84.
17. Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly es-

timate the prevalence ratio. *BMC Med Res Methodol*. 2003;3(1):21.
18. Deddens J, Petersen M, Lei X, editors. Estimation of prevalence ratios when PROC GENMOD does not converge. Paper 270–28. Proceedings of the 28th Annual SAS Users Group International Conference Cary NC: SAS Institute Inc. Available from: URL: http://www2 sas com/proceedings/sugi28/270-28 pdf); 2003.
19. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev*. 1991; 59(1):25-35.
20. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Stat Methods Med Res*. 2004;13(4):309-23.
21. Fitzmaurice GM, Lipsitz SR, Arriaga A, Sinha D, Greenberg C, Gawande AA. Almost efficient estimation of relative risk regression. *Biostatistics*. 2014;15(4):745-56.
22. Greenland S, Rothman KJ. Introduction to stratified analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia (PA): Lippincott,Williams& Wilkins; 2008: 258-82.
23. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies. *J Natl Cancer Inst*. 1959;22(4):719-48.
24. Yu B, Wang Z. Estimating relative risks for common outcome using PROC NLP. *Comput Methods Programs Biomed*. 2008;90(2):179-86.
25. Zhang J, Kai FY. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*. 1998;280(19):1690-1.
26. Osborn J, Cattaruzza MS. Odds ratios and relative risk for cross-sectional data. *Int J Epidemiol*. 1995;24(2):464-5.
27. Schouten EG, Dekker JM, Kok FJ, Cessie SL, Van Houwelingen HC, Pool J, et al. Risk ratio and rate ratio estimation in case-cohort designs: Hypertension and cardiovascular mortality. *Stat Med*. 1993;12(18):1733-45.
28. Yelland LN, Salter AB, Ryan P. Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. *Int J Biostat*. 2011;7(1):1-31.
29. Zou G. A modified poisson regression approach to prospective studies with binary data. *Am J Epidemiol*. 2004;159(7):702-6.
30. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology*. 2010;21(6):855-62.
31. Zou G, Donner A. Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Stat Methods Med Res*. 2013;22(6):661-70.
32. Carter RE, Lipsitz SR, Tilley BC. Quasi-likelihood estimation for relative risk regression models. *Biostatistics*. 2005;6(1):39-44.
33. Blizzard L, Hosmer W. Parameter Estimation and Goodness-of-Fit in Log Binomial Regression. *Biom J*. 2006;48(1):5-22.
34. Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol*. 1998;27(1):91-5.
35. Petersen MR, Deddens JA. A comparison of two methods for estimating prevalence ratios. *BMC Med Res Methodol*. 2008;8(1):9.
36. Santos CA, Fiaccone RL, Oliveira NF, Cunha S, Barreto ML, do Carmo MB, et al. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. *BMC Med Res Methodol*. 2008;8(1):80.
37. Yelland LN, Salter AB, Ryan P. Relative risk estimation in cluster randomized trials: a comparison of generalized estimating equation methods. *Int J Biostat*. 2011;7(1):1-26.

Appendix: Implementation using popular statistical software

In this Appendix, we will show how log-binomial and log-Poisson regression models can be implemented in popular statistical software.
We assume that y indicates a binary outcome, x1 is an exposure/ treatment variable and there is a covariate (x2) and we want to estimate the effect of exposure/ treatment on the outcome controlling the covariate variable.
We have used the statistical packages: Stata (version 12), R (version 3.2), SAS (version 9.3) and SPSS (version 22).
Stata
For independent data setting, we can use the command **glm** for running log-binomial regression. A log-Poisson regression with robust standard errors can be estimated by **glm** command using the **robust** option. Also, we can use **Clustered robust** option in **glm** and specify a unique subject identifier (id) as the **Cluster variable**. Another option in Stata is to use **xtgee** command and specify a unique subject identifier (id) as the **Panel ID variable.**
glm y x_1 x_2, family(binomial) link(log) eform
glm y x_1 x_2, family (poisson) link(log) vce (robust) eform
glm y x_1 x_2, family(poisson) link(log) vce (cluster id) eform
xtset id
panel variable: id
xtgeey x_1 x_2, family(poisson) link(log) corr(independent) vce(robust) eform
For clustered data, we can use **xtgee** command in Stata software to estimate the adjusted risk ratio by log-binomial or log-Poisson regression models. We need to specify a unique cluster identifier (cluster) as the **Cluster variable** in **Panel ID variable** in **xtgee** and define the correlation structure. We used the independent working correlation structure for this example.
xtset cluster
panel variable: cluster
xtgee y x_1 x_2, family(binomial) link(log) corr(independent) vce(robust) eform

xtset cluster
panel variable: cluster
xtgee y x_1 x_2, family(poisson) link(log) corr(independent) vce(robust) eform
SAS
In the independent data setting, the log-binomial regression can be conducted simply by specifying binomial distribution and log link in **PROC GENMOD**. To compute robust standard errors for log-Poisson regression, we need to use **REPEATED** statement and specify a unique subject identifier (id) as the **SUBJECT** in the **REPEATED** statement.
proc genmod data=data;
model y = x1 x2/dist=binomial link=log;
run;
proc genmod data=data;
class id;
model y = x1 x2/dist=poisson link=log;
repeated subject=id;
run;
In the clustered data setting, both log-binomial regression and log-Poisson regression can be implemented in in **PROC GENMOD** using **REPEATED** statement and specifying a unique cluster identifier (cluster) as the **SUBJECT** in the **REPEATED** statement. We also need to define a working correlation structure. In this example, we have used the default structure in SAS which is an independent working correlation structure.
proc genmod data=data;
class cluster;
model y = x1 x2/dist=binomial link=log;
repeated subject=cluster;
run;
proc genmod data=data;
class cluster;
model y = x1 x2/dist=poisson link=log;
repeated subject=cluster;
run;
R
If data areindependent, we can use **glm()** function in R for implementing log-binomial regression. Log-Poisson regression with robust variances can be calculated in two different ways. We can use **glm()** function with Poisson distribution and log link and then use library (**sandwich)** to get robust standard errors. We can also use **geeglm()** function in library (**geepack**) to get robust standard errors. We need to specify a unique subject identifier (id) in this function as shown below:
glm(y ~ x_1+ x_2, data = data, family=binomial(link="log"))
poiss<-glm(y ~ x_1+ x_2, data = data, family = poisson (link="log")) library(sandwich)
library (lmtest) # to test coefficients coeftest (poiss, vcov = sandwich)
library(geepack)
geeglm(y ~ x_1+ x_2 , data = data , family = poisson, id = id)

For the clustered data setting, we can use **geeglm()** function in library (**geepack**) and specify a unique cluster identifier (cluster) as follows:
geeglm(y ~ x_1 + x_2, data = data, family=binomial(link="log"), id=cluster, corstr="independence")
geeglm(y ~ x_1 + x_2, data = data, family=poisson(link="log"), id=cluster, corstr="independence")
SPSS
To run the log-binomial regression in SPSS for an independent dataset, we need to use **Generalized Linear Models** menu and in the **Type of Model** sub-menu, we should activate the **Custom** option and then select binomial for the **Distribution** and log for the **Link function**. Log-Poisson regression can be estimated similar to log-binomial regression and to get robust standard errors, we only need to choose the **Robust estimator** in the **Estimation** sub-menu.
In the clustered data setting, we need to use Generalized Estimating Equations and specify a unique cluster identifier (cluster) as **Subject variables** in the **Repeated** sub-menu.

Mount Tochal is at an elevation of 3,933 m, in the Alborz mountain range in northern Tehran, Iran.
(*photo by: M. H. Azizi MD, August 2015*)