Duke GLOBAL HEALTH INSTITUTE | Research Design & Analysis Core

Core Guide: Fitting Regression Models with a Binary Outcome

Part of a series addressing common issues in statistical and epidemiological design and analysis

Purpose

When the purpose is inference, and your goal is to maximize the precision and minimize the bias of your primary effect (treatment, exposure, etc.).

Logistic (logit link) or log-risk/log-binomial (log link) regression are the most common GLM to fit to a binary outcome. A linear risk/linear probability (identity link) model can also be used to estimate the risk difference; however, this is somewhat less common. For associations that can be assumed to be causal, this risk difference can be interpreted as the 'attributable risk' of the exposure on the outcome.

Deciding between a logit or log-risk model

1. If the data have been generated from a case-control study that employed 'survivor sampling' or when the underlying distribution of the exposure in the source population cannot otherwise be estimated, then a log-risk model cannot be used, because the risk ratio cannot be estimated. Instead, logistic regression should be used to estimate the odds ratio.
2. For a cross-sectional study, log-risk regression can be used but the resulting effect parameter should be termed the 'prevalence ratio' to clarify that the estimate of risk is from cross-sectional data, and therefore, not truly a 'risk'.
3. For a cohort study, the most common parameter estimate is the risk ratio, and therefore, a log-risk model is generally preferred to a logistic model.
4. When there is a choice, we generally recommend use of the risk ratio over the odds ratio. While both are appropriate measures, risk is more interpretable and intuitive than the odds. However, the odds may be preferred for several reasons, one of which may simply be the choice of some journals/editors.
5. Another option for modeling a binary outcome with a log link is the modified Poisson regression approach, as described by Zou[1]. In this method, the statistician simply fits the model using Poisson regression with a robust error variance. In a simulation study, Knol et al.[2] showed that the modified Poisson approach and log-binomial (log-risk) approach yielded correct risk ratios and confidence intervals. Thus, the modified Poisson approach may be used if the log-binomial model does not converge.

Considerations when deciding between a logit or log-binomial model.

Many studies have argued that RCTs and cohort studies should generally not use the logit link and instead use a link that allows direct estimation of the risk difference or risk ratio[1-3]. Often in the literature, odds are interpreted as risks. However, with increasing prevalence of exposure in a population comes increasing disparities between the odds ratio and risk ratio, with the odds ratio being pulled much farther from unity[1,2,4]. Not only is this so, but odds are much less intuitive than risks, and the very people who use the results of research in practice, such as clinicians, often mistake odds for risks[4].

In addition, some high profile studies using odds ratios have been misinterpreted by the authors, the media, and the public. For example, in a study published in 1999, the authors found that black women had an odds ratio of 0.4 compared to white men in being referred to cardiac catheterization, and conclude that black women are 60% less likely to be referred to the "best line of care", thus indicating discrimination in medical referral practices[5]. However, Schwartz et al point out several flaws in their interpretation (and the subsequent media interpretation)[3]. The authors of the original study interpreted the odds ratio as a risk ratio. When the actual risk ratio is computed, it is 0.87! The large discrepancy is due to the high prevalence of referral (~90%) in this population.

Steps in developing your model.

The below methods generally apply to both logit and log-link (and linear risk) GLM types.

1. Develop a clear hypothetical association between your primary exposure variable (E), and your outcome (O), as well as any confounding variables (C, measured and unmeasured), mediating variables (M), and effect modifiers (EM).
2. Draw a causal diagram to best represent this network of associations. Ultimately, **your model should be heavily grounded in theory**, and should not simply be a fishing expedition for possible variables with small p-values.
3. To assess the possibility of confounding, your covariate (C) needs to be associated with (E) and causally associated with (O).
    a. First, summarize the raw data in contingency tables, figures, and other tables to visually assess the relationships between (E), (O) and (C). For example, if these variables are dichotomous, produce 2 2x2 tables showing the relationship between each of (E) and (C) and (O) and (C). Note: this assumes that you have already established that there is a crude association between (E) and (O). Then consider calculating a Chi-squared statistic (or other similar test) to formally assess these relationships. The p-value from this Chi-squared statistic will give you evidence for a crude association. You could also forgo this step and move straight into univariable modeling if you prefer. For example, a logistic regression model of (C) on (E) will give you similar results as a Chi-squared test of the same relationship. For the modeling approach, the parameter estimate and the p-value should be used together to assess whether the variables are crudely associated. Variables from these analyses that show an association with the outcome, either as primary exposure or confounders should then be considered for inclusion in a multivariable model. For continuous variables (E and C), consider using x/y plots of your

data, correlation coefficients, tables and t-tests to crudely assess the data. Inclusion of possible mediators and effect modifiers are beyond the scope of this Guide and are discussed separately.

4. Before fitting a multivariable model, go back to your casual diagram, expert knowledge of the subject matter, and common sense, and make sure that each of your variables are appropriate to include in the same model. For example, are any two of your covariates highly correlated ($r > 0.6$)? If so, then it may not be appropriate to fit these two variables in the same model. In this case, fit separate models for each of the two collinear variables, or choose one of the variables because, *a priori*, it's more important to your work. Furthermore, based on your casual diagram, does any one variable act as a collider between (E) and (O)? If so, then do NOT include this variable in your model. Or test by including and excluding this variable and see how the parameter estimate of your primary exposure is affected.
5. Fit a saturated model, including all covariates that proved significant in crude analysis (and using your best judgment and guidance in step #4 above). Write down or otherwise save the log-likelihood (LL) of the model, and the parameter estimate and standard error for (E). Based on these values, calculate the Mean Squared Error (MSE) for (E) of this fully saturated model (*see resource guide on MSE for details*).
6. Look at the parameter estimates (and to a much lesser degree, the p-value) of all covariates in the model in Step #5. Any parameters that have very little effect (e.g., RR ~1.0) should be considered for removal from the model. This requires your best judgment as there are no firm rules on this. For example, if a parameter has a p-value of 0.2 but the RR=2.5, you may want to keep this variable. The only general rule is that if a parameter estimate for (E) changes by more than 10% by removing a covariate, then consider keeping this covariate. Removal of a single variable, in general, should (hopefully only slightly) increase the bias of (E) but should also increase the precision of (E). The MSE approach is attractive because it finds the model that maximizes the precision and minimizes the bias of (E).
7. **Note**: we do not recommend using automated procedures in Stata, SAS or other software to fit final models. These procedures are based only on the smallest p-values of parameter estimates, which is a simplistic and often incorrect way of finding your best model.
8. Once you have fit several different models based on Step #6, then compare the MSE of each model. The model with the smallest MSE value may be your final model.
9. You can also perform a Likelihood Ratio Test (LRT) between two **nested** models. The LRT assesses the overall model fit and is not directly measuring the precision and bias of an estimate. In general, the more covariates you add, the better the fit. As a result, you would often choose the larger model if you use the LRT method. As a result, the Akaike Information Criterion (AIC), or other 'IC' values can be used instead of the LRT, because they penalize a model with more covariates. The rule of thumb is that a model with an AIC value of 2 or more points **lower** than another model is a better fit (and more parsimonious) and should be considered over the other model. However, for both the LRT and AIC methods, the number of observations for each model **must be exactly the same** or the models cannot be comparable.

The Chi-squared statistic and p-value from the LRT can then be assessed to determine the better fitting model.

10. In general we recommend using both the AIC and the MSE methods to determine your final model. Oftentimes these results will agree.
11. In summary, model fitting is difficult, nuanced, and requires practice. There are few rules of thumb or set algorithms for finding your 'final' model. It relies on your best judgment and should be firmly rooted in *a priori* theory of the causal relationships you are testing.

Prepared by: Joseph Egger, PhD and John Gallis, MSc
Reviewed by: Duke Global Health Institute | Research Design & Analysis Core
Version: 1.1, last updated September 19, 2017

References

1. Zou G. A modified poisson regression approach to prospective studies with binary data. American journal of epidemiology 2004;159:702-6.
2. Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP, Groenwold RH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. Canadian Medical Association Journal 2012;184:895-9.
3. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. The New England journal of medicine 1999;341:279.
4. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. Journal of clinical epidemiology 1994;47:881-9.
5. Schulman KA, Berlin JA, Harless W, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. New England Journal of Medicine 1999;340:618-26.

Appendix.

Example STATA Code for fitting the models

Log-binomial (log-risk) regression

```
glm Y X Z, family(binomial) link(log) eform
```

Modified Poisson regression

```
glm Y X Z, family(Poisson) link(log) vce(robust) eform
```

Linear risk (Linear probability) model

```
glm Y X Z, family(Gaussian) link(identity)
```

Note: OLS should not be used to fit linear risk models.

Example SAS Code for fitting the models

Log-binomial (log-risk) regression

```
proc genmod data=dataset descending;
   model Y=X Z / dist=bin link=log;
   estimate 'X label' X 1 /exp;
```

Modified Poisson regression

```
proc genmod data=dataset;
   class id;
   model Y=X Z / dist=poisson link=log;
   repeated subject=id/type=ind;
   estimate 'X label' X 1 /exp;
```

Linear risk (Linear probability) model

```
proc genmod descending;
   model Y=X Z / dist=bin link=identity;
```

Note: OLS should not be used to fit linear risk models.