Relative Risk Calculations in R

Robert E. Wheeler
ECHIP, Inc.

October 20, 2009

Abstract

This package computes relative risks for both prospective and retrospective samples using GLM and Logit-Log transformations. It will always produce relative risk estimates, although it may fail sometimes to produce confidence intervals.

1 Introduction

Calculating relative risks with GLM is full of misery. Too often an attractive set of data will be regurgitated by estimating software, with words that can be understood only by a high priest of numericity. This package attempts to ease these difficulties, by providing some automated calculations that will always produce estimates, and usually confidence intervals. The appropriate model for relative risk is the log model, shown in equation (2). In this model, the exponentials of the coefficients are relative risks. See section (5), for details. The workhouse program `glm()` is used to estimate relative risk for the log model by using it in a variety of ways. It is used first with the Logit-Log starting values described in section (5), which closely resemble the correct values. If this fails, `glm()` is run without starting values. If this fails, the data is modified, as described in section (3), and `glm()` is run both with and without starting values. If this fails, the Logit-Log estimates are reported together with confidence intervals obtained by a jackknife.

This paper describes the methodologies used in this package, and illustrates its functionality with a few examples.

2 Simple Example

It's always good to start with a simple example. The data is the Berkeley graduate admissions data set, available as a standard data set in R.

The R command to analyze this data is as follows.

```
> library(RelativeRisk)
> data(gradData)
> aa <- est.rr(Count | Admitted ~ Dept * Male, gradData, indexed = TRUE)
> aa$table

                   rr 2.5 % 97.5 %     or
(Intercept)     0.070 0.048  0.104  0.076
Dept(1)        11.709 7.884 17.388 61.871
Dept(2)         9.662 6.037 15.462 28.068
Dept(3)         4.840 3.239  7.232  6.824
Dept(4)         4.963 3.295  7.477  7.091
Dept(5)         3.398 2.224  5.194  4.152
Male            0.838 0.479  1.466  0.828
Dept(1):Male    0.899 0.509  1.587  0.422
Dept(2):Male    1.106 0.593  2.065  0.969
Dept(3):Male    1.293 0.718  2.329  1.369
Dept(4):Male    1.130 0.625  2.044  1.113
Dept(5):Male    1.384 0.737  2.599  1.476
```

This output displays the ratio of two proportions in the `rr` column. The first five rows show the proportion of admissions for the various departments with respect to the admission proportion for Department 6. Such proportion ratios are of course "relative risks," which are always less extreme than "odds ratios," as is illustrated by comparing the first and last columns[1]. The seventh row shows that the proportion of admissions for males is only 84% of that for females. The other rows suggest Simpson's paradox, since the admissions of males for most departments is higher than that for females, yet the overall admission of males is less.

3 convergence

Convergence is not always assured for GLM with the log link: a major problem with Newton like methods is the presence of zeros in the data. The easiest way to understand the difficulty is by contrasting the logistic and log models. The logistic model assumes that a probability p is related to the environmental variable vector x by the model:

$$logit(p) = \log(\frac{p}{1-p}) = \beta_0 + \beta' x, \tag{1}$$

while the log model assumes

$$\log(p) = \beta_0 + \beta' x. \tag{2}$$

If l is the binomial likelihood, then estimates are obtained by setting its derivative equal to zero:

$$\frac{\partial l}{\partial \beta_r} = \sum_i \frac{\partial p_i}{\partial \beta_r} \frac{(y_i - m_i p_i)}{p_i(1-p_i)} = 0, \tag{3}$$

[1] The `Dept(4):Male` row is a numerical aberration

where y_i is the binomial response with index m_i.

For the logistic model, $\frac{\partial p_i}{\partial \beta_r} = x_i p_i(1 - p_i)$, so that equation (3) becomes

$$\frac{\partial l}{\partial \beta_r} = \sum_i (y_i - m_i p_i) x_i = 0,$$

while for the log mode, $\frac{\partial p_i}{\partial \beta_r} = x_i p_i$, equation (3) becomes

$$\frac{\partial l}{\partial \beta_r} = \sum_i \frac{(y_i - m_i p_i)}{(1 - p_i)} x_i = 0.$$

If p_i is near unity, the ith summand in the likelihood summation becomes large, which places undue emphasis on it and makes convergence difficult. Estimation for the logistic model suffers from no such problem.

There have been a number of solutions offered for this problem. They are summarized in Lumley et al. [2006]. The solution used in this package has the flavor of a continuity correction, which is achieved by inflating the counts and adding 1 where there are zeros. In particular, if the m_i are all unity, then the Bernoulli response y_i is replaced by a two column matrix with entries $[(y_i K + (1 - y_i)), y_i + K(1 - y_i)]$, where the columns denote "reaction" and "no reaction," and K is some large number: the default is 1000. If m_i is not unity, the binomial data will require two columns, and the counts in both columns are multiplied by K and zeros are replaced by unities. This modification does not change the scale of the estimates, but it inflates the log likelihood and deflates the variances by K. The estimates converge to the correct values as K increases, but for any finite K, the estimates are slightly biased.

4 Retrospective Data

There is less difference between the analyses of prospective and retrospective data than is generally supposed. Prospective data, of course, is data collected on a fixed sets of subjects, say those treated and untreated, and the response is the proportion reacting. Retrospective data is collected for subjects who have reacted and the response is the proportion who fall in the treated or untreated categories. When the reaction is infrequent, the odds ratio and the relative risk are approximately equal, but when the reaction is not rare, these statistics differ. The odds ratio is not affected by the retrospective sample sizes, which is not true for the crude estimate of relative risk that may be calculated from retrospective samples. This has led many to prefer the odds ratio as a summary statistic. A better estimate of relative risk may be obtained by taking into account the marginal frequency of the reaction in the population, and this estimate is not affected by the retrospective sample sizes.

[Breslow and Day, 1980, p203]have given an argument leading to an estimate of relative risk conditional on the sampled individuals. A more interesting argument in the same pattern may be obtained by considering the marginal frequencies of the reaction in the population. For retrospective sampling, suppose that the “cases” are a random sample from the population of reactors and that the marginal probability of reactors is ϕ, and suppose that the “controls” are also randomly selected from the remainder of the population, and that marginal probability is $1-\phi$. If $P(X|C)$ and $P(X|\tilde{C})$ are the conditional probabilities of the environmental vector X, given case and control, respectively, then it follows via a Bayes argument using $P(X|C)\phi = pP(X)$, that

$$p = \frac{P(X|C)}{P(X|C) + P(X|\tilde{C})\theta},$$

where $\theta = (1-\phi)/\phi$. If n_C and $n_{\tilde{C}}$ are the case and control samples sizes and a and b counts for X, then $P(X|C)$ is estimated by a/n_C, and $P(X|\tilde{C})$ by $b/n_{\tilde{C}}$, and one has the estimate

$$\hat{p} = \frac{a}{a + b\theta n_r},$$

where $n_r = n_C/n_{\tilde{C}}$. This is clearly unaffected by scaling – increasing $n_{\tilde{C}}$ also increases b.

It follows from the above that retrospective data may be analyzed by multiplying the control samples by θn_r.

An interesting thing about this estimate is the relative unimportance of θ. If θ is large, as it is when the reaction is rare, then relative risk and odds ratios are essentially the same. If θ is small, say in the 2 to 10 range, corresponding to ϕ's from 1/3 to 1/11, then the relative risk estimates are very close to the crude estimates from the retrospective data.

5 Logit-Log translations

It is possible to translate the coefficients of the logistic and log models from one to the other by setting all variables but one to zero and equating the expressions for p. To distinguish the coefficients in the two models, rewrite equation (2) as

$$\log(p) = \alpha_0 + \alpha' x. \tag{4}$$

The exponential of the coefficients in this model are relative risks when the variables are coded (0,1), and the exponentials of the coefficients in equation (1) are odds ratios for the same coding. This may be seen by subtracting the models differing in the levels of one variable only. When the variables are coded (ρ_l, ρ_u), the odds ratio is $\exp(\rho\beta_i)$ and the relative risk is $\exp(\rho_u\alpha_i)$, where $\rho = \rho_u - \rho_l$.

Setting all variables but one to zero and equating the models gives:

$$\exp(\rho\alpha_i) = \exp(\rho\beta_i)\frac{1+\exp(\beta_0+\rho_l\beta_i)}{1+\exp(\beta_0+\rho_u\beta_i)}, \tag{5}$$

and the reverse translation:

$$\exp(\rho\beta_i) = \exp(\rho\alpha_i)\frac{1+\exp(\alpha_0+\rho_l\alpha_i)}{1+\exp(\alpha_0+\rho_u\alpha_i)}. \tag{6}$$

These are not mathematical inverses. Equation (5) is not the reciprocal of equation (6), which can lead to small numerical differences when the output of one is input into the other. For example, `to.rr()` is the implementation of equation (5) and `to.rr()` is the implementation of equation (6), and using these with the default variable limits of $(0, 1)$ gives

```
> to.rr(to.or(c(0.3, 1.5, 2, 0.7, 0.3)))
```

```
[1] 0.3 1.5 2.0 0.7 0.3
```

while using them with the limits $(1, -1)$ gives

```
> round(to.rr(to.or(c(0.3, 1.5, 2, 0.7, 0.3), limit = c(1, -1)),
+     limit = c(1, -1)), 3)
```

```
[1] 0.300 1.502 2.016 0.699 0.286
```

The true relative risks are of course (0.3,1.5,2.0,0.7,0.3).

For retrospective data, one simply replaces the unity in the numerator and denominator by θn_r: see section (4) for the definition of these parameters.

The relative risks estimates are biased, but their biases are much smaller than those for odds ratios. These estimates will be calculable so long as is the logistic model: they are most useful for providing starting values for GLM calculations and as a backup when all else fails. The jackknife Miller [1974] works well for estimating confidence intervals even for small numbers of strata. It of course fails when removing a stratum results in a singularity.

6 Data Input

The package attempts to accommodate most types of input that are expressed in matrix form, with observations as rows and columns as variables. The response may be either a single column of $(0, 1)$ values, a pair of columns denoting "reaction" and "no reaction," or a column of counts with an auxiliary variable to indicate "reaction" and "no reaction." The columns may be external to the function or in the matrix of data input to the function. For retrospective data, two columns are required, one for "case" and one for "control." The `tally.data()` function will process the data into two columns, according to strata as defined by the values of the other variables in the data. Needless to say, variables with a great many levels will produce a great many strata.

7 Examples

The `Chocolate` dataset gives consumer preferences for chocolate additions. It has a single response, `prefer` which assumes two values, 1 for preference and 0 for not. Ten subjects were presented with eight bars made up of all possible combinations of four ingredients, and asked their preferences. The analysis is

```
> data(Chocolates)
> est.rr(prefer ~ ., Chocolates)$table
```

```
                rr 2.5 % 97.5 %    or
(Intercept) 0.034 0.006  0.212 0.050
Subject     1.853 0.439  7.830 1.000
dark        4.025 1.014 15.975 5.943
soft        0.104 0.014  0.768 0.072
nuts        2.680 0.827  8.690 3.210
```

which indicates the the subjects preferred hard, dark, nutty chocolates. The `subjects` variable is not very informative because it has 10 levels. This variable would be more informative if one could compare the individual subjects, which can be done by making it a factor. Factors use `contr.treatment()` with the last level set as the base; hence, the output shows relative risks of each level with respect to the last level.

```
> ac <- as.afactor("Subject", Chocolates)
> est.rr(prefer ~ ., ac)$table
```

```
                rr 2.5 % 97.5 %    or
(Intercept) 0.054 0.010  0.281 0.050
Subject(1)  0.451 0.071  2.864 1.000
Subject(2)  1.000 0.874  1.144 1.000
Subject(3)  0.416 0.058  2.978 1.000
Subject(4)  0.451 0.071  2.864 1.000
Subject(5)  0.437 0.066  2.907 1.000
Subject(6)  1.000 0.874  1.144 1.000
Subject(7)  0.437 0.066  2.907 1.000
Subject(8)  1.000 0.874  1.144 1.000
Subject(9)  1.000 0.874  1.144 1.000
```

```
dark        5.269 1.346 20.622 5.943
soft        0.087 0.012  0.631 0.072
nuts        3.525 1.154 10.771 3.210
```

To change the base level, `as.afactor()` may be used. Thus to make the second subject the base, redefine the levels:

```
> ac <- as.afactor("Subject", Chocolates, levOrder = c(1, 3:10,
+     2))
> est.rr(prefer ~ ., ac)$table
```

```
               rr 2.5 % 97.5 %    or
(Intercept) 0.054 0.010  0.281 0.050
Subject(1)  0.451 0.071  2.864 1.000
Subject(3)  0.416 0.058  2.978 1.000
Subject(4)  0.451 0.071  2.864 1.000
Subject(5)  0.437 0.066  2.907 1.000
Subject(6)  1.000 0.874  1.144 1.000
Subject(7)  0.437 0.066  2.907 1.000
Subject(8)  1.000 0.874  1.144 1.000
Subject(9)  1.000 0.874  1.144 1.000
Subject(10) 1.000 0.874  1.144 1.000
dark        5.269 1.346 20.622 5.943
soft        0.087 0.012  0.631 0.072
nuts        3.525 1.154 10.771 3.210
```

One can also reverse the preference variable by using the function `as.twolevel`.

```
> ac <- as.twolevel("prefer", ac, 1)
> est.rr(prefer ~ ., ac)$table
```

```
               rr 2.5 % 97.5 %     or
(Intercept) 0.864 0.739  1.009 20.108
Subject(1)  1.000 1.000  1.000  1.000
Subject(3)  0.999 0.937  1.065  1.000
Subject(4)  1.000 1.000  1.000  1.000
Subject(5)  0.999 0.950  1.051  1.000
Subject(6)  1.000 1.000  1.000  1.000
Subject(7)  0.999 0.950  1.051  1.000
Subject(8)  1.000 1.000  1.000  1.000
Subject(9)  1.000 1.000  1.000  1.000
Subject(10) 1.000 1.000  1.000  1.000
dark        0.891 0.771  1.030  0.168
soft        1.158 0.991  1.353 13.846
nuts        0.981 0.904  1.064  0.312
```

which shows the relative risks for not prefer.

More commonly the response will have two columns. The columns may contain counts, as in the `simData` dataset, where the columns represent “success” and “failure.” One analysis of this dataset is

```
> data(simData)
> est.rr(Success | Failure ~ ., data = simData)$table
```

```
               rr 2.5 % 97.5 %    or
(Intercept) 0.728 0.666  0.796 2.917
X1          1.094 1.031  1.161 1.721
X2          1.235 1.140  1.337 2.806
X3          0.925 0.874  0.980 0.743
X4          0.645 0.575  0.725 0.193
```

Another analysis of this dataset might take into account that it is a case-control sample, with 600 cases and 400 controls. In the general population, 31% fall into the case category, thus the appropriate parameter for analysis is `theta=2.2`.

```
> est.rr(Success | Failure ~ ., data = simData, theta = 2.2)$table
```

```
               rr 2.5 % 97.5 %    or
(Intercept) 0.298 0.257  0.345 2.917
X1          1.448 1.295  1.619 1.721
X2          2.026 1.780  2.305 2.806
X3          0.718 0.645  0.800 0.743
X4          0.311 0.265  0.366 0.193
```

which, among other things, suggests that there is less difference that ordinarily supposed between prospective and retrospective analyses for case-control data – the case-control sample size ratio is 600/400 or 150% as opposed to 31% for the population, which are substantially different, and yet the conclusions about significance that would be drawn from the relative risk estimates are the same in the two analyses. This dataset is the output of a simulation, in which the true relative risks were `(0.3,1.5,2,.7,.3)`.

Another type of two column response, contains counts in one column and an indicator variable in the second column. The indicator variable is used to divide the counts into “success” and “failure” columns. Counts of the survivors of the Titanic shipwreck illustrate this point. Here the variable `survived` is used to index the counts in the `count` column. The fact that this second column is an index column is signaled by setting the parameter `indexed` to `TRUE`.

```
> data(TitanicMat)
> aa <- est.rr(count | survived ~ ., indexed = TRUE, data = TitanicMat)
> aa$log
```

```
  [,1]
1 "The data was tallied."
2 "2  rows were removed because response had all zeros."
3 "GLM successful."

> aa$table

               rr 2.5 % 97.5 %    or
(Intercept) 0.876 0.768  0.999 9.466
class(1st)  1.141 1.001  1.302 2.358
class(2nd)  1.020 0.885  1.176 0.852
class(3rd)  0.782 0.683  0.896 0.398
sex(Male)   0.419 0.377  0.466 0.089
age(Adult)  0.872 0.824  0.923 0.346
```

The log shows the progress of the calculations. In this case, the GLM calculation was successful, even thought two rows of the data were eliminated. The results indicate that the probability of survival for males was about half that of females, and adults about 87% of children. The probability of survival for first and second class passengers was greater than that of the crew.

One can reverse the relative risk values by changing the order of the levels for a factor. The “sex” factor levels in the above analysis are `("Male","Female")`. To change the order use `as.afactor()` as follows:

```
> TitanicMatR <- as.afactor("sex", TitanicMat, levOrder = c("Female",
+     "Male"))
> est.rr(count | survived ~ ., indexed = TRUE, data = TitanicMatR)$table

            LL-rr     or
(Intercept) 0.457  0.842
class(1st)  1.455  2.358
class(2nd)  0.914  0.852
class(3rd)  0.550  0.398
sex(Female) 1.979 11.247
age(Adult)  0.493  0.346
```

Because of two missing rows, GLM was unable to solve the equations with the reversed levels, and the jackknife was unable to calculate confidence intervals, so the Logit-Log estimates are shown. The “sex” factor was, however, reversed as desired.

References

N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research: Volume 1 - The Analysis of Case-Control Studies.* Number 32. International Agency for Research on Cancer, Lyon, 1980.

T. Lumley, R. Kronmal, and M. Shuangge. Relative risk regression in medical research: Models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper*, 293: 1–24, 2006.

R.G. Miller. The jackknife - a review. *Biometrika*, 62(1):1–15, 1974.