Model-based Estimation of Relative Risk

Søren Lophaven
Background

• A clinical trial to determine whether nitrofurazone-impregnated urinary catheters reduce the incidence of urinary tract infections in patients that were admitted directly from the accident scene to the Trauma Center in Copenhagen

• The primary endpoint was the proportion of patients developing an urinary tract infection after surgery

• Analyzed using a logistic regression model with treatment group and gender as explanatory factors

• Reviewer’s comment: "We ask that you use log-binomial models because events were common, so ORs from a logistic regression model will overstate risk estimates"
1. Introduction
2. Model-based estimation of relative risk
3. Simulation experiments
4. Implementation in SAS and R
5. An algorithm for estimating relative risk
6. Points for discussion
Probability and odds

- Probability is intuitive (range from 0 to 1)
- Odds are not intuitive (range from 0 to ∞)
- The probability of the event occurring divided by the probability of the event not occurring, i.e. probability/(1 − probability)
- Probability of an event occurring is 0.05, then the odds of that event is 0.05/(1 − 0.05) = 0.053 = 1/19
- Probability of an event occurring is 0.5, then the odds of that event is 0.5/(1 − 0.5) = 1
Probability and odds
Odds ratio and relative risk

- Probability of outcome is 0.66 in one group and 0.33 in the second group
- Treatment effect: The first group is twice as likely to have the outcome as the second group
  \[ \text{RR} = \frac{p_1}{p_2} = 2 \]
- Odds ratio = the ratio of two odds
  \[ \text{OR} = \frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} = \frac{p_1 \cdot (1 - p_2)}{p_2 \cdot (1 - p_1)} = \frac{0.66 \cdot (1 - 0.33)}{0.33 \cdot (1 - 0.66)} = 4 \]
- Does not mean that the outcome is 4 times as likely in the first group as in the second group
- Odds of the outcome is 4 times higher in the first group than the second group
Why is OR so frequently reported?

- Estimated by logistic regression which employs the so-called canonical link for binary outcome data

\[ \text{logit}(\rho) = \log(\text{odds}) = \log \left( \frac{\rho}{1 - \rho} \right) \]

- Natural choice for binary outcome data = Mathematically prettier

- Logistic regression always produce probabilities in the range 0 to 1

- Odds have an unlimited range, and any positive odds ratio will still yield a valid probability (RR=2 can only apply to probabilities below 0.5)

- Convergence problems are rare
What is the problem?

• Relative risk has a natural interpretation. The odds ratio has not
• Many examples of researchers misinterpreting the odds ratio as a relative risk
• Holcomb et al. (2001) report that 26% of authors in top-tier medical journals explicitly misinterpret odds ratios as though they were relative risks
• Thomas Lumley: *Most people don’t have any feel for the meaning of an odds ratio (and those that do mostly can’t communicate it to those who don’t)*
• If the outcome is common (> 10% with event) the odds ratio becomes a poor approximation of the relative risk
• In this case the odds ratio overestimates the relative risk
OR versus RR

The relationship between risk ratio (RR) and odds ratio by incidence of the outcome.
Primary care physicians were shown videotaped interviews with an actor portraying a patient with chest pain, and given data on cardiac risk factors and results of thallium stress test. There were 18 scenarios, and each was portrayed by 8 actors (race $\times$ sex $\times$ age combinations). The outcome was whether the doctors recommended referral to cardiac catheterization.

<table>
<thead>
<tr>
<th></th>
<th>Odds ratio 95 % CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td>0.6 (0.4-0.9)</td>
</tr>
<tr>
<td>Men</td>
<td>1</td>
</tr>
<tr>
<td>Women</td>
<td>0.6 (0.4-0.9)</td>
</tr>
</tbody>
</table>

So the odds of referral were 40% lower for women than for men and for blacks than for whites.
Example (Schulman et al., 1999)

**Nightline**  "In our main analysis we found that blacks were 40% less likely to be referred for cardiac catheterization compared to whites"

**USA Today**  "Heart Care Reflects Race and Sex, not Symptoms"

**Washington Post**  "Doctors are far less likely to recommend sophisticated cardiac tests for blacks and women than for white men with identical complaints"

**LA Times**  "Authors suggest the differences are the consequences of race and sex bias"

**NY Times**  "Doctors are only 60% as likely to order cardiac catheterization for women and blacks as for men and whites"
Example (Schulman et al., 1999)

• After many readers pointed out that a relative risk of 0.93 isn’t a 40 % reduction the New England Journal of Medicine and many of the news sources carried a revision of the story. The original article is still being cited as evidence of widespread and serious racial bias in medicine.

• This was a particularly public and embarrassing example, but the underlying problem is much more common.
Model-based estimation of relative risk
Early Relative Risk Estimation - Mantel-Haenszel

- Calculating the relative risk from a single 2 x 2 table is a trivial task
- The Mantel-Haenszel method averaged estimates across strata
- Given I strata each contributing a 2 x 2 table, i.e.

<table>
<thead>
<tr>
<th></th>
<th>(1) Exposed</th>
<th>(0) Unexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Disease</td>
<td>$n_{i11}$</td>
<td>$n_{i10}$</td>
<td>$n_{i1}$</td>
</tr>
<tr>
<td>(0) No disease</td>
<td>$n_{i01}$</td>
<td>$n_{i00}$</td>
<td>$n_{i0}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{i,1}$</td>
<td>$n_{i,0}$</td>
<td>$n_{i}$</td>
</tr>
</tbody>
</table>

- The log relative risk can be estimated by

$$
\log \hat{RR} = \frac{\sum_{i=1}^{I} w_i \log \hat{RR}_i}{\sum_{i=1}^{I} w_i}, \quad w_i = \frac{1}{\text{Var}(\log \hat{RR}_i)} = \left[ \frac{n_{i01}}{n_{i11} \cdot n_{i,1}} + \frac{n_{i00}}{n_{i10} \cdot n_{i,0}} \right]^{-1}
$$

- Unable to deal with continuous explanatory variables
The log-binomial model

- Relative risks arise naturally from the regression model
  \[ \log P[Y = 1 | X] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \]

- If \( P[Y = 1 | X] \) is small then
  \[ \log P[Y = 1 | X] \approx \log \frac{P[Y = 1 | X]}{1 - P[Y = 1 | X]} = \text{logit} P[Y = 1 | X] \]

- Partial derivatives of the binomial likelihood for the logistic model is
  \[ \frac{\partial l}{\partial \beta_r} = \sum_i (y_i - m_i p_i) x_i = 0 \]

- Partial derivatives of the binomial likelihood for the log mode is
  \[ \frac{\partial l}{\partial \beta_r} = \sum_i \frac{(y_i - m_i p_i)}{1 - p_i} x_i = 0 \]

- \( p_i \) near unity \( \Rightarrow \) \( i \)th summand is large \( \Rightarrow \) convergence becomes difficult
The log-binomial model

• Unlike the logistic regression model, the log-binomial model requires constraints on $\beta$ to ensure that fitted probabilities remain in the interval $[0, 1]$

• All model-based relative risk estimation flows from this model

• Other models for relative risk estimation should only be used if problems with the log-binomial model are identified, e.g. failed convergence or invalid fitted probabilities

• The methods below can be seen as ways of approximating the maximum likelihood estimate originating from the log-binomial model
Logit-Log translations

• Consider

\[
\text{logit}(p) = \log \left( \frac{p}{1-p} \right) = \beta_0 + \beta' x
\]

and

\[
\log(p) = \alpha_0 + \alpha' x
\]

• Coefficients of the logistic and log-binomial model can be translated from one to the other by setting all variables but one to zero and equating the expressions for \( p \)

• Relative risk obtained from the results of a logistic regression:

\[
\exp(\alpha_i) = \exp(\beta_i) \frac{1 + \exp(\beta_0)}{1 + \exp(\beta_0 + \beta_i)}
\]
The COPY method

• Suggested by Deddens and Peterson (2003)
• Computes maximum likelihood estimates from a new expanded data set that contains $c - 1$ copies of the original data and one copy of the original data with the outcome values interchanged (1’s changed to 0’s and 0’s changed to 1’s)
• Lumley et al. (2006) pointed out that this is equivalent to creating a new data set which consists of one copy of the original data set with weight $w = (c - 1)/c$ and one copy of the original data set with the outcome values interchanged with weight $1 - w = 1/c$, and then performing a weighted log-binomial regression
• In many simulations the method converges on the modified data set with standard software
• Precise reasons for the successful convergence by the COPY method have not been shown
Modifying data

- Suggested by Wheeler (2009)
- Modify data by replacing $y_i$ with
  \[ y_iK + (1 - y_i); y_i + K(1 - y_i) \], $K = 1000$
- Does not change the scale of the parameter estimates
- It inflates the log likelihood
- It deflates the variances by $K$
- The estimates converge to the correct values as $K$ increases
- For any finite $K$, the estimates are slightly biased
- The method has the flavor of a continuity correction
Poisson regression

• Can be used to estimate relative risks from binary data.
• Poisson regression gives standard errors that are too large, because the variance of a Poisson random variable is always larger than that of a binary variable with the same mean.
• The sandwich estimator was suggested by Zou (2004) and Carter et al. (2005) for removing this bias.
Cox regression

• Cox regression usually used for time-to-event data
• Construct an artificial dataset where every subject has the same observation time and all events occur at the same time
• Tied event times should be dealt with, e.g. as suggested by Breslow (1974)
• Then, the hazard ratio estimated by Cox regression approximates the relative risk
• When every individual has the same artificial observation time this approximation results in the same estimating equations as Poisson regression, i.e. the same parameter estimates and the same upwardly-biased standard errors are calculated
Poisson/Cox regression

- No guarantee that the fitted probabilities will remain in the allowable range
- No warning given by standard software if fitted probabilities outside the allowable range
Simulation experiments
Data-generation process

• In accordance with Savu et al. (2010)

• Model with four independent explanatory variables: treatment ($E$) is from a 0/1 bernoulli distribution with equal probabilities, $X_1$ is from a 0/1 bernoulli distribution with equal probabilities, $X_2$ is from a three-category multinomial distribution with values 0, 1 and 2 and corresponding probabilities 0.3, 0.3 and 0.4, and $X_3$ is from a uniform distribution on the interval (-1,2)

• Outcomes for non-exposed subjects ($E=0$):

\[
\log P( Y_1 = 1 | E = 0, x_1, x_2, x_3 ) = -2.1 - x_1 - x_2(1) - x_2(2) - x_3
\]

\[
\logit P( Y_3 = 1 | E = 0, x_1, x_2, x_3 ) = -1.7 - x_1 - x_2(1) - x_2(2) - x_3
\]

• Outcomes for exposed subjects ($E=1$):

\[
P( Y_k = 1 | E = 1, x_1, x_2, x_3 ) = 3 \cdot P( Y_k = 0 | E = 0, x_1, x_2, x_3 )
\]
Data-generation process

- Intercepts selected by requiring that 0.95 is the largest value of \( P( Y_k = 1 | E = 0, x_1, x_2, x_3 ) \)
- Thereby \( P( Y_k = 1 | E, x_1, x_2, x_3 ) \leq 1 \)
Investigation of convergence

- Investigated convergence for the log-binomial model with and without starting values obtained by Logit-Log translations, Poisson regression, the COPY method, and the modification of data algorithm with and without starting values obtained by Logit-Log translations.
- Investigated convergence for different sample sizes.
- 1000 simulations.
Investigation of convergence

- Convergence problems not caused by small sample sizes as suggested by Scott (2008)
- Rate of non-convergence consistent with literature
Investigation of convergence

Data model = 1

Data model = 3

Overall event rate (%)
## Investigation of performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Data model 1</th>
<th>Data model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log($RR$)</td>
<td>Relative bias (%)</td>
</tr>
<tr>
<td>Log-binomial</td>
<td>1.1446</td>
<td>4.19 %</td>
</tr>
<tr>
<td>Poisson</td>
<td>1.10396</td>
<td>0.49 %</td>
</tr>
<tr>
<td>COPY method</td>
<td>1.1240</td>
<td>2.31 %</td>
</tr>
<tr>
<td>Modifying data</td>
<td>1.12169</td>
<td>2.10 %</td>
</tr>
<tr>
<td>Logit-Log translations</td>
<td>1.09064</td>
<td>-0.73 %</td>
</tr>
</tbody>
</table>
Implementation in SAS and R
Log-Binomial Model:

\[
\text{glm}(y \sim \text{treatment} + x_1 + x_2 + x_3, \text{family}=\text{binomial}(\text{link}="\text{log}"), \text{data} = \text{data}, \text{start}=\text{start})
\]

Logistic regression:

\[
\text{glm}(y \sim \text{treatment} + x_1 + x_2 + x_3, \text{family}=\text{binomial}(\text{link}="\text{logit}"), \text{data} = \text{data})
\]

Poisson Regression:

\[
\text{library(sandwich)}
\text{model.output<-glm(}y \sim \text{treatment} + x_1 + x_2 + x_3, \text{family}=\text{poisson}(\text{link}="\text{log}"), \text{data} = \text{data})
\text{sandwich(model.output)}
\]
COPY method:

```r
weight <- 0.999
original <- cbind(data,weight)
weight <- 0.001
new <- cbind(data,weight)
new$y <- (1-new$y)
combined <- c(original,new)
glm(y~treatment+x1+x2+x3, family(binomial(link="log"), data = combined,
weights = weight))

or:

library(geepack)
relRisk(y~treatment+x1+x2+x3, id=id, data=data)
```
Modifying data:

\[
\begin{align*}
y_{ValT} &= y_{ValC} \times \text{modConstant} + (1 - y_{ValC}) \\
y_{ValC} &= \text{cbind}(y_{ValT}, y_{ValC} + \text{modConstant} \times (1 - y_{ValC})) \\
\text{glm}(y_{ValC} \sim \text{treatment} + x1 + x2 + x3, \text{family}(\text{binomial}(\text{link}="log")), \text{data}=\text{data}, \text{start}=\text{start})
\end{align*}
\]

or:

\[
\begin{align*}
\text{library}(\text{RelativeRisk}) \\
\text{est.rr}(y \sim \text{treatment} + x1 + x2 + x3, \text{data}=\text{data}, \text{start}=\text{start})
\end{align*}
\]
Log-Binomial Model:

```sas
proc genmod data=data;
  class treatment x1 x2;
  model y=treatment x1 x2 x3 / dist=bin link=log;
  estimate 'RR Treatment' treatment -1 1/ exp;
run;
```

Logistic regression:

```sas
proc genmod data=data;
  class treatment x1 x2;
  model y=treatment x1 x2 x3 / dist=bin link=logit;
  estimate 'RR Treatment' treatment -1 1/ exp;
run;
```

Poisson Regression:

```sas
proc genmod data=data;
  class id treatment x1 x2;
  model y=treatment x1 x2 x3 / dist=poisson link=log;
  repeated subject=ID/type=ind;
```
COPY method:

```
data original;
  set data;
  weight=0.999;
run;

data new;
  set data;
  y=1-y;
  weight=0.001;
run;

data combined;
  set original new;
run;

proc genmod descending;
  model y=treatment x1 x2 x3 / dist=binomial link=log;
  weight=weight;
run;
```
An algorithm for estimating relative risk
An algorithm for estimating relative risk

As suggested by Wheeler (2009):

1. Run a log-binomial model with starting values obtained from the Logit-Log translations
2. If this fails then run a log-binomial model without starting values
3. If this fails, the data is modified, and the log-binomial model is run both with and without starting values on the modified data
4. If this fails, the estimates obtained from Logit-Log translations are reported
Points for discussion
Points for discussion

• Do we use logistic regression too much?
• Do we communicate the meaning of odds ratios correctly?
• Should we estimate relative risks in clinical trials?
• If ’yes’, should it be the primary or secondary analysis?

What about removing explanatory variables from the model?


