Poisson Regression: Let me count the uses!

Lisa Kaltenbach

November 20, 2008
Poisson Reg may analyze

1. Count data (ex. no. of surgical site infections)
2. Binary data (ex. received vaccine(yes/no) as alternative to logistic reg
3. Time-to-event data (ex. time-to-stroke with time dependent covar) as alternative to survival analysis
Outline

• Binary Data
  • Motivating ex.
  • Background
  • RR vs OR: Common Misunderstandings
  • How to estimate

• Poisson Reg
  • Binomial and Poisson Dist
  • Simple Poisson Reg for 2x2 Tables

• Time-to-event data
  • Background
  • Motivating ex.
  • Simulation ex. by Patrick Arbogast
  • How it works
Movitating ex: Non-rare event

- E+, E-: Exposed, Unexposed (placebo group)
- D+, D-: Disease, Non-disease

<table>
<thead>
<tr>
<th></th>
<th>E+</th>
<th>E-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+</td>
<td>112</td>
<td>24</td>
<td>136</td>
</tr>
<tr>
<td>D-</td>
<td>28</td>
<td>36</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>140</td>
<td>60</td>
<td>200</td>
</tr>
</tbody>
</table>

Risk: \( \frac{112}{140} = 0.8 \) \( \frac{24}{60} = 0.4 \)

RR: \( \frac{0.8}{0.4} = 2 \)

Odds: \( \frac{112}{28} = 4 \) \( \frac{24}{36} = 0.667 \)

OR: \( \frac{112 \times 36}{28 \times 24} = 6 \)
Binary Data

- Logistic reg is most popular method for binary data (with fixed follow-up)
- Logistic reg estimates odds ratios adjusted for covariates
- \[ \log \frac{Pr(Y=1|X)}{1-Pr(Y=1|X)} = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p \]
- If rare event \( Pr(Y = 1) < 10\% \), then odds ratio (OR) \( \approx \) relative risk (RR)
- Logistic reg is also used to model common events, where RR could be directly estimated and is NOT close to OR
RR is preferable to OR

- Extensive discussion in lit reached conclusion (Greenland 1987, Sinclair 1994, Nurminen 1995)
- If event rate is not low < 10% then conversion of OR to RR will provide invalid CIs and produce inconsistent estimates for RR
- Difficulty of explaining the correct interpretation of odds ratios
- Despite rare event assumption OR are often interpreted as RR
OR misinterpreted as RR

- Schulman et al. (1999) claimed that the "race and sex of a patient independently influence how physicians manage chest pain."
- Several major US news media (including ABC’s Nightline) overstated the effects
- 7% reduction in referral rate for Black women was mistakenly reported at 40% (Schwartz 1999)
**Table 1. Rate of Referral for Cardiac Catheterization, Odds of Referral, Odds Ratio, and Risk Ratio According to Sex and Race.**

<table>
<thead>
<tr>
<th>Patients</th>
<th>Mean Referral Rate</th>
<th>Odds of Referral</th>
<th>Odds Ratio (95% CI)</th>
<th>Risk Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four strata</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White men†</td>
<td>90.6</td>
<td>9.6 to 1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Black men</td>
<td>90.6</td>
<td>9.6 to 1</td>
<td>1.0 (0.5–2.1)</td>
<td></td>
</tr>
<tr>
<td>White women</td>
<td>90.6</td>
<td>9.6 to 1</td>
<td>1.0 (0.5–2.1)</td>
<td></td>
</tr>
<tr>
<td>Black women</td>
<td>78.8</td>
<td>3.7 to 1</td>
<td>0.4 (0.2–0.7)</td>
<td>0.87 (0.80–0.95)</td>
</tr>
<tr>
<td>Aggregate data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White†</td>
<td>90.6</td>
<td>9.6 to 1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>84.7</td>
<td>5.5 to 1</td>
<td>0.6 (0.4–0.9)</td>
<td>0.93 (0.89–0.99)</td>
</tr>
<tr>
<td>Men†</td>
<td>90.6</td>
<td>9.6 to 1</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>84.7</td>
<td>5.5 to 1</td>
<td>0.6 (0.4–0.9)</td>
<td>0.93 (0.89–0.99)</td>
</tr>
<tr>
<td>Overall</td>
<td>87.7</td>
<td>7.1 to 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Referral rates for the four strata were inferred from aggregate rates and odds ratios reported by Schulman et al.1 The odds of referral were calculated according to the following formula: referral rate ÷ (100% – referral rate). The risk ratio was calculated as the referral rate for the group in question divided by the referral rate for the reference group. CI denotes confidence interval.
Solution: Estimate RR

- Poisson reg with robust error variance to estimate RR directly
- Recall: RR cannot be directly estimated in case-control studies and ORs in cohort and case-control studies are identical.
- Simple 2x2 table justifies the validity of this approach
- Simulations show reliable even with N=100
- Assume that time-to-event info is either unavailable or inappropriate to answer study questions
Hypothetical dataset: UCLA Stat website

Outcome: lenses, indicator for use of corrective lenses by age 30

Assume all participants enter study at age 10 w/out lenses

Want to know if lenses is assoc with having a gene which causes craving for carrots (assume not having this gene results in the opposite), and that we screened everyone for this carrot gene at baseline (carrot = 1 if they have it, = 0 if not)

All values were assigned using a random number generator
Estimators of RR

- Show estimators that give consistent results and valid SEs can be seen as a series of robust generalizations of the MLE
- Lumley argues that MLE is insufficiently robust to model misspecification
- Neither reasonable nor necessary to assume that the binary var has a Poisson dist
Estimators of RR 2

- Log-linear model is natural choice, since it estimates incidence rate ratio and most medical applications of Poisson dist arise via Poisson approx to binomial dist
- Estimating eq are unbiased when the response var is binary rather than Poisson, and thus give a consistent estimation of the RR
- When used to estimate RR from binary, Poisson reg gives SEs that are too big, because the variance of Poisson random var is always larger than that of a binary variable with the same mean
  - Ex: If $\mu = .5$ then $\sigma_{\text{poisson}} = .5$ but $\sigma_{\text{binomial}} = .5 \times .5 = .25$
- Remove bias by using model-robust SE estimates
Summary: RR reg

- RR naturally arise from reg model
- \( \log(\Pr(Y = 1|X)) = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p \)
- \( \exp \beta_i \) is RR contrasting levels of \( X_i \) that differ by 1
- If \( \Pr(Y = 1|X) \) is small, then
  \[
  \log(\Pr(Y = 1|X)) \approx \log \frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)} \equiv \text{logit}(\Pr(Y = 1|X))
  \]
- But if \( \Pr(Y = 1|X) > 1\% \), then
  - \( \beta_{\text{logistic}} \) and \( \beta_{\text{RR}} \) will differ materially
  - \( |\beta_{\text{logistic}}| > |\beta_{\text{RR}}| \)
- Unlike logistic model, RR model has constraint on \( \beta \) to ensure fitted probabilities remain in \([0,1]\)
- If RR is of interest, then calculate it directly using Poisson reg with robust error variance
Ex: no. of times students place a beer can on statue during 1 yr follows a Poisson dist with parameter $\lambda$
Binominal and Poisson Dist

• Binomial Dist
  • Suppose that $n$ unrelated events are observed among $N$ patients.
  • Let $p$ be the probability that any patient has an event.
  • Let $K$ be a random variable for the number of events among $N$ patients.
  • Then $K$ is dist. as Bin($N,p$)
  • $Pr(K = n) = \frac{N!}{(N-n)!n!} p^n(1-p)^{N-n}$

• Poisson Dist
  • The probability of $n$ events occurring in a time period $t$ for a Poisson random variable with parameter $\lambda$ is
  • $Pr(X = n) = \frac{(\lambda t)^n \exp(-\lambda t)}{n!}, \ n=0,1,2,\ldots$
  • Where $\lambda$ is the expected number of events per time unit
  • Poisson showed that when $N$ is large and $p$ is small the distribution of $n$ is approximately a Poisson distribution.
**Terminology**

<table>
<thead>
<tr>
<th></th>
<th>E+</th>
<th>E-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td>$d_1$</td>
<td>$d_0$</td>
<td>$d_1 + d_0$</td>
</tr>
<tr>
<td><strong>Person-time</strong></td>
<td>$n_1$</td>
<td>$n_0$</td>
<td>$n_1 + n_0$</td>
</tr>
<tr>
<td><strong>Incidence Rate</strong></td>
<td>$\frac{d_1}{n_1}$</td>
<td>$\frac{d_0}{n_0}$</td>
<td>$\frac{d_1+d_0}{n_1+n_0}$</td>
</tr>
<tr>
<td><strong>Incidence Rate Ratio</strong></td>
<td>$\frac{d_1/n_1}{d_0/n_0}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 95 percent CI for IRR is $IRR^{exp(\pm 1.96*s)}$
- Where $s^2 = \frac{1}{d_1} + \frac{1}{d_0}$
- Suppose $\pi_i$ is true event rate, if $\pi_i$ is small then assume $d_i$ follows a Poisson dist
Simple Poisson Reg for 2x2 Tables

- $R = \frac{\pi_1}{\pi_0}$ is the relative risk of event assoc. with E+
- $\hat{\pi} = \frac{d_i}{n_i}$ is the estimated event rate for group i
- $E(d_i|x_i) = n_i\pi_i$
- $E(\hat{\pi}_i|x_i) = E\left(\frac{d_i}{n_i}|x_0\right) = \frac{E(d_i|x_i)}{n_i} = \pi_i$
- Then $\log(\pi_0) = \log(E(d_0|x_0)) - \log(n_0)$
  $\log(\pi_1) = \log(E(d_1|x_1)) - \log(n_1)$
- Thus $\log(R) = \log(\pi_1) - \log(\pi_0)$
- Let $\alpha = \log(\pi_0)$ and $\beta = \log(R)$
- Then $\log(E(d_i|x_i)) = \log(n_i) + \alpha + \beta x_i$
Poisson Reg is a GLM

- Recall any GLM
  - Random component
  - Linear predictor
  - Link function
- Poisson
  - Random component: $d_i$ the no. of events in ith group of $n_i$ patients or patient-years
  - Linear predictor: $\log(n_i) + \alpha + \beta x_i$
  - Link function: log
Using Poisson for time-to-event data
TN Medicaid Study: NSAIDs and risk of CHD

- Large dataset (about 3G) 5 million records, 200 var
- Time varying covariates
- Time-to-first event
- Want to study many outcomes
  - Composite outcomes:
    - AMI or CHD or Stroke
    - AMI or Stroke
    - Stroke or CHD
    - AMI or Stroke
    - Stroke or CHD
  - Each outcome alone
- Cox Proportional Hazards model is so computationally intensive that it may not converge or may take a day to fit each model!
Solution: Use Poisson instead

- For grouped survival data, if the event rate per unit of time is low then fit Poisson reg models and save run time
- Typically we fit both Poisson reg and Cox reg for primary outcome to show that our results are similar (and answer reviewers concerns)
- Same data format works for both models
- Useful if data arrive in events per person-yr format (census data) instead of individuals
Published Papers using this Approach

Current Project: ADHD Study

- **5 Sites** (Medicaid: TN and WA, Kaiser: North and South, Ingenix)
- Retrospective Cohort Study
- ADHD medications and risk of serious cardiovascular disease in children and youth
- Endpoints: sudden cardiac death, serious coronary disease, serious cardiovascular disease, MI, Stroke
- Estimated event rate 0.0001
- Estimated exposure prevalence 2-4%
- Need lots of data to detect an association
- Analysis Plan: Use Poisson reg models to estimate IRR
- Patrick Arbogast is lead statistician designed simulation to show that Poisson Reg is a reasonable alternative to time-dependent Cox Reg
Simulation comparing Poisson Reg to time-dependent Cox Reg

- N=100,000 patients
- E+ incidence rate of 0.0002; E- incidence rate of 0.0001 (ie, IRR=2)
- Censoring: .25
- Patients randomly assigned to one of four exposure patterns over time:
  - Unexposed for entire follow-up
  - Exposed for entire follow-up
  - Unexposed initially, then exposed for remaining follow-up
  - Exposed initially, then unexposed for remaining follow-up
- Each exposure pattern had equal probability of occurring
- Survival time simulated assuming an exponential distribution
- Fit Poisson and time-dependent Cox regression
## Simulation Results

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta_E$</th>
<th>$\exp(\beta_E)$</th>
<th>p-value</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>0.7447 ± 0.0900</td>
<td>2.106</td>
<td>$\leq .0001$</td>
<td>1.96 sec</td>
</tr>
<tr>
<td>Cox</td>
<td>0.74446 ± 0.09005</td>
<td>2.105</td>
<td>$\leq .0001$</td>
<td>2 min</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>26.12 sec</td>
</tr>
</tbody>
</table>
Grouped Survival Data

- Grouped failure time data consist of the total number of failures (occurrence) and the total time at risk (exposure) for each covariate pattern.

- Reasons for grouping:
  - Large sample sizes (epi), population groups
  - To economize on data transmission and storage, to reduce computation, to protect privacy of individual records, or to account for limitations of a measurement instrument
  - Some datasets are publicly released in grouped form only
  - Difficult to obtain exact lifetimes, because ethical, physical, or economic restrictions in research design allow subjects in follow-up to be monitored only periodically

- Grouped data can involve censoring (rgt, left, or double) and truncated data

- Can include covariates
Life Tables

- Oldest and most commonly used methods of presenting lifetime data
- Useful for presenting, summarizing, and estimating survival function, the pdf and the hazard function along with the variance of these estimators
- Logrank test used to compare survival probabilities by exposure (or covar strata)

<table>
<thead>
<tr>
<th>Failure</th>
<th>E+</th>
<th>E-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>$d_{1k}$</td>
<td>$d_{0k}$</td>
<td>$d_k$</td>
</tr>
<tr>
<td>No</td>
<td>$n_{1k} - d_{1k}$</td>
<td>$n_{0k} - d_{0k}$</td>
<td>$n_k - d_k$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{1k}$</td>
<td>$n_{0k}$</td>
<td>$n_k$</td>
</tr>
</tbody>
</table>
Why Poisson works for time-to-event

- Many authors give the MLE for grouped data
- Natural estimate of unknown hazard rate $\lambda_{rk}$ is $\hat{\lambda}_{rk} = \frac{d_{rk}}{Y_{rk}}$
  occurrence/exposure rate
- The grouped data based hazard estimator can be obtained by maximizing an approximation to Cox's partial likelihood Laird and Oliver (1981) showed that this estimator can be interpreted as the MLE in a Poisson reg model
Conclusions

- Poisson reg is useful for more than just count data
- Calculate (adjusted) RR for rare and non-rare events with Poisson reg with robust SE
- Computationally fast and efficient for analyzing grouped survival data
• Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *Am J Epidemiol* 2004;
Thanks!

Patrick Arbogast
Nate Mercaldo
Tebeb Gebretsadik