Generalized Logistic Regression

Randy Johnson

3/2/2017

Generalized Linear Regression

GLM

Generalized Linear Models (GLMs) allow us to relax the multivariate normality assumption.

Assumptions:

- Model fit is correct
- No/little multi-collinearity
- No overly influential variables

```
args(glm)
```

```
## function (formula, family = gaussian, data, weights, subset,
##     na.action, start = NULL, etastart, mustart, offset, control = list(...),
##     model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts = NULL,
##     ...)
## NULL
```

Linear Regression

- Response: Continuous
- Family: gaussian(link = 'identity')
- Interpretation:
  - β_0 is the mean response in the reference group (Intercept).
  - $\beta_{1...n}$ are the mean differences comparing those exposed to X with the reference group.
- Caveat: Same as if we used `lm()`, so the multivariate normality assumption should hold.

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Log-linear Regression

- Response: Continuous, Time to event, Count
- Family: poisson(link = 'log')
- Interpretation:
  - β_0 is the log mean response in the reference group.
  - $\beta_{1...n}$ are the log ratios of the means comparing those exposed to X with the reference group.
- Caveat: `mean(y)` is assumed to be equal to `var(y)` – use quasipoisson(link = 'log') in the event that this assumption is not valid.

$$\log \mathrm{E}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Logistic Regression

- Response: Binary
- Family: binomial(link = 'logit')
- Interpretation:
  - β_0 is the log odds for the reference group (cohort study only).
  - $\beta_{1...n}$ are the log odds ratios (ORs) comparing those exposed to X with the reference group. These are sometimes referred to as LOD scores.
- Caveats:
  - β_0 doesn't have any real-world interpretation for case-control studies.
  - The OR will slightly over estimate the Relative Risk Ratio. For rare disease this difference is negligible.

$$\log odds(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

Odds vs Relative Risk

Given A equal to the number of events and B equal to the number of non-events,

$$odds(Y|X) = \frac{\mathrm{P}(Y|X)}{1-\mathrm{P}(Y|X)}$$
$$= \frac{\frac{A}{\cancel{A+B}}}{\frac{B}{\cancel{A+B}}}$$
$$= \frac{A}{B}$$

$$RR(Y|X) = \mathrm{P}(Y|X)$$
$$= \frac{A}{A+B}$$

Thus one of the major advantages of modeling the odds is that you don't need to know the prevalence or incidence in the population. This is why we use Odds Ratios in case control studies instead of RR Ratios.

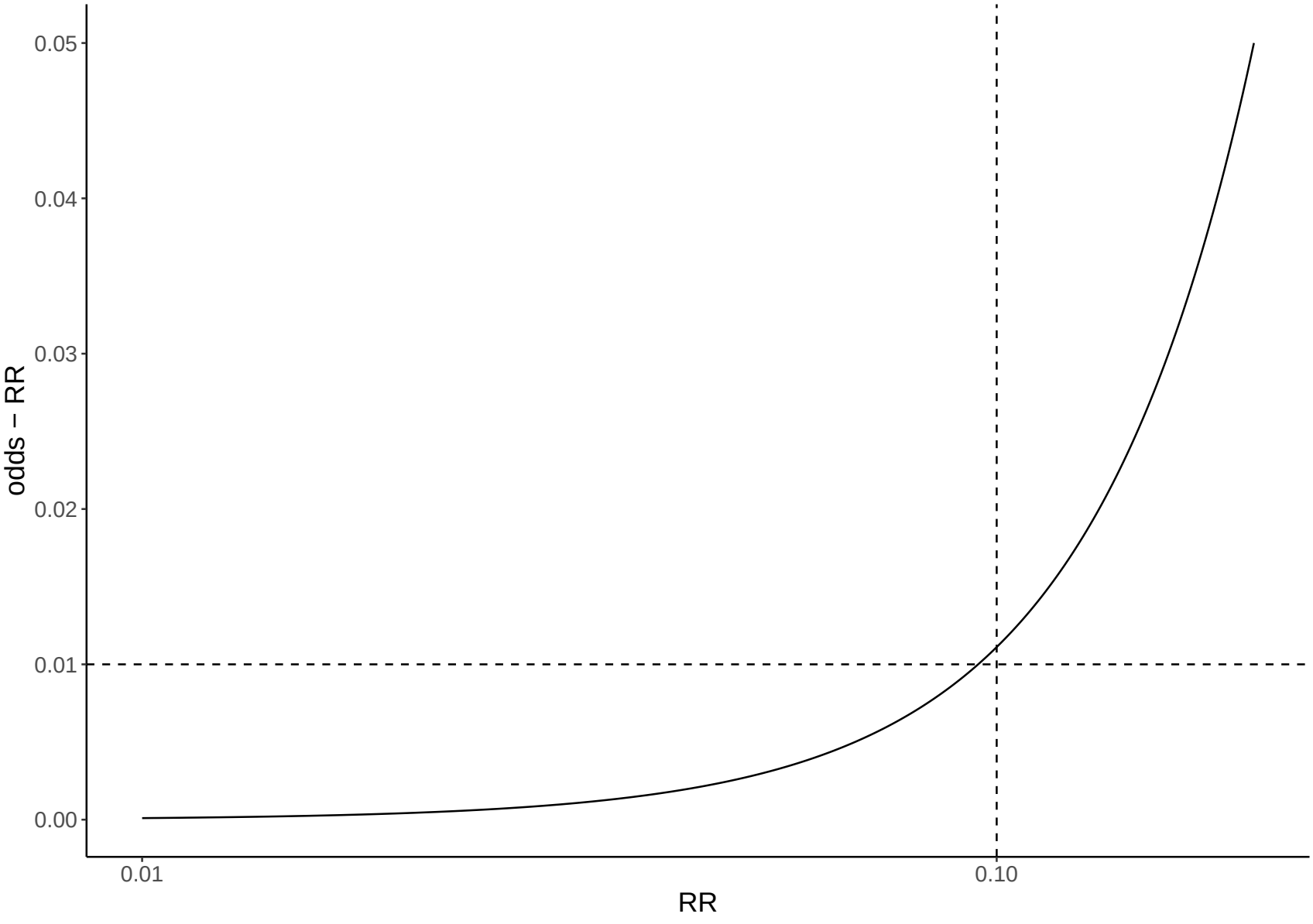
Odds vs Relative Risk

The odds will always overestimate the relative risk,

$$
\begin{aligned}
odds(Y|X) &> RR(Y|X) \\
\frac{A}{B} &> \frac{A}{A+B} \\
\frac{A}{B} &\approx \frac{A}{A+B}
\end{aligned}
$$

but they will be approximately equal if the number of events, A, is small relative to B (i.e. when the event is rare). How rare does the event need to be?

Odds vs Relative Risk



Log-binomial Regression

- Response: Binary
- Family: binomial(link = 'log')
- Interpretation:
  - β_0 is the log incidence/prevalence for the reference group.
  - $\beta_{1...n}$ are the log incidence/prevalence ratios comparing those exposed to X with the reference group.
- Caveat:
  - This model only makes sense in the context of cohort studies.
  - It tends to have more stability issues than logistic regression.

$$\log \mathrm{P}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$