



Choosing a regression for dichotomous outcomes

Ken Kleinman

Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute

Modeling dichotomous outcomes

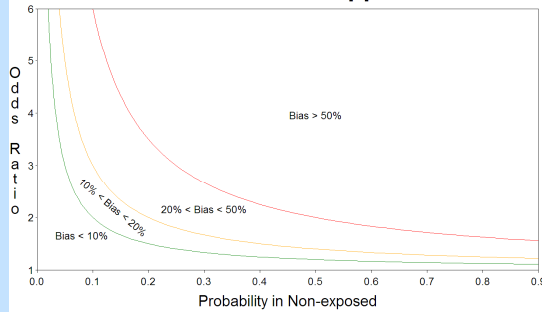
- Logistic regression is a default tool for most data analysts, epidemiologists, and statisticians
- Many useful features
- Results in an odds ratio
- But the odds itself is a ratio, so the odds ratio is a ratio of a ratio! Hard to interpret.**

$$\log \left[\frac{\Pr(y=1)}{\Pr(y=0)} \right] = \sum xB \quad \text{so} \quad \Pr(y=1) = \frac{e^{B_0 + x_1 B_1 + \dots + x_k B_k}}{1 + e^{B_0 + x_1 B_1 + \dots + x_k B_k}}$$

• e^{B_1} is the odds ratio for x_1 $e^{B_1} = \frac{\text{Odds}(y|x=1)}{\text{Odds}(y|x=0)}$

It's OK— Odds ratio is like risk ratio?

Percent bias in OR as an approximate RR



Odds ratios are really awkward

What does an odds ratio of 2 mean?

- If $\Pr(y=1|x=0) = .1$ (rare), then odds = $.1/.9 = .11$
- So $\Pr(y=1|x=1) = .1818$: odds = $.1818/.8181 = .22$

X\Y	0	1
0	90	10
1	82	18

OR=2 Note: RR = $.18/.1 = 1.8$

But...

- If $\Pr(y=1|x=0) = .6$ (common) then odds = $.6/.4 = 1.5$
- So $\Pr(y=1|x=1) = .75$: odds = $.75/.25 = 3$

X\Y	0	1
0	40	60
1	25	75

OR=2 Note: RR = $.75/.25 = 1.25$

- An odds ratio in a vacuum isn't very informative
- When the baseline probability ($\Pr(y=1|x=0)$) changes, the risk ratio changes, if the OR is constant.

Instead, "convert" an OR to a RR?

"Correction factor" (Zhang and Yu, JAMA 1998) estimates the RR from a logistic regression.

But if there are additional covariates, each value for them changes the baseline probability. The baseline probability is now $\Pr(y=1|x_1=0, x_2=?)$: different for each value of x_2 .

For example, suppose $\Pr(y=1) = \frac{e^{B_0 + x_1 B_1 + x_2 B_2}}{1 + e^{B_0 + x_1 B_1 + x_2 B_2}}$

and $B_0=1, B_2=4$

- Then $P(y=1|x_1=0, x_2=0) = e/(1+e) = .73$
- But $P(y=1|x_1=0, x_2=1) = e^5/(1+e^5) = .99$

But when the baseline probability changes, the risk ratio changes, if the OR is constant!

And in a multivariate logistic regression, of course, the OR is constant.

In other words, if the logistic regression is the right thing to do, seeking a single risk ratio is the wrong thing to do!

Better: Generalized Linear Models

- Modern statistics allow us to fit models such as $\log(\Pr(y=1)) = B_0 + x_1 B_1 + \dots + x_k B_k$ while retaining the binomial distribution.

- The "log-binomial" model
- A risk ratio of 2 is impossible if $\Pr(y=1|x=0) = .75$. ($\Pr(y=1|x=1) = 1.5$) This model will often fail to converge, for common outcomes.
- In these cases, you can fit a Poisson model
- Either Poisson or binomial result in $RR = e^{B_1}$
- We can also fit the "linear-binomial" model

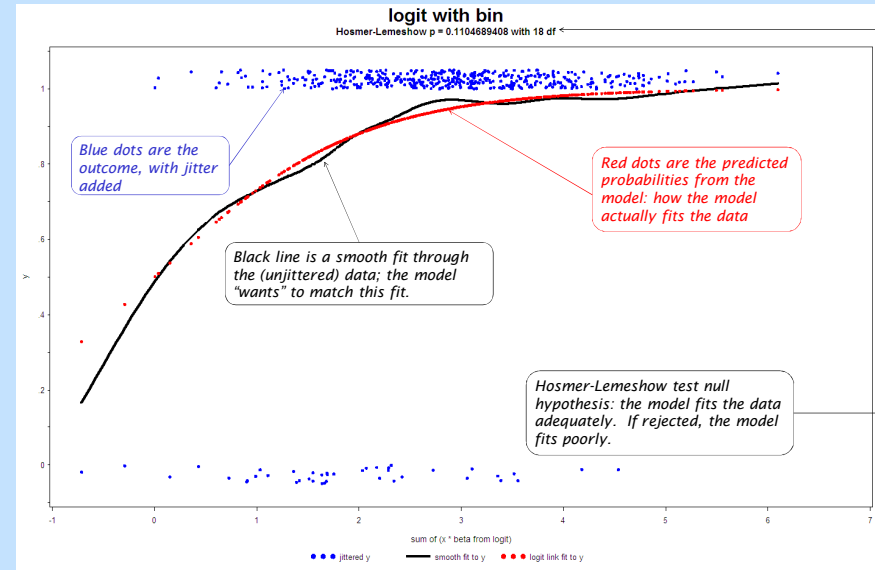
$$\Pr(y=1) = B_0 + x_1 B_1 + \dots + x_k B_k$$

where B_1 estimates the difference in probability. But we shouldn't choose a model because we like the meaning of the parameter estimates.

An adequate fit is a pre-requisite to interpretation. Chances are that some models fit better than others. Choose the model that best describes the data, first.

Choosing a good model

I propose a visual diagnostic for assessing the fit of the model.



Using the diagnostic

- In practice, use this diagnostic to compare logistic, log, linear models
- SAS Macro (R version also) does this
 - Uses Poisson if binomial doesn't fit
 - Shows Hosmer-Lemeshow test
 - Ken_Kleinman@hms.harvard.edu
 - Simulated data are logistic: fit of log and linear models are both poor.

Recommendations

- Use this diagnostic
- Models with a large range of predicted probabilities tend to be best fit by logistic
- If more than one fits well, choose by interpretation
- For all models, discuss some predicted probabilities, to give context

