

Logistic regression: Why we often can do what we think we can do

Maarten Buis

19th UK Stata Users Group meeting, 10 Sept. 2015

Introduction

- In 2010 Carina Mood published an overview article called **“Logistic regression: Why we cannot do what we think we can do, and what we can do about it”**
- Her main conclusions were:
 1. It is problematic to interpret odds ratios as substantive effects, because they also reflect unobserved heterogeneity.
 2. It is problematic to compare odds ratios across models with different independent variables, because the unobserved heterogeneity is likely to vary across models.
 3. It is problematic to compare odds ratios across groups, because the unobserved heterogeneity can vary across the compared groups.
- This article had a big impact in sociology making many researchers unsure whether logistic regression and odds ratios have any use at all.

History

- The article by Mood is a review article, so the points she made are not new:
 - There was an active debate in the bio-medical sciences (e.g. Gail et al. 1984).
 - The issue is known in economics (e.g. Lee 1982), but does not play a big role as odds ratios are not popular there anyhow.
 - Some early work did happen in sociology (e.g. McKelvey and Zavoina 1975), but never reached a large audience.

The first indication

- Many studies start with the observation that if one adds a variable to a logistic regression model the remaining coefficients will change **even if** the added variable is uncorrelated with the other variables.
- The easiest explanation of this phenomenon starts with the latent variable representation of logistic regression.
- This assumes that there is a latent propensity for experiencing a 'success'.
- One experiences the success if the propensity passes a threshold (0).
- The propensity of success is a linear function of the explanatory variables plus an error term

$$y_i^* = x_i\beta + \varepsilon_i$$

- So, the probability of success is:

$$\begin{aligned} \Pr(y_i = 1|x) &= \Pr(y_i^* > 0) \\ &= \Pr(\varepsilon_i > -x_i\beta) \\ &= \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)} \quad \text{if } \varepsilon \sim \lambda\left(0, \frac{\pi}{\sqrt{3}}\right) \end{aligned}$$

The scale of the latent variable

- Notice, that I needed to fix the standard deviation of the error term
- The reason is that y^* is latent, so its scale is unknown
- The scale of the latent variable is fixed by fixing the standard deviation of the error term
- $\text{var}(y^*) = \text{var}(x_i\beta) + \text{var}(\varepsilon)$
- What happens when we add a variable to our model?
- That extra variable is 'removed' from the error term, so the variance of the error term decreases (assuming that the variable is uncorrelated with the error term).
- But the scale of the dependent variable was defined by fixing the scale of the residual.
- So the scale of the dependent variable depends on which variables are in the model.

Mood's three problems

- problem 1** The scale of the dependent variable depends on which variables are in the model, making it problematic to interpret the resulting coefficients 'the effect'.
- problem 2** The scale of the dependent variable changes when one adds variables, making comparison of effects problematic.
- problem 3** Say we want to compare groups and the residual variances is likely to differ across groups (e.g. comparing effects on the probability of some labor market outcome between men and women)
- problem 3** In that case the scale of the dependent variable differs, making the comparison of effects problematic.

Is the scale of the dependent variable really unidentified?

- There is a different way of looking at logistic regression that does not involve a latent dependent variable
- In that view logistic regression is a linear model for the log odds of success.
- An odds is just an alternative way to quantify how likely a success is:
- Instead of considering the expected proportion of successes (probability) you look at the expected number of successes per failure (odds)
- So an odds can take any value larger or equal to zero
- The log of the odds can take any value
- The scale of the log odds is known and does not change when adding or removing variables or comparing groups.
- However, this does not solve everything as the coefficients still change when we add or remove uncorrelated variables.

previous attempts at solving the problem

- use a linear probability model and approximations thereof (average marginal effects) as these don't react to uncorrelated additional variables. (`regress` and `margins`)
- standardize the dependent variable (`listcoef` in the `spost` package)
- estimate the variance of the residual (`oglm`)

My attempt

- current state: I had an intuition for a long time and finally taken time to write a it down in a working-paper. However, a more formal representation would improve that paper.
- I start with the question: Is there is a problem that needs solving?
- One can think of logistic regression as trying to model the proportion of successes
- The proportion is a characteristic of a group.
- One can turn it into an individual level characteristic by saying that it is a probability
- A probability is an assessment of how likely we think that the event happens
- If we add (relevant) information this assessment should change
- Adding information in case of logistic regression means adding variables.
- So probabilities only exist within the context of a specific model.
- For example, adding a variable does not lead to an improved estimate of the probability, but leads to an estimate of a different probability.

How should effect react to new information?

- If we become surer because of the new information our probabilities can become
 - closer to 0 if we become surer that the event will not happen
 - closer to 1 if we become surer that the event will happen
- So after adding new information there is more room for a variable to have an effect and the effects should increase
- The more relevant the new information is, the larger the increase
- The odds ratios from logistic regression show exactly this behavior
- So the odds ratios can be a meaningful effect size.
- However, one needs to specify which variables were in the model, but that makes sense in this interpretation of the dependent variable.
- Linear probability models and Average Marginal Effects have been proposed as 'solutions' because they don't change when adding non-confounding variables.
- According to my argument, this would actually make them problematic.

Comparing groups

- Say we compare the effect on a labor market outcome between men and women.
- The labor market experiences for men tend to be more predictable than that of women.
- So we are probably surer in the group of men
- So the predicted probabilities can be closer to 0 and 1 for men compared to women
- and there is thus more room for a variable to have an effect
- The odds ratios from logistic regression show exactly this behavior
- So a comparison of odds ratios across groups provides an accurate **description** of the difference in effect.

Causality in group comparisons

- What would happen to the effect if we turn a man into a woman (without the stigma usually associated with such a change)?
- The uncertainty the probability measures comes from many variables we did not observe
- The sum of the effects of each of these variables is the error term, and the variance of that error term captures the amount of uncertainty
- The variance could differ across groups
 - because the variances of the unobserved variables differ and/or
 - because their effects differ.
- If the difference in residual variance is only due to a difference in effects of the unobserved variables,
- then a man draws his unobserved variables from the same distribution as the women, only their effects differ.
- So the effects in the female sub-sample represents the counter-factual effects (assuming all other conditions are met).

Comparing across models

- Say we want to explain who enters university and we have parental status and previous school performance
- part of the effect of parental status is due to the fact that
 - high status children tend to do well in school and
 - those that do well are more likely to enter university
- An attempt to quantify this indirect of parental status through school performance is often done by comparing
 - the effect of parental status in a model with only parental status with
 - the effect of parental status in a model with parental status and previous school performance
- This won't work in logistic regression, as the effect of parental status changes not just because of the correlation between parental status and school performance but also because of the extra certainty obtained from adding school performance.
- We can use `knb` or `ldecomp` for these problems.





Conclusions

1. The odds ratio is a meaningful effect-size. The fact that it is dependent on which variables are included in the model is not a problem but actually a requirement for an effect on a probability.
2. It is indeed problematic to compare coefficients across models with different sets of explanatory variables, since effects on probabilities are supposed to change when variables are added to the model even if they are uncorrelated with the other explanatory variables.
3. A comparison of odds ratios across groups provides an accurate description of the difference in effects across these groups, and under special circumstances can also be given a causal interpretation.
 - Points where I am still uncertain are:
 - How does this argument work when the added information from a new variable is inconsistent with what we knew before?
 - In my argument the linear probability model is not appropriate, but it is usually more helpful to say what it does measure.
 - My hope is that a more formal representation of this argument will help solve those questions.

Thank you!

- This is work in progress, so I welcome all questions and comments
- There is a working paper available from
<http://www.maartenbuis.nl/wp/oddsratio.html>

References

-  [M. H. Gail, S. Wieand, and S. Piantadosi](#)
Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates
Biometrika, 71(3):431–444, 1984.
-  [L.-F. Lee](#)
Specification error in multinomial logit models: Analysis of the omitted variable bias
Journal of Econometrics, 20(2):197–209, 1982.
-  [R. D. McKelvey and W. Zavoina](#)
A Statistical Model for the Analysis of Ordinal Level Dependent Variables
Journal of Mathematical Sociology, 4(1):103–120, 1975.
-  [C. Mood](#)
Logistic regression: Why we cannot do what we think we can do, and what we can do about it
European Sociological Review, 26(1):67–82, 2010.