Logistic regression: When can we do what we think we can do?

Universität Konstanz

Maarten Buis

maarten.buis@uni.kn

Carina Mood (2010) concluded:

It is problematic to interpret odds ratios as substantive effects, because they also reflect unobserved heterogeneity.
It is problematic to compare odds ratios across models with different independent variables, because the unobserved heterogeneity is likely to vary across models.
It is problematic to compare odds ratios across groups, because the unobserved heterogeneity can vary across the compared groups.

Is this the end of logistic regression, log-linear models, and the odds ratio?

1. The Problem

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.

This can be understood using the latent variable representation of logistic regression.

Assume that there is a latent propensity for experiencing a 'success', and one experiences the success if the propensity passes a threshold (0).

The scale of the latent variable is fixed by fixing the standard deviation of the error term

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.

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.

If we compare groups and the residual variance is different across groups, then the scale of the dependent variable will differ across groups.

2. The puzzle

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: it is the expected number of successes per failure.

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.

Moreover, regardless of which way we think about logistic regression, we get exactly the same parameter estimates.

How can it be that the scale of the dependent variable across groups is simultaneously the same and different?

Is there a problem that needs solving?

With logistic regression we try to model the degree of certainty that an event happens, i.e. an assessment of how likely we think that the event happens. This degree of certainty could be quantified as either an odds, a log-odds, or a probability.

Our assessment of how likely an event is should depend on the available information, which in logistic regression is captured by the variables in our model.

If we become surer our probabilities can become closer to 0 or closer to 1. 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.

However, one needs to specify which variables were in the model, but that makes sense in this interpretation of the dependent variable.

Also, chances refer to things that are as yet unknown. So they don't refer to the observations in the data, but to similar units who have not yet experienced the event.

4. Are we interested in effects on a 'degree of certainty'?

Yes, for example questions that have to do with inequality of opportunity: There are unequal opportunities when someone from a lower background is less likely to attain a high level of education or a high occupation or marry a 'desirable' partner.

In that case our interest is in the degree of certainty that someone attains the favourable position and how this degree differs between social backgrounds. This degree of certainty is more a characteristic of the society than the person.

We can compare odds ratios across societies to find out which society is more closed. In this case the interest is at the society level, not the individual level.

Yes, for example when we want to use the results to make decisions. It is these degrees of certainty that we want to use to inform our decisions, but we have to be aware of how these are affected by the available information we have and potential differences across groups.

No, for example questions that have to do with understanding individual choices: We are not interested in an external observer's assessment of how likely a choice is. These types of questions fit naturally in the latent variable representation, with all the problems that come with it.

Carina Mood's (2010) objections apply to some but not all applications of logistic regression

If we are interested in effects on a 'degree of certainty', then:

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.

It is problematic to estimate direct and indirect effects by comparing 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.

Odds ratios can be compared across groups, as that provides an accurate description of the difference in effects across these groups.

If we are interested in understanding individual choice, then:

The objections by Mood (2010) hold.

Linear probability models/average marginal effects won't answer the question of interest as they measure effects on a 'degree of certainty'.

Instead, **effects on the (standardized) latent variable get closer** to the question of interest, but has the problem that the scale is unidentified.

– maartenbuis.nl/wp/oddsratio.html