Logistic regression: Why we often can do what we think we can do[1].

Augst 8th 2015

Maarten L. Buis,

University of Konstanz,

Department of History and Sociology

maarten.buis@uni.konstanz.de

> "All propositions are true or false, but the knowledge we have to them depends on our circumstances; and while it is often convenient to speak of propositions as certain or probable, this expresses strictly a relationship in which they stand to a corpus of knowledge, actual or hypothetical, and not a characteristic of the propositions in themselves. A proposition is capable at the same time of varying degrees of this relationship, depending upon the knowledge to which it is related so that it is without significance to call a proposition probable unless we specify the knowledge to which we are relating it." (Keynes, 1921 [2004], pp. 8-9)

Abstract

In a much cited overview article Mood (2010) criticized many of the ways in which the raw coefficients and odds ratios from logistic regression have been used. However, logistic regression has an unusual dependent variable: a probability, which measures how certain we are that an event of interest happens. This degree of certainty is a function of how much information we have, which in case of logistic regression is captured by the variables we add to the model. If the dependent variable is interpreted in that way many of the problems with logistic regression pointed out by Mood (2010) turn out to be desirable properties of the logistic regression model.

Introduction

Since the appearance of the overview article by Mood (2010) there has been a growing concern within sociology that the odds ratio can no longer be used in research. In particular Mood (2010, pp. 67-68) concluded that:

[1] I thank Michelle V. Jackson, Sebastian E. Wenz, and Richard T. Campbell for useful comments.

1. "It is problematic to interpret log-odds ratios or odds ratios as substantive effects, because they also reflect unobserved heterogeneity
2. It is problematic to compare log-odds ratios or odds ratios across models with different independent variables, because the unobserved heterogeneity is likely to vary across models.
3. It is problematic to compare log-odds ratios and odds ratios across samples, across groups within samples, or over time – even when we use models with the same independent variables – because the unobserved heterogeneity can vary across the compared samples, groups, or points in time."

In this article I will argue that the first and the last of these problems actually represent desirable properties of the logistic regression model if one interprets the dependent variable, the probability, as an assessment of how likely it is that the event of interest happens. I will start with a general description of the logistic regression model, followed by a description of the problems discussed by Mood (2010) and others (for example: Allison, 1999; Auspurg & Hinz, 2011; Gail, Wieand, & Piantadosi, 1984; Karlson, Holm, & Breen, 2012; Lee, 1982; Neuhaus & Jewell, 1993; Williams, 2009; Wooldridge, 2010). I will then discuss why many of these "problems" are not problems at all.

Logistic regression

There are two ways in which one can think about the logistic regression model (for example: Long, 1997; Maddala, 1986). The first way starts with the observation that a probability is a number between zero and one, and if we just used a linear regression on a probability we could easily end up with predictions outside that range. With logistic regression we apply the logit transformation to the probabilities, meaning that we have a linear model for the log-odds of success instead of the probability of success. This is shown in equation (1) where *y* is the binary dependent variable, *P(y=1)* is the probability that the dependent variable takes the value 1, the *x*s are the explanatory variables and the βs their effects.

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (1)$$

The odds can be interpreted as the expected number of "successes" per "failure", and the log-odds is the logarithm of that. The log-odds has no upper or lower bound, so a linear model cannot give invalid predictions. Moreover, the exponentiated coefficients can be interpreted directly as odds ratios, that is, the ratio by which the odds change per unit increase in the explanatory variable.

This can also be seen as a way of fitting an "S-shaped" effect of the explanatory variables on the probability of success, as equation (1) can be rewritten as equation (2). This way the predicted

probabilities from a logistic regression model are guaranteed to remain within the allowable range for probabilities.

$$\mathrm{P}(\mathrm{y}=1)=\frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2)} \tag{2}$$

The second way of thinking about logistic regression is to assume that there is an unobserved latent propensity to experience an event. If that latent propensity passes a threshold (typically 0) then the event will occur. What is observed is whether or not the event occurs and not the propensity. The latent variable is influenced by the explanatory variables and an error term, as in equation (3), where y^* is the latent propensity and ε is the error term.

$$y^*=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon \tag{3}$$

The probability of experiencing the event is in this model the probability that the latent propensity is larger than zero, which according to equation (3) can be written as

$$P(y=1)=P(\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon>0) \tag{4}$$

Since both the βs and the *x*s are fixed, the only uncertain part is the error term. As a consequence equation (4) can be rewritten as equation (5):

$$P(y=1)=P(\varepsilon>-[\beta_0+\beta_1 x_1+\beta_2 x_2]) \tag{5}$$

The probability that a random variable, in this case ε, is larger than some number can be derived from the cumulative distribution function. If we assume that the error term follows a logistic distribution with mean 0 and a variance equal to $\pi^2/_3$, then equation (5) will become equation (2), thus leading to logistic regression.

The problem

One characteristic of logistic regression that has troubled many authors (for example: Allison, 1999; Auspurg & Hinz, 2011; Gail et al., 1984; Hauck, Neuhaus, Kalbfleisch, & Anderson, 1991; Lee, 1982; Mood, 2010) is that if we add a variable to a logistic regression model, the effects of the other variables will change, *even if* this additional variable is uncorrelated with the other variables. This property of logistic regression is typically explained with the presence of unobserved heterogeneity.

The simplest version of this explanation starts with the latent variable representation of the logistic regression model. As discussed above, the dependent variable in that representation of logistic regression is a latent propensity to experience the event. The fact that the dependent variable is

latent means that the unit is unknown; it does not have a known scale like meters, euros or minutes. Instead the unit of that latent variable is fixed by fixing the variance of the error term. So if something changes the residual variance, then that also changes the scale of the dependent variable. One way to change the residual variance is to add a new explanatory variable to the model. That new variable was previously part of the residual. So adding that variable to the regression reduces the residual variance (assuming the explanatory variable is uncorrelated with the error term). The scale of the dependent variable thus changes by adding a variable to our model.

This is pretty damning; if the scale of the dependent variable changes then a comparison of coefficients becomes meaningless. If the scale of the dependent variable is so dependent on which control variables you add, even if those control variables are uncorrelated with the explanatory variable of interest, then how can the coefficient of a logistic regression represent the effect of a variable? This is the basis for Mood's first claim that it is problematic to interpret log-odds ratios or odds ratios as substantive effects and second claim that it is problematic to compare across models with different independent variables.

This problem also affects comparisons of logistic regression coefficients across groups. To borrow an example from Allison (1999): Say, we are estimating one logistic regression to explain some labor market outcome for men and another for women. He makes a plausible argument that women's career paths and labor market experiences tend to be more variable than those of men, leading to a residual variance that is larger for women than for men. As a consequence the scale of the dependent variable in the sample of men is different from the scale of the dependent variable in the sample of women. This is the basis for Mood's third claim that it is problematic to compare log-odds ratios or odds ratios across groups.

The biggest problem is that the scale of latent dependent variable changes when the residual variance changes. There is an easy solution to that, as there is another way of looking at logistic regression: logistic regression models linear effects of explanatory variables on the log-odds of experiencing the event. So in that interpretation the dependent variable is no longer a latent variable but the log-odds. That scale is known: the logarithm of the odds, which is the logarithm of the expected number of "successes" per "failure". This scale does not change when we add or remove variables from our model and it is the same across groups.

However, this does not solve everything: the effects, the odds ratios, will still change if you add an unconfounding variable. Adapting a numerical example by Buis (2011): Say we have two countries, A and B, and within each country we look at the effect of a binary variable *x* on a binary variable *y*, as described in Table 1. These cross tabulation make for convenient examples, as the parameters of

logistic regression model can be easily computed by hand: The odds of getting a high *y* is larger in country A compared to country B, but in both countries the odds of getting a high *y* is three times larger for high *x* individuals compared to low *x* individuals. This factor three is the odds ratio one would get in a logistic regression model. Another important feature of this example is that the distribution of *x* is the same in each country, so country is not a confounding variable.

Table 1: Hypothetical example while controlling for country

		y		N	probability	odds	Odds ratio
country	*x*	Low	High				
A	Low	200	200	400	.5	1	3
	High	100	300	400	.75	3	
B	Low	300	100	400	.25	.333	3
	High	200	200	400	.5	1	

If one were interested in the effect of *x* and had used a linear regression then it would not matter whether one controlled for country or not, as country and *x* are in this example completely uncorrelated. However, this is not true with logistic regression. Not controlling for country means collapsing Table 1 to Table 2: If we do not want to control for country then it is sufficient to know that there are 300 + 200 =500 observation with low *x* and low *y*, 200+100=300 observations with low *x* and high *y*, etc. When we do that the odds ratio drops from 3 to 2.778. The reason for that change in effect of *x* is that collapsing a table means averaging the probabilities. For example the probability of being a high *y* given a low *x* in countries A and B is respectively 200/400=.5 and 100/100=.25. After collapsing Table 1 to Table 2 the probability of being a high *y* given a low *x* is 300/800 = (.5 + .25)/2 = .375. However, with logistic regression we work with odds and odds ratios, and these are non-linear transformations of probabilities. So averaging probabilities does not correspond to averaging odds. To continue the example: (1+.333)/2=0.667≠0.6.

Table 2: Hypothetical example without controlling for country

	Y		N	probability	odds	Odds ratio
x	Low	High				
Low	500	300	800	.375	0.6	2.778
High	300	500	800	.625	1.667	

So even though the scale of the dependent variable is comparable across models and groups, the size of the effect is still dependent on which variables were added to the model even if those added variables are non-confounding variables. In particular, in the model without the extra variable the probabilities tend to be less extreme (more in the middle and further away from either one or zero), and the effects are thus smaller compared to the model with the extra variable (Neuhaus & Jewell, 1993).

A solution

Much of the effort in this area has focused thus far on removing the effect of non-confounding variables, by for example using average marginal effects or linear probability models (Auspurg & Hinz, 2011; Mood, 2010), using standardized regression coefficients (Karlson, 2015; Winship & Mare, 1984), or directly estimating the degree of heterogeneity and controlling for it (Allison, 1999; Williams, 2009). Instead I propose to take a step back and first consider what this effect of unconfounding variables means and whether it is desirable to remove it.

To do so, I will start with reconsidering the meaning of the dependent variable in a logistic regression. In a pure mechanical sense logistic regression models the proportion of successes. This proportion is necessarily a property of a (hypothetical) group and not a property of an individual; the "proportion" of successes in a single trial can only take two values zero or one, fractional proportions can only occur when it is based on multiple trials. However, in the way logistic regression is used in sociology, we often want to say something about individuals. We are often interested in why some individuals are more successful in school, get married, become unemployed, etc. One way of turning this group level proportion into an individual level probability is to view a probability as a way of quantifying someone's assessment of how likely it is that an individual will experience an event. In logistic regression this assessment is then based on the predicted proportion of people that experience the event who share the same values on the explanatory variables.

Looking at probability in this way has consequences: a probability of an event is not an absolute property of either that event or of the person doing the evaluating, but is conditional on the information available to the person doing the evaluating; if someone learns more, then her or his assessment should change. (Keynes, 1921 [2004]) The more (mutually consistent) information we have, the surer we are, the closer the probabilities should be to either zero or one.

In logistic regression the "available knowledge" translates to the set explanatory variables included in the model. So the probability is defined by the choice of control variables rather than something that exists outside the model. This links back to the scale of the latent variable being dependent on which control variables are added to the model. However, instead of this dependence being a problem, it is now a desirable property of the logistic model in the sense that the probabilities we get out of a logistic regression react to new information (new variables) as one would expect.

This way of thinking about probability not only involves an effect of extra information on probabilities, but also a change in the effects of variables on probabilities due to adding extra information. Before adding extra information all the probabilities should be close together, that is, further away from the extremes of zero and one. There is thus less room for variables to have an effect compared to after adding extra information. So not only should one expect the probabilities to change in reaction to extra information, but also effects on that probability. This means that adding unconfounding variables should increase the effect of other variables, which is exactly what happens in logistic regression.

Since a change in effects due to adding unconfounding variables is now desirable, methods and models that do not have that property become problematic. For example, both the linear probability model and average marginal effects are not influenced by unconfounding variables, and thus cannot measure effects on probabilities as conceived of above. In fact, exactly because they are not influenced by unconfounding variables they were previously put forward as "solutions" (Auspurg & Hinz, 2011; Mood, 2010). Since the problem disappeared by changing the interpretation of the dependent variable, the solutions became problems.

A comparison of effects across groups is also influenced by this interpretation of the dependent variable. If one group is more predictable than another, then that should lead to larger effects in the more predictable group. The logic is the same as with adding an unconfounding variable: In a group where one is less sure the range of possible predicted probabilities should be more restrictive (further away from zero or one) compared to a group where one is surer. As a consequence there is more room for a variable to have an effect in a group where one is surer. Consider again the example by Allison (1999) where we investigate the difference between men and women in effect of a

variable on the probability of some labor market outcome. The labor market experiences tend to more predictable for men than for women. This should make the effects of variables for men larger than those effects for women. As was discussed above the logistic regression model has exactly this characteristic. So rather than that being a problem, this is again a desirable characteristic and a comparison of odds ratios across groups will thus give an accurate description of differences in effects across these groups.

Whether this difference in odds ratio across groups has a causal interpretation is more complex. To borrow an example from (Auspurg & Hinz, 2011), say we are comparing effects in different countries, Switzerland and Germany. The causal interpretation involves a thought-experiment where we pick up a person in Germany, put her in Switzerland and see how much the effect changes just because of that move. It is now helpful to consider where the uncertainty that a probability measures comes from. We are uncertain because we have not seen everything that influences our outcome. These things could in principle have been captured by variables, but we did not do that and these variables are unobserved. The sum of the influences of these different unobserved variables is the source of our uncertainty and is the error term in equation (3). The amount of uncertainty is captured by the variance of that error term. That variance could differ across countries because the distribution of the unobserved variables is different across countries and/or the effects of the unobserved variables are different across countries.

A simple comparison of odds ratios has a causal interpretation (assuming all the other requirements for a causal effect are met) if we think that the distribution of the unobserved variables is the same across countries but the institutions differ across countries making the unobserved variables have different effects. In that case, we assume that the people in Switzerland draw their unobserved variables from the same distribution as the people in Germany, but the effects of these unobserved variables differ. As a consequence, the effects in the Swiss sample correspond to counterfactual effects of a German who was moved to Switzerland.

Alternatively, one can think that the effects of the unobserved variables are the same in both countries, but that their distribution differs. In that case we can try and estimate the difference in residual variances between the countries and adjust the odds ratio correspondingly. This can be done with heterogeneous choice models (Allison, 1999; Williams, 2009). However, the most likely scenario is one where both the distribution of the unobserved variables and their effects differ across groups. In that case it is hard to think of a strategy that allows one to identify a causal effect.

Mood (2010) also considered the comparison of coefficients across models with different explanatory variables problematic. Such a comparison is typically done to quantify the indirect effect

of a variable. For example, we may want to know how much of the effect of parental background on the probability of entering university can be explained by the fact that children with a privileged parental background tend to perform better at school prior to entering university, and those children that performed better at school are more likely to enter university. In a linear regression model we could just estimate two models: one with just parental background and one with parental background and prior academic performance. The difference in effect of parental background between these models quantifies the indirect effect of parental background on entering university via prior academic performance. However, this trick does not work with logistic regression: If adding a variable should change the effect of parental background regardless of whether it is correlated with family background, then the comparison of coefficients between models with and without the intervening variables is not a good way of quantifying the indirect effect. Instead one can scale the effects in both models such that they refer to the uncertainty present without adding the intervening variables (Buis, 2010; Erikson, Goldthorpe, Jackson, Yaish, & Cox, 2005) or scale the effects in both models such that they refer to the uncertainty present when also adding the intervening variables (Karlson et al., 2012). Either can be reasonable. The logic behind the former is that the explanandum is the total effect, which is the effect of the variable of interest without adding the intervening variables. The logic behind the latter is that the intervening variable is part of the model, so the uncertainty that remains after adding the intervening variable is the relevant amount of uncertainty. A consequence of that is that the total effect using the method by Karlson et al. (2012) will differ depending on which intervening variables one wants to model. So if one wants to compare indirect effects for different intervening variables then that would be easier using the method by Erikson et al. (2005) or Buis (2010).

Summary and conclusions

There has been a long discussion on whether the influence of unobserved heterogeneity makes the interpretation of logistic regression coefficients problematic. Mood (2010, pp. 67-68) summarized that discussion by naming three problems:

1. “It is problematic to interpret log-odds ratios or odds ratios as substantive effects, because they also reflect unobserved heterogeneity
2. It is problematic to compare log-odds ratios or odds ratios across models with different independent variables, because the unobserved heterogeneity is likely to vary across models.
3. It is problematic to compare log-odds ratios and odds ratios across samples, across groups within samples, or over time – even when we use models with the same independent

variables – because the unobserved heterogeneity can vary across the compared samples, groups, or points in time."

This article adds to this debate by stating that unobserved heterogeneity does have the effects as summarized by for example Mood (2010) and Auspurg and Hinz (2011), but that that does not have to be a problem and in many cases is actually desirable. The key difference with previous contributions is that a probability is considered to be an assessment of how likely it is that an event occurs. A probability should thus be dependent on the information available for making such an assessment; the more (mutually consistent) information we have the surer we are. In logistic regression the set of explanatory variables represents the information used for making such an assessment. So adding variables to a model that predicts a probability should result in different coefficients.

To be more precise, consider what one would expect if one adds an additional variable *z*, and that variable is relevant for predicting a probability. In that case one will be more certain after adding it. In other words, the probabilities should be closer to either zero (I am more certain that the event does not happen) or one (I am more certain that the event does happen). This should also influence the effect of other variables. Say I am interested in the effect of a variable *x*. Before adding the additional variable *z* the predicted probabilities should be further away from either zero or one compared to after adding that additional variable. As a consequence there is less room for the variable *x* to have an effect before adding *z* than after adding *z*. So adding the additional variable *z* should increase the effect of a variable, even if *x* and *z* are uncorrelated. The log-odds ratios and odds ratios from a logistic regression show exactly this behavior. So Mood (2010) was right when she noted in the first problem that these coefficients are dependent on which variables are included in the model even if those additional variables are uncorrelated with the variables of interest, but that property of logistic regression is actually desirable instead of being a problem.

The size of the effect on a probability is a function of how certain we are that the event of interest happens. This degree of certainty can change by adding a variable as discussed above, but it can also differ from group to group. Here we would expect stronger effects in groups where are more certain. Within groups were we are more certain the predicted probabilities can get closer to zero or one, so there is more room for a variable to have an effect. Logistic regression coefficients have exactly this property, as can be most clearly seen in the latent variable representation of logistic regression. So rather than making odds ratios incomparable across groups, this property of logistic regression ensures that a comparison of odds ratios give an accurate description of the difference in effects across groups.

Under special circumstances this comparison can also be given a causal interpretation. The difference in certainty we have across groups is captured by the (unobserved) difference in variance of the residual across groups. If one assumes that this difference in variance is due to a difference in effects of the unobserved variables that make up the error term and the distribution of these unobserved variable is the same across groups, then a comparison of odds ratios also capture the counterfactual thought experiment where a person is moved from one group to another. If the effects of the unobserved variables are the same across groups but the distribution of these variables differ, then one can use a heterogeneous choice model to get an estimate of the causal effect. If both the effects and the distribution of the unobserved variables differ across groups the causal effect is unidentified. So Mood's third problem is not as bad as she puts it: A comparison of odds ratios is an accurate way of describing differences of effects across groups, and under special circumstances this comparison can be given a causal interpretation.

The comparison coefficients across models is more complicated. The purpose of such a comparison is often to see how much of an effect can be explained by a set of intervening variables. One first estimates a model with the explanatory variable of interest but without the intervening variables and then a model with both the explanatory variable of interest and the intervening variables. The difference in effect of the variable of interest between these models is interpreted as the part of the effect that is explained by the intervening variables. However, this does not work with logistic regression, since adding variables will change the effect even if the added variables are uncorrelated with the explanatory variable of interest. So Mood's second problem is a real problem.

To summarize, the three problems with logistic regression stated by Mood (2010) are not as bad if we interpret the dependent variable, the probability, as an assessment of how likely it is that the event of interest occurs. In that case one can conclude that:

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.

References

Allison, P. D. (1999). Comparing Logit and Probit Coefficients across Groups. *Sociological Methods & Research, 28*(2), 186-208.

Auspurg, K., & Hinz, T. (2011). Gruppenvergleiche bei Regressionen mit binären abhängigen Variablen—Probleme und Fehleinschätzungen am Beispiel von Bildungschancen im Kohortenverlauf. *Zeitschrift für Soziologie, 40*(1), 62-73.

Buis, M. L. (2010). Direct and indirect effects in a logit model. *Stata Journal, 10*(1), 11-29.

Buis, M. L. (2011). The consequences of unobserved heterogeneity in a sequential logit model. *Research in Social Stratification and Mobility, 29*(3), 247-262.

Erikson, R., Goldthorpe, J. H., Jackson, M., Yaish, M., & Cox, D. R. (2005). On class differentials in educational attainment. *Proceedings of the National Academy of Sciences of the United States of America, 102*(27), 9730-9733.

Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika, 71*(3), 431-444.

Hauck, W. W., Neuhaus, J. M., Kalbfleisch, J. D., & Anderson, S. (1991). A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology, 44*(1), 77-81.

Karlson, K. B. (2015). Another Look at the Method of Y-Standardization in Logit and Probit Models. *The Journal of Mathematical Sociology, 39*(1), 29-38.

Karlson, K. B., Holm, A., & Breen, R. (2012). Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit: A New Method. *Sociological Methodology, 42*, 286-313.

Keynes, J. M. (1921 [2004]). *A Treatise on Probability*. Mineola, NY: Dover Publications, Inc.

Lee, L.-F. (1982). Specification error in multinomial logit models: Analysis of the omitted variable bias. *Journal of Econometrics, 20*(2), 197-209.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics*. Cambridge: Cambridge university press.

Mood, C. (2010). 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.

Neuhaus, J. M., & Jewell, N. P. (1993). A Geometric Approach to Assess Bias Due to Omitted Covariates in Generalized Linear Models. *Biometrika, 80*(4), 807-815.

Williams, R. (2009). Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups. *Sociological Methods & Research, 37*(4), 531-559.

Winship, C., & Mare, R. D. (1984). Regression Models with Ordinal Variables. *American Sociological Review, 49*(4), 512-525.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT press.