Logistic regression: Uncovering unobserved heterogeneity

Carina Mood*

April 17, 2017

1 Introduction

The logistic model is an elegant way of modelling non-linear relations, yet the behaviour and interpretation of its effect estimates is more complex than for coefficients in linear regression. Mood (2010) explained why coefficients, i.e., odds ratios (OR) or log-odds-ratios (LnOR) from logistic regression cannot be compared across regression models or groups using the same intuition as for linear regression. The reason is that these coefficients reflect not only the effects on the outcome, but also unobserved heterogeneity, which may differ across groups or models. Since the publication of Mood (2010), there has been a widespread change of practice in reporting, with OR and LnOR giving way to effect measures expressed on the probability scale. While this is desirable, the understanding of the underlying rationale is lagging behind practice. Many researchers still see unobserved heterogeneity as a mysterious concept that is hard to grasp, and in this article I seek to de-mystify the concept in order to help empirical researchers better understand what unobserved heterogeneity is, how we can measure it, and why we normally want effect measures unaffected by it. I show that an LnOR can be seen as consisting of three parts: The average percentage unit effect (AME), the proportion with the outcome (base rate), and the proportion unexplained variance. These components can be manipulated to adjust LnOR for unobserved heterogeneity but I argue that this is seldom particularly useful given that measures based on percentage units normally speak better to our questions.

The argument here is not that the logistic model or its associated OR and LnOR are wrong or problematic in themselves.[1] The problems related to unobserved heterogeneity lie entirely in the interpretation of the results, and the gist of the problem is the discrepancy between the questions we normally purport to answer and the questions that the LnOR and OR respond to. Too often, researchers are concerned with choosing the method that is "right" in some *general* sense, while the concern should

*Swedish Institute for Social Research, Stockholm University

[1] I make this very clear statement early so that even casual readers get this point, because critique concerning how logistic regression is used is often misconstrued as critique of the model itself.

rather be to make it very clear what the actual question is, and thereafter choose a method and way of reporting that best speak to this question. Very often, we cannot do what we think (or, increasingly, thought) we can do with logistic regression, and to solve this discrepancy we must either adapt our analysis to correspond to our questions, or re-formulate our questions so that they match the answers provided. Although the first strategy is in my view often the soundest, one can envisage cases where the questions should be reconsidered. The crucial point is that the analytical output and questions must be aligned.

2 What logistic regression does

A common way of explaining unobserved heterogeneity in logistic regression is in terms of the omission of one variable from a true model, where the true model is normally defined in terms of a latent variable. For example, the true model may be:

$$y_i^* = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \tag{1}$$

where i indexes individuals, y^* denotes the value of the (unobserved) latent variable, x_1 and x_2 are independent variables that are uncorrelated with each other, β_1 and β_2 are effects (to be estimated) on the latent variable for a one unit change in the respective independent variable, and ϵ is an unobserved individual error term with a standard logistic distribution with variance 3.29. The omission of x_2 from equation (1) would lead us to estimate β_{1R}, an underestimation of β_1, according to equation (2) (for more details, see Mood (2010), p.69):

$$\beta_{1R} = \beta_1 \frac{\sqrt{3.29}}{\sqrt{3.29 + \beta_2^2 var(x_2)}} \tag{2}$$

This is a useful way of explaining the logic behind, but it is not so helpful in practice. The latent variable formulation has a mystical flavor that many find intimidating - we measure effects on something that is not observed or measured, does not have a known scale, and which perhaps does not even exist. In addition, the framing of unobserved heterogeneity in terms of the deviation from a true latent model works well in theory and in simulations, but does not reflect the reality faced by the empirical researcher. In many cases it is difficult to say if there is such a thing as a latent propensity, and even when this may be warranted we rarely know which variables that would go into the true model. For the rare case where we think that we know the true model in theory, we are seldom able to observe all the variables that go into it. This means that in practical research with real data, we almost always face unobserved heterogeneity, and unlike in simulations we cannot determine how our coefficients compare to those of the (presumably) true model.

There is however an alternative and, I think, both more intuitive and more useful way of explaining the problem, which I develop in this article. I show how we can

understand unobserved heterogeneity in terms of the observed empirical data without any appeal to underlying latent variables. In this approach, the focus is on the predictive capacity (or explained variance) of the model, which is measurable and hence opens up possibilities to quantify the unobserved heterogeneity and to estimate its impact on effect estimates. Using this perspective, the term unexplained variance is more apt than unobserved heterogeneity, and henceforth I use these terms interchangeably.

The normal reason that researchers choose logistic regression is that the scale of the dependent variable is binary, taking either of the values 1 (does have the outcome) or 0 (does not have the outcome). Yet what we estimate from logistic regressions are effects on log-odds, odds or probabilities of having the outcome, and these are not observed on the individual level. Let y_1 represent the binary (0/1) dependent variable and p_i the probability that individual i has ($y_1 = 1$), and consider a logistic regression of the form:

$$ln\frac{p_i}{1-p_i} = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} \tag{3}$$

Where x_1 and x_2 are observed independent variables (continuous or categorical) and α, β_1 and β_2 are parameters to be estimated. The lefthand side here is not the individual-level observed value of y_1 but a prediction of the model – the logarithm of the odds of having the outcome ($y_1 = 1$) vs. not having it ($y_1 = 0$). If we exponentiate this predicted logged odds of having the outcome, we obtain the odds of having the outcome vs. not having it, but what we are ultimately interested in is normally neither the log-odds nor the odds, but rather the probability of having the outcome. To obtain this, we translate the predicted log-odds from Equation 3 to predicted probabilities through equation (4):

$$p_i = \frac{exp(ln(p_i/(1-p_i))}{1 + exp(ln(p_i/(1-p_i))} \tag{4}$$

So the three scales – log-odds, odds and probabilities – represent three ways of expressing the likelihood of an outcome, and each can be transformed to another (except for probabilities of 0 and 1, that cannot be expressed as an odds or log-odds). None is observable on the individual level.

The input to the logistic model consists of individual-level observed y_1-values of zero or one in combination with the same individuals' values on x_1 and x_2, so the information content that the model has is the distribution of y_1 across values of x_1 and x_2. Consequently, p_i equals the predicted proportion of people with y_1=1 at certain values of x_1 and x_2. The model seeks to estimate the parameters (LnOR) that maximizes the likelihood of observing the given empirical distributions, under the assumption that the association between y_1 and the independent variables follows a logistic functional form.[2]

The logistic probability density function (PDF) is $p(1-p)$, and it is illustrated in

[2]If the data contain a sufficient number of individuals for each combination of values on x_1 and x_2, one can work directly with aggregate proportions in a least squares model (Berkson, 1944; Aldrich and Nelson, 1984).

Figure 1 along with the CDF (cumulative distribution function). The PDF is largest, 0.25, when the probability is 0.5, which occurs at a log-odds of 0.

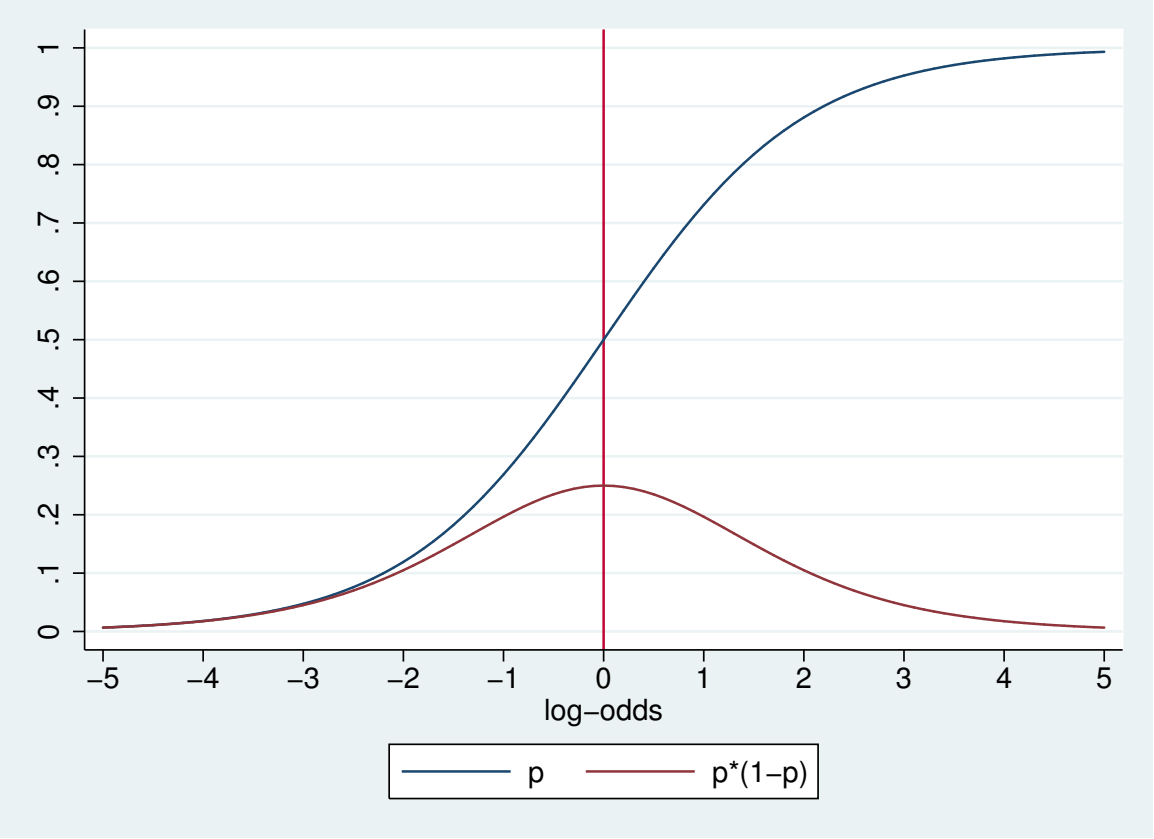

Figure 1: Logistic probability density function (PDF, $p \times (1-p)$) and cumulative distribution function, (CDF, p) by log-odds

Assuming x_1 is continuous,the estimated percentage unit effect of x_1 for a given individual is $\beta_1 \times p_i(1-p_i)$. Thus, the estimated average percentage unit effect of x_1 is simply:

$$\frac{1}{n}\sum_{i=1}^{n} \beta_1 \times p_i(1-p_i) \tag{5}$$

Which is normally labelled the average marginal effect (AME). AME (and its discrete alternatives, see Mood (2010)) is known to be strongly robust to differences in unobserved heterogeneity (Cramer, 2007), and to understand why this is, consider the case where x_1 and x_2 are uncorrelated. If x_1 does not vary with x_2, the average probability of $y_1 = 1$ at different values of x_1 should logically be unaffected by whether or not we observe x_2 (and vice versa). Because these observed distributions are the fixed input that the model seeks to reproduce, the predicted probabilities across values of x_1, and hence the estimated change in these probabilities for changes in x_1, should be similar whether or not x_2 is in the model.

To exemplify, let us revisit Figure 1 in Mood (2010), here Figure 2, which shows predicted probabilities from a logistic regression model (simulated data) of transition to university on IQ, first in a bivariate model (model 1, bold line) and then in a model conditional on sex (model 2, girls solid line, boys dashed line). The model also shows the average predicted probability from model 2, across both boys and girls, at different values of IQ (dotted line).

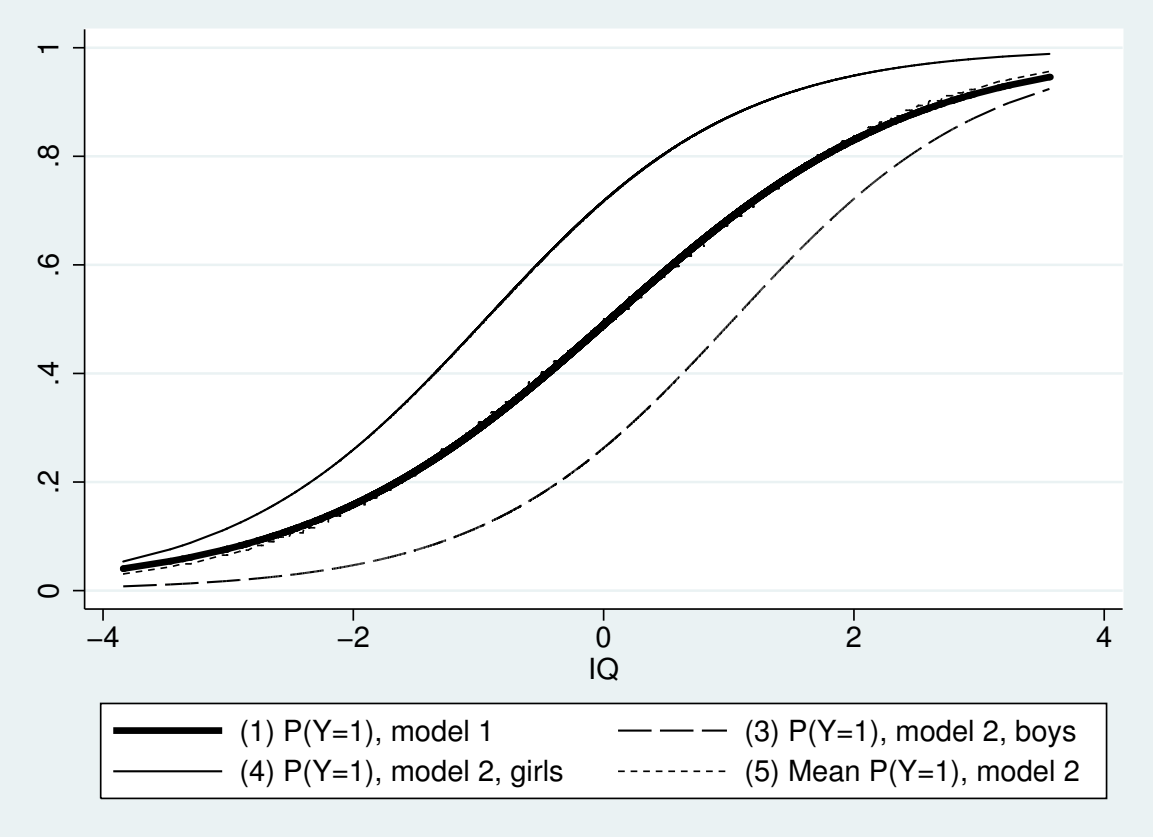

Figure 2: Probability of transition to university by IQ and sex (simulated data)

We see that at all levels of IQ, girls are more likely to go to university than boys. But, importantly, the average predicted probabilities from model 2, i.e., the average of the curves for boys and girls is at all levels of IQ as good as identical to the predicted probabilities from model 1 (which did not include sex), as seen by the fact that the dotted curve overlaps almost perfectly with the bold curve. The average slopes of the bold and dotted curves give the AME from models 1 and 2, respectively, and obviously these are also as good as identical in the two models despite a large increase in the LnOR and OR in model 2. What we observe here is an example of the nice feature of percentage unit differences being collapsible (Greenland, Robins, and Pearl, 1999), that is – in contrast to LnOR and OR – they are unchanged by the inclusion of uncorrelated predictors. Because logistic regression assumes a particular functional form, AME are not always robust in this way, but they tend to be so if the logistic functional form assumption is not gravely wrong (Cramer, 2007).

If we think of x_1 as IQ and x_2 as sex in Equation 3, Model 1 in Figure 1 corresponds to estimating Equation 3 without x_2, i.e.:

$$ln\frac{p_{Ri}}{1-p_{Ri}} = \alpha + \beta_{1R}x_{1i} \tag{6}$$

Where R is used as a subscript to show that these estimates come from a reduced model. The corresponding AME is:

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{1R} \times p_{Ri}(1-p_{Ri}) \tag{7}$$

With knowledge of both x_1 and x_2, we can predict the outcomes better than when we know only x_1, so the predicted probabilities from Equation 3 are closer to the individual outcomes (0 or 1) than the predicted probabilities from Equation 6. As a consequence, $p_i(1-p_i)$ is on average smaller than $p_{Ri}(1-p_{Ri})$.

So we get better predictions and smaller $p_i(1-p_i)$ with x_2 in the model, but at the same time we know that the average predicted probabilities of $y_1 = 1$ at different levels of x_1 should be similar regardless of whether x_2 is in the model, meaning that we should also get similar end results from Equations 5 and 7. In order for the predicted probabilities of the model to match the given distribution of y_1 over x_1 (remember that these are the fixed inputs that the model adapts its estimates to) when the predictions move closer to 0 and 1, the model must "compensate" the smaller average $p_i(1-p_i)$ by estimating a larger β_1. So the estimated coefficients have to increase in proportion to the difference between $p_i(1-p_i)$ and $p_{Ri}(1-p_{Ri})$, resulting in the relation:

$$\frac{\sum_{i=1}^{n} p_{Ri}(1-p_{Ri})}{\sum_{i=1}^{n} p_i(1-p_i)} = \frac{\beta_1}{\beta_{1R}} \tag{8}$$

So, for example, if x_2 reduces average $p_i(1-p_i)$ by half, β_1 will double. This relationship is not exact as it depends on the fit of the models to the assumed functional form. For probit models, Wooldridge (2002) proved the equality between ratios of derivatives and ratios of β under the assumption that the functional form is correctly specified, but although the intuition is the same for logit models the relationship has so far only been shown by simulation by Cramer (2007). He finds that even blatant functional form misspecification has only a minor impact, meaning that Equation 8 can be treated as valid for most practical purposes. We can think of the logic behind the sensitivity of LnOR and OR to unobserved heterogeneity as a consequence of these quantities being what is "flexible" in the model, that is they are what can change in order to match what is exogeneously given – the observed distribution of y_1 over the independent variables.

3 Predicted probabilities and explained variance

In logistic regression, the individual-level outcome of interest (y_1) can only take on values 0 or 1, and the average probability of a positive outcome is simply the average of y_1 over all individuals: $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} y_{1i}$. Just as in an OLS, we can measure the extent to which our model improves the prediction of y_1 in terms of the proportionate reduction of residual variance:

$$R^2_E = 1 - \frac{\sum_{i=1}^{n}(y_{1i} - p_i)^2}{\sum_{i=1}^{n}(y_{1i} - \bar{p})^2} \tag{9}$$

R^2_E in Equation 9 defines what is normally labelled *Efron's* R^2 (e.g., Mittlböck and Schemper, 1996), and is analogous to the standard R^2 in OLS in terms of the use of squared deviations, with $\sum_{i=1}^{n}(y_{1i} - p_i)^2$ representing the residual sum of squares (RSS), i.e. the total variance around the regression predictions, and $\sum_{i=1}^{n}(y_{1i} - \bar{p})^2$ representing the total sum of squares (TSS), i.e., the total variance around the mean.

Another way to express the total variance in a binary variable (TSS) is $\bar{p}(1-\bar{p})$, which is also the logistic probability distribution function (PDF) at the average of y_1. If the logistic functional form is correct, the residual variance is the average of $p_i(1-p_i)$, that is, the average of the PDF at the predicted probabilities from the model, averaged over all individuals. So, assuming the logistic functional form is correct, Efron's R^2 in Equation 9 is identical to Equation 10:

$$R^2_G = 1 - \sum_{i=1}^{n} \frac{p_i(1-p_i)}{\bar{p}(1-\bar{p})} \tag{10}$$

which has been labelled *Gini's concentration* R^2 (Mittlböck and Schemper, 1996) (I will here refer to it as *Gini's* R^2 for short). To understand the logic behind, consider a sample with mean $y_1 = 0.4$. At this point $\bar{p}(1-\bar{p})$ is 0.24, i.e., $(0.4 \times (1-0.4))$. Say that we estimate a model with a dummy independent variable that is associated to y_1, resulting in predicted probabilities that are 0.2 and 0.6 (in equal proportions), giving $p_i(1-p_i)$ of 0.16, i.e., $(0.2 \times (1-0.2))$; and 0.24, i.e., $(0.6 \times (1-0.6))$, averaging to 0.20, i.e., smaller than the 0.24 we started with. The R^2_G would in this case be $1 - (0.20/0.24) = 0.17$.

If the true functional form is not logistic, R^2_G will differ from R^2_E, with R^2_E giving the correct reflection of the explained variance by the model but an underestimation of the explained variance that we would get with a more correct functional form. If the two R^2 differ, it is thus a sign that the logistic functional form is inappropriate, and that, if we want to draw substantive conclusions from this model we need to test other functional forms. In practice, however, the two R^2s are normally close (Mittlböck and Schemper, 1996).

So changes in R^2 reflect the decrease in the proportion residual variance, which – if the logistic specification is correct – is the same as the ratio of the PDF at the model's predicted values to the PDF at the average of the outcome. Given Equation 8, we can then express the relation between β_1 and β_{1R} as:

$$\beta_1 = \beta_{1R} \sum_{i=1}^{n} \frac{p_{Ri}(1-p_{Ri})}{p_i(1-p_i)} \tag{11}$$

Which can also be written as:

$$\beta_1 = \beta_{1R} \frac{1-R^2_{GR}}{1-R^2_G} \tag{12}$$

Thus, there is a straightforward relationship between the predicted probabilities, the explained variance and the LnOR at a given base rate.

The above also shows that the better the predictive capacity of the model, the larger will the difference be between AME and marginal effects evaluated at the mean of y_1. The PDF has its maximum at the mean of y_1, meaning that marginal effects are largest when evaluated at this point. Hence, marginal effects at the mean of y_1 can be misleading if used to approximate the AME.

4 What an odds ratio or log-odds ratio consists of and the consequences for comparisons across models, groups, etc.

An OR gives the multiplicative effect on the odds of an outcome, and an LnOR gives the additive effect on the log of this odds. The natural scale for a 0/1 outcome is however the probability, and a given average effect on the probability can correspond to different LnOR and OR. Above we saw that the size of the LnOR or OR is affected by the predictive capacity of the model. This mechanism is a manifestation of the more general principle in logistic regression that the LnOR "weighs" the percentage unit effect by the relative difficulty of getting an effect of this size given the logistic PDF. If we have an outcome that is very rare (e.g., only 5 percent have a positive outcome) or very common (e.g., 95 percent have a positive outcome), the logistic slope is flatter and hence a given percentage unit effect requires a higher OR or LnOR than a same-sized percentage unit effect on an outcome that is more close to a 50/50-distribution. But if we can predict the outcome with the 50/50 distribution very well, so that our predicted probabilities lie close to 0 or 1, we are again at places in the logistic distribution where the slope is flatter, and a one percentage unit effect will correspond to a higher LnOR and OR than if our model has weak predictive power. Both these mechanisms reflect the non-linear shape of the logistic PDF, seen in Figure 1.

Combining these insights, it is easy to show that OR and LnOR estimating the effect of x_1 on y_1 can vary because of: (1) Differences in percentage unit effects of

x_1 on y_1; (2) Differences in base rates (i.e. averages of y_1), and (3) Differences in explained variance. We can see this by noting that the average of $p_i(1-p_i)$ is equal to $\bar{p}(1-\bar{p}) \times (1-R^2_G)$, and rewriting Equation 5 as:

$$\beta_1 = \frac{AME_1}{\bar{p}(1-\bar{p}) \times (1-R^2_G)} \tag{13}$$

The numerator is the AME, and the denominator consists of the total variance $\bar{p}(1-\bar{p})$ (determined by the base rate of the outcome) multiplied by the proportion unexplained variance. In other words, the denominator is the total unexplained variance in absolute terms, so one can also simply say that the LnOR is equivalent to the estimated average percentage unit effect divided by the total unexplained variance in the outcome. The total unexplained variance can under no condition be higher than 0.25, a maximum which will occur only if the base rate is 0.5 and the R^2 is close to zero.[3]

Equation 13 shows how effect estimates depend on different communicating parts, and we can manipulate these parts in different ways depending on our purposes: For example, we can calculate what LnOR a given AME would correspond to, or what AME a given LnOR would result in, at a higher explained variance and/or a higher base rate. When the base rate is constant, e.g., when we compare different OR and LnOR from nested models in a given sample, differences can be due only to differences in AME, or in the proportion unexplained variance. So changes in LnOR when introducing an additional variable x_2 to a model can be due either to this variable being correlated to both x_1 and y_1, and hence acting as a confounder or suppressor when not included, or to the increase in explained variance that x_2 gives rise to. If we want to neutralize the effect of differences in explained variance, we can use Equation 14 to estimate what a given LnOR (β_1) would be if we had the proportion explained variance of another model.

$$\beta_{1alt} = \beta_1 \frac{(1-R^2_G) \text{ in model}}{(1-R^2_G) \text{ in alternative model}} \tag{14}$$

There is however a problem here: A different R^2 for a given model is not compatible with the observed distribution of the independent and dependent variables in a given sample, so the underlying rationale for Equation 14 must therefore be in terms of addition to the model of an unexplained variance parameter, correlated to the dependent variable but uncorrelated to the independent variable, which drives the R^2 up to the level of an alternative model.

In the case of comparisons of LnOR more generally, i.e., across groups, samples, outcomes, etc., we additionally need to take differences in base rates into account. Here, it also becomes important to distinguish between the bivariate and the multivariate case. In his seminal contribution on unobserved heterogeneity and comparisons across groups, Allison (1999) did not make this distinction, but the implications of unobserved heterogeneity is quite different in these two cases. If we estimate an

[3]LnOR divided by 0.25 is the definition of the *Lambda* effect measure (see Hellevik (2009), which thus essentially gives the maximum percentage unit effect that a given LnOR can correspond to.

LnOR in a bivariate model, i.e., studying associations between x_1 and y_1 with no further independent variables), differences across groups depend on differences in (1) the effect size of x_1 in percentage units, (2) variances in x_1 (affecting the spread in predicted values and hence residual variance) and (3) averages of y_1 (determining total variance). Thus, in the bivariate case OR and LnOR can be used as a measure of association, combining the effect size and the explained variance attributed to the independent variable.

The multivariate case is however different, because LnOR and OR for any given variable will be affected by the variance explained by all independent variables. For example, in a model with x_1, x_2 and y_1, the LnOR or OR for x_1 will depend also on (1a) the effect size of x_2 and (2a) the variance in x_2, as these contribute to the explained variance. If our question concerns the association between x_1 and y_1, it matters whether the coefficient for x_1 is higher because of x_1 or because of x_2, yet the coefficient itself cannot tell us the difference.

5 Unobserved heterogeneity and a special case turned upside up

One argument sometimes raised is that OR or LnOR in a regression of y_1 on x_1 are only affected by unobserved heterogeneity caused by an omitted variable x_2 if there is a correlation between x_1 and x_2 conditional on y_1. This argument, it turns out, is merely an upside-down perspective on the general principle described above. It is upside-down from a theoretical and causal point of view because we presume x_1 and x_2 to affect y_1 and not the other way around, meaning that the correlation between x_1 and x_2 given y_1 is not an independent feature of the process. If x_1 and x_2 are uncorrelated in the population, and if both have independent effects on the prediction of y_1, there will be a correlation between x_1 and x_2 conditional on y_1, and this correlation is caused by the independent effects of x_1 and x_2, respectively, on y_1. The observed correlation between x_1 and x_2 given y_1 shows that both x_1 and x_2 contribute independently to the prediction of y_1, and it is this independent contribution of x_2 to the prediction of y_1 that is the mechanism behind the change in OR and LnOR – the lack of a correlation between x_1 and x_2 given y_1 is just a secondary effect of it.

The simple principle is that the size of any increase in the LnOR and OR for x_1 when including x_2 in the model depends on how much x_2 increases the ability to predict y_1 (i.e. how much we reduce the unobserved heterogeneity in y_1 as compared to when only x_1 was in the model). If the LnOR and OR for x_1 are not affected when including the uncorrelated x_2 in the model, it means that x_2 does not add much to our ability to predict y_1 once we know x_1. Thus, the correlation between x_1 and x_2 given y_1 is not the condition for the attenuation of the LnOR for x_1, but the correlation and the attenuation are common outcomes of the contribution of x_2 to the prediction of y_1.

For example, consider Figure 2 above (or Figure 1 in Mood (2010)), where IQ

and sex have independent effects on the transition to university. Among those who go to university (y_1 =1), women will have on average lower IQ than men because it takes a higher IQ for a man to go to university, so we have an association between sex and IQ at y_1=1. At the same time, men who choose not to go to university have on average higher IQ than women who choose not to go to university, so we will have an association between sex and IQ at y_1=0 as well. In fact, men will on average be more intelligent than women *both among those who go to university and those who do not do so*, but men are not more intelligent than women on average in the population. This seeming paradox is caused by the independent effects of IQ and sex on the transition to university.

6 Understanding the impact of unobserved heterogeneity: Example analyses

In order to illustrate how one can measure unobserved heterogeneity and assess its role on the effect estimates of interest, I use the dataset on careers of biochemists that has previously been used in the literature about unobserved heterogeneity in logistic regression (Allison, 1999; Williams, 2009; Long, 2009).[4] The dataset is longitudinal, consisting of person-years, and the dependent variable here is whether an assistant professor is tenured (1) or not (0), and the particular question in focus in Allison (1999) was whether men and women had different payoff of their number of published articles on their chances of tenure. He additionally included variables measuring time since the start of the assistant professorship (duration in years, also included in squared form), the selectivity of the college where the undergraduate degree was obtained (range 1-7), and the prestige of the current department (range 0.65-4.60). For a more detailed description of data and variables, see Allison (1999, p.187). Table 1 reproduces the results from Allison (1999:188, Table 1), but also bivariate models, information about AME (estimated by Stata's margins command), proportion observations with a positive outcome in each group, and R^2 (Gini's and Efron's). Statistical significance is reported using the same criteria as in Allison's article.

LnOR give the additive effects on the natural log of the odds of being tenured, so for example, in the bivariate model for men, we see that the log of the odds of being tenured increases by, on average, 0.102 units for each article. If we think that it is appropriate to see tenure as generated by a latent propensity for tenure, we can also interpret this as saying that the latent propensity to be tenured increases by, on average, 0.102 units for each article. This effect size does not tell us about the differences in the probability of being tenured, and it is difficult to give it a substantive interpretation, because the scale has no upper or lower bound. As put by Aldrich and Nelson (1984, p. 41) the LnOR has only ordinal, not cardinal meaning, so if we show only the LnOR the reader can only conclude that number of articles increases the probability of tenure, but not by how much in any substantive sense.

[4]Retrieved via code in Williams (2009) from "http://www.indiana.edu/ jslsoc/stata/spex_data/tenure01.dta"

If we exponentiate 0.102, we get the OR, which gives the multiplicative effect on the odds of being tenured. In this case it is 1.11, saying that for each article, the odds of tenure increases by a factor of 1.11 for each article, or in other words, for one extra article the ratio of tenured to non-tenured men increases by a factor of 1.11. It is extremely common to misunderstand or misrepresent OR as a multiplicative effect on the probability, which is only true if the outcome is rare.

The AME gives the average effect on the probability of the outcome. In the first model, we see that each article is associated to a 1.1 percentage unit increase in the probability of tenure, which is a substantial effect given that the average probability of tenure among men is 13.2. For women, the corresponding effect is clearly smaller at 0.4 percentage units, and their average probability of tenure is 10.9 percent.

Table 1: Example analysis: Logistic regression of tenure for men and women

	Men				Women			
	LnOR	AME	LnOR	AME	LnOR	AME	LnOR	AME
Number of articles	0.102***	0.011***	0.074***	.007***	0.047***	0.004***	0.034**	.003**
	(0.010)	(0.001)	(0.012)	(0.001)	(0.010)	(0.001)	(0.013)	(0.001)
Duration			1.909***	0.035***			1.408***	0.027***
			(0.214)	(0.003)			(0.257)	(0.003)
Duration squared			-0.143***				0.096***	
			(0.019)				(0.022)	
Undergraduate selectivity			0.216***	0.020***			0.055	0.005
			(0.061)	(0.006)			(0.072)	(0.006)
Job prestige			-0.431***	-0.040***			-0.371*	-0.032*
			(0.109)	(0.010)			(0.156)	(0.013)
N	1741		1741		1056		1056	
Efron's R^2	0.058		0.171		0.008		0.097	
Gini's R^2	0.075		0.181		0.027		0.108	
Proportion tenured	0.132		0.132		0.109		0.109	

Note: AME for duration is calculated using the combined information from duration and duration squared.
*** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.05$

6.1 Unexplained variance across models

In the bivariate regressions, the LnOR for number of articles is 0.102 (men) and 0.047 (women). When controlling for the other independent variables, it shrinks to 0.074 (men) and 0.034 (women), suggesting that the estimated effect of articles on the log-odds of tenure can to some extent be accounted for by these other variables. However, the coefficients from the two models cannot be straightforwardly compared, because they are estimates of two different parameters, one that is averaged over the distribution of duration, undergraduate selectivity and job prestige, and one that is conditional on these variables (cf. Mood 2010:72). If the variables added in model 2 contribute independently to the prediction of tenure, the LnOR or OR for articles in the two models would differ even if articles was uncorrelated to the other independent variables.

R^2 increases from the bivariate to the multivariate model from 0.058 and 0.075 to 0.171 and 0.181 (men, Efron's/Gini's) and from 0.008 and 0.028 to 0.097 and 0.108 (women, Efron's, Gini's), which serves to make the LnOR and OR for articles in model 2 higher than it would be had the explained variance been the same as in model 1.[5] Therefore, the extent to which duration, selectivity, and prestige account for the bivariate association between articles and tenure is underestimated by a comparison of coefficients between models 1 and 2. From Equation 13, we know that each LnOR consists of three components: AME, $\bar{p}(1-\bar{p})$ and $(1-R^2_G)$, and these components are shown in Table 2.

Because the base rate is the same across models in a fixed sample, changes in the LnOR can only be driven by changes in the AME and in the proportion unexplained variance. Here, we see that the increase in explained variance in the multivariate model only suppresses the denominator slightly, from 0.106 to 0.094 for men, and from 0.094 to 0.087 for women. In order to understand the impact of this reduction in unobserved heterogeneity for comparisons across models, we can use Equation (14) and multiply the coefficient from the bivariate model with the ratio of the unexplained variance in the bivariate model to the unexplained variance in the multivariate model (see Appendix for a demonstration of how this can be done in a regression framwork).

For men, this changes the bivariate coefficient for articles from 0.102 to 0.116, and for women from 0.047 to 0.051. Relating these to the multivariate coefficients, we confirm that the proportion of the coefficient accounted for by the other independent variables was attenuated by unobserved heterogeneity: When re-scaling to the higher R^2, the reduction in the coefficient is 36 rather than 27 percent (men) and 34 rather than 28 percent (women) of the bivariate association. As noted above, it is important to note that with only articles in the model, the counterfactual R^2 is not compatible with the observed distribution of tenure over articles. If articles is the only variable in the model, R^2 can only be higher if the probability of tenure is more strongly affected by the number of articles, or if the base rate of tenure is further from 0.5. Neither of these can be true within a given sample with given distributions of independent

[5] Note that Efron's and Gini's R^2 differ substantially in the bivariate models, which suggests that the logistic functional form is not entirely appropriate. Had we had a substantial interest in this analysis we would investigate this more closely.

Table 2: Components of the log-odds-ratio for articles

	Men		Women	
	Bivariate	Multivariate	Bivariate	Multivariate
AME (numerator)	0.011	0.007	0.004	0.003
$\bar{p}(1-\bar{p})$	0.114	0.114	0.097	0.097
$1-R^2$	0.925	0.819	0.973	0.892
Total denominator	0.106	0.094	0.094	0.087
Resulting LnOR	0.102	0.074	0.047	0.034

and dependent variables, so we should think of this operation as an addition to the bivariate model of an unexplained variance parameter, correlated to tenure but uncorrelated to articles.[6]

6.2 Unexplained variance and group comparisons

The focus of Allison's (1999) article was the differences between the LnOR for articles between men and women in the multivariate model. The difference (0.074 vs. 0.034) was statistically significant, but Allison argued that this could be due to differences in unobserved heterogeneity, i.e., that women have more heterogeneous career patterns than men, making unmeasured variables more important for women (note that this is an argument about the *proportion* unexplained variance). He proposed the estimation of a parameter capturing the difference in unobserved heterogeneity by means of adding a constant weight to the interaction terms of sex with all independent variables, and letting the model estimate the size of this weight. This method builds on the insight that unobserved heterogeneity has a uniform effect on all coefficients, but it is problematic in that it confers *any* systematic tendency for women to have smaller LnOR only to differences in unexplained variance (cf. Williams (2009)). As discussed above, two other factors can affect the differences in LnOR: Women can have a different base rate of tenure, and they can have smaller percentage unit effects. So Allison's method controls for more than the differences in the proportion unobserved heterogeneity, and it may therefore underestimate the difference between men and women.

Using the approach discussed here, we can straightforwardly measure the three different components and assess their contribution to the difference in the LnOR for articles. First, R_{2G} tells us that there is indeed a difference in the proportion unexplained variance: 89 percent of the variance in tenure among women is due to unobserved factors, with the corresponding number being 82 percent for men. Second, the base rate differs slightly, with tenure observed in 13.2 percent of the observations among men and 10.9 percent among women. Third, the AME differs strongly: Among men, the probability of tenure increases by an average of 0.7 percentage points for

[6] Although they do not frame it in terms of explained variance, this procedure operates in the same way and gives the same result as the method proposed by Karlson, Holm, and Breen (2012) for comparisons of changes in coefficients across models.

every article, and among women it increases by only 0.3 percentage points. Using the same procedure as in Table 2, we can see that:

The LnOR for articles among men equals: $\frac{0.7}{0.132\times(1-0.132)\times 0.819}$

The LnOR for articles among women equals: $\frac{0.3}{0.109\times(1-0.109)\times 0.892}$

The AME, the numerator, contains the information that is often of fundamental interest: Do men, on average, get a higher payoff on their number of articles in terms of the probability of getting tenure?[7] We see, as before, that indeed they do (and the difference in AME is also statistically significant at $P < 0.05$). But the difference in LnOR reflects more than this, because there are also differences in the denominator, i.e., in the total variance (which is a direct function of the base rate), and in the proportion of this total variance that is unaccounted for by the model.

We can test to which extent the difference in LnOR between men and women is due to men's lower level of unexplained variance through comparing men's LnOR with a counterfactual LnOR for women, where their R^2 is set to the same level as for men (see appendix for how to do this in a regression framework).

$$\beta_{1w} = \frac{AME_{1w}}{\bar{p}_w(1-\bar{p}_w) \times (1-R^2_m)} \tag{15}$$

The resulting counterfactual LnOR is 0.037, so the adjusted sex difference is 0.074-0.037=0.037 as compared to the original difference (0.074-0.034=0.040). Thus, less than 8 percent of the sex difference in LnOR is due to differences in unexplained variance. As noted above, this counterfactual comparison is equivalent to adding an unobserved heterogeneity parameter to the regression for women, uncorrelated to articles, that lifts their R^2 to the same level as men.

If we would also like to assess what the difference would be if women had both the same R^2 *and* the same base rate as men, we can calculate the counterfactual LnOR as:

$$\beta_{1w} = \frac{AME_{1w}}{\bar{p}_m(1-\bar{p}_m) \times (1-R^2_m)} \tag{16}$$

This gives an LnOR of 0.031, and a *larger* gap between men and women than the observed one. So the higher coefficient for men is almost entirely driven by their higher AME, and to a small extent by their lower unobserved heterogeneity – but it is suppressed by their higher base rate.

The above calculations adjust for differences in the proportion explained variance regardless of its source. Our interest, however, may be in adjusting for differences only in the part of R^2 that is due to variables other than the one in focus. As discussed above, if LnOR is higher for men because articles explain more of the variation in tenure, we can still discuss the differences between men and women in

[7]For a discussion about the linear probability model as an alternative to AME, see Mood (2010) and Hellevik (2009).

Table 3: R^2 and its components

	Men	Women
R2, only articles	0.075	0.027
R2, all variables except articles	0.143	0.096
R2, all variables	0.181	0.108
R2-components		
Due to articles only	0.038	0.012
Due to other variables only	0.106	0.081
Due to overlap	0.037	0.015

terms of differences in the impact of articles, but if men's LnOR for articles is higher because duration, selectivity and prestige explains tenure better among men this is no longer true.

In order to discern to which extent the LnOR for one variable reflects the predictive capacity of other variables in the model, we can further decompose the LnOR in a multivariate model to:

$$\beta_1 = \frac{AME_1}{\bar{p}(1-\bar{p}) \times (1 - R^2art - R^2oth - R^2overlap)} \tag{17}$$

with R^2art being the part of R^2 due only to the variance in articles, R^2oth being the part of R^2 that is due only to the variance in variables other than articles, and $R^2overlap$ is the part of R^2 due to the variance in articles that overlaps with the variance in one or more of the other variables. We can get these different components through successive comparisons in R^2, shown in Table 3.

The R^2 components are obtained through simple subtraction. For example, for men we obtain the component due to articles only given by 0.181-0.143=0.038; the component due to all variables except articles by 0.181-0.075=0.106; and the component due to their overlap by the remainder 0.181-0.106-0.038=0.037. If we would have only articles in the model, we would capture the contribution of articles and the contribution of the overlap, so the part of R^2 due to other variables is the increase in explained variance we get when we move from a bivariate to a multivariate model. This increase is due only to variation in the other variables that is uncorrelated to articles, and its impact on the LnOR for articles comes from the better predictions and hence the reduction of $p(1-p)$.

To see how much of the difference in the LnOR for articles between men and women that is accounted for by other independent variables in the model, we can estimate a new counterfactual coefficient for women:[8]

$$\beta_{1w} = \frac{AME_{1w}}{\bar{p}_w(1-\bar{p}_w) \times (1 - R^2art_w - R^2oth_m - R^2overlap_w)} \tag{18}$$

[8]We can of course estimate a counterfactual LnOR for men instead, and compare this to the observed LnOR for women)

With w denoting women and m denoting men. In this case, we find that if other variables would explain women's tenure to the same (higher extent) as for men, all else (including the AME for articles) being equal, their LnOR would be slightly higher at 0.035 rather than 0.034. Thus, if we would compare men and women with both having R^2 from other variables set to men's value, we would get a difference of 0.039, which is very close to the observed difference of 0.40. Hence, in this case, the difference in the LnOR for the variable of interest across groups was not driven by the other variables in the model.

The above tests can be straightforwardly done in a regression framework (see Appendix). However, I believe that there are few questions that motivate these or other kinds of rescaling of LnOR and OR, because these quantities are seldom interesting as endpoints in themselves but only as intermediary steps in the analysis (cf. Greenland, 1986), and the quantities that are normally of ultimate interest are identified without rescaling (cf. Long, 2009; Angrist, 2001). With binary outcomes, the observed scale is 0-1, and the substantive meaning of the results is for most purposes best conveyed on this or the equivalent 0-100 scale – in our example, the ultimate interest is generally in the probability of tenure rather than in the odds or logged odds of tenure.

The most straightforward percentage unit measure is AME, which gives us the average percentage unit effect. This is almost always a good starting point, as it neatly summarizes the substantive effect. If our question is whether men have a higher payoff of articles than women do, the fact that the average increase in the probability of tenure for each article is 0.7 percentage units for men and 0.3 percentage units for women clearly speaks directly to that question. The AME says which increase in the proportion women or men with tenure that we can expect for a one unit increase in number of articles, given the values of other observed variables in the respective group. This tells us the average effect given the actual situation that exists in the population (if the sample is representative).

We can however complement our original question and ask, e.g,: What is the difference in the payoff per article for men and women in equally prestigious departments, or with an equal number of years since PhD, or perhaps with equal values on all independent variables except articles? These questions may all be relevant, but they require a range of estimates for different values of the relevant variables, and they tell us what increase in the probability of tenure that we would expect for a change in output only among those in this specific situation. Another potential question is: What would the difference be in the counterfactual state of women having men's base rate or men's level of explained variance? This can be estimated using the reverse of the procedure used above to rescale LnOR (in this case holding the LnOR rather than the AME fixed).

The best choice of effect measure, or set of measures, can only be determined in light of the research question, and it is important to clearly distinguish between average and conditional percentage unit effects and the questions that they speak to in the case at hand. Long (2009) discusses different ways of comparing groups using effects on the probability scale, and Stata's margins command allows great flexibility for estimating percentage unit effects and testing for differences across groups under

different scenarios. Given that we can easily work directly with measures on the probability scale, I see few cases where techniques that rescale LnOR and OR are needed.

7 Relation to latent variables and "true models"

I noted in the beginning that the perspective promoted here differs from the alternative of assuming an underlying "true model", expressed in latent variable terms. In the latent variable perspective, we think of the observed binary outcome as generated by an unobserved continuous variable, as in Equation 1. In the above example, we would think of tenure as generated by an underlying propensity for tenure, denoted as y^*, and assuming that we believe that the bivariate model reflects the true data-generating process we would write the "true" models for men and women as:

Men:

$$y^*_{i\,m} = \alpha_m + \beta^*_{1\,m} articles_i + \epsilon_i \tag{19}$$

Women:

$$y^*_{i\,w} = \alpha_w + \beta^*_{1\,w} articles_i + \epsilon_i \tag{20}$$

The total variance of y^*_i consists of two parts: (1) The variance of the predictions from the model, $y*_i$, resulting from the β^*_1 and the variance in articles, and (2) The variance of ϵ. These components may differ in size, meaning that there is no fixed variance and no fixed scale for y^*_i. So even when the two "true" β^*_1 have the same numerical value for men and women, they can mean quite different things because the underlying propensity can be measured in different units. For example, if the variance in ϵ_i is larger among women than among men, while the variance in the predicted values is similar, the "true" scale of y^*_i is larger among women. A given β^*_1 therefore represents a weaker association between articles and the propensity for tenure among women (and, consequently, a smaller R^2 if we could estimate the latent model).

In the latent variable perspective, these are the "true" underlying models, and at some threshold value of the y^*_i (normally assumed to be 0), the observed binary outcome changes from 0 to 1. When we use logistic regression to estimate these models, however, we force the variance of ϵ to be the same in the two groups, which means that the variance in y^*_i can only differ due to differences in the strength of the association between articles and the propensity for tenure. Any difference in scale between men and women that is due to the absolute size of the error variances is thus eliminated, and in order to reflect the same association between articles and y^*_i, the coefficients must change in proportion to this change in scale. The coefficients in the "true" model and the logistic model thus estimate the same association between the independent and the dependent variable, and their substantive implications are the same. To say that the coefficients of the underlying model are "true" and the coefficients of the logistic model "biased" implies that we think that men's and women's propensity of tenure should be measured with different rather than standardized scales, which I believe is hard to justify. It would be akin to saying that in a linear regression of

number of articles on earnings, with men's earnings measured in dollars and women's earnings in hundreds of dollars, these raw coefficients are "true" and those expressed on a common scale are biased.

So in a bivariate model, the logistic regression coefficients can be interpreted as measures of association between the independent and dependent variable (but as noted above they are not intuitive as effect measures). The problem comes when we move to multivariate models. This is because the scale of y_i^* is still not fully standardized: Differences in the absolute size of the error variances are neutralized, but the scale can vary between men and women because of the size of the variance in the predicted values, caused by different effects of, and different variances in, the independent variables in the model. When the variance in predicted values increase, the total variance (and thus the scale) increases and the proportion unexplained variance decreases. With more than one independent variable in the model, all these affect the scale of the dependent variable and hence the coefficient for any given variable.

For example, say that we compare the coefficients β_1 and β_{1R} in Equations 3 and 6 above, with $var(xb)$ and $var(xb)_R$ being the variances in the predicted values. Then, still under the assumption that the logistic functional form is correct:[9]

$$\frac{\beta_1}{\beta_{1R}} = \sqrt{\frac{3.29 + var(xb) - var(xb)_R}{3.29}} \tag{21}$$

So we reach the same conclusion as above: The coefficients increase in response to changes in the proportion unexplained variance. The substantive implications of unobserved heterogeneity are thus the same whether we have a latent variable perspective or not. However, my experience is that the latent variable motivation for logistic regression tends to lead to a reification of the unobserved variable and its arbitrary scale. This often makes us forget what the actual information content of the model is, and what the results mean in terms of the probability of having the outcome of interest.

8 Conclusions

This article has provided an intuitive explanation of what unobserved heterogeneity in logistic regression is, when it matters and how. I have demonstrated how LnOR can be seen as consisting of three different components, and that these components have different information about the association and the model. Laying out these different parts, we can see that unobserved heterogeneity can be measured, and its impact on LnOR or OR can be straightforwardly assessed if we so wish.

[9] We can note that this is slightly different from the relation assumed by y-standardization. Y-standardization is discussed in Mood (2010) but with an error in the definition: On p.73 in Mood (2010) the definition of the standard deviation in y* should read "The total estimated sdY is the square root of (1) the variance of the predicted logits and (2) the assumed variance of the error term (which is always 3.29)", meaning that the assumed relation is: $\frac{\beta_1}{\beta_{1R}} = \sqrt{\frac{3.29+var(xb)}{3.29+var(xb)_R}}$

The total variance in a binary variable is a direct function of the mean, and unexplained variance is – under the assumption that the logistic functional form is correct – the same as the average derivative at the predicted probabilities of the model. This means that the mean of the dependent variable and the spread of predicted probabilities around this mean gives us the information we need to understand why and when unobserved heterogeneity matters for LnOR and OR in logistic regression. Thus, the fundamental limitation of binary variables, that their means and variances are not separately identified, is in fact the key to understanding unobserved heterogeneity in logistic regression.

In this perspective, it also becomes clear that the robustness of AME to unobserved heterogeneity uncorrelated to the respective variable is an intrinsic feature of the model, because what the estimation does is to strive to reproduce the observed (fixed) distributions of the dependent variable across the independent variables. If AME are found not to be robust to inclusion of uncorrelated controls, it means that the functional form is severely misspecified in at least one of the models and one should then seek to re-specify the model. In practice, however, logistic regression is robust even under rather extreme misspecifications (Cramer, 2007), and the robustness of AME is exactly what forces the log-odds ratios or odds ratios to be non-robust, because they are the quantities that the model can change in order to match the given observed distribution when information about other independent variables is added to the model.

Building on this understanding we could easily rescale a given LnOR to reflect what it would be under a different base rate, a different level of unexplained variance, or a different level of unexplained variance due to all variables but the one of interest. However, I see these counterfactual exercises primarily as a way of making unobserved heterogeneity in logistic models visible and understandable. Even if we adjust for unobserved heterogeneity, the log-odds or the odds scales are seldom the ones of ultimate interest, nor are they intuitive. The choice of scale for reporting of effects should be based on how it corresponds to the question we have, and on how the answers to our research questions are best expressed in order for their substantive meaning to be understood by ourselves and our readers. For most research questions, measures based on effects on the probability are to prefer over LnOR and OR (cf. Greenland, 1986). AME are easy to interpret and robust to omitted variables in the same way as coefficients in linear regression, and it is normally a good standard measure that gives an overall summary of the effect size on the scale of ultimate interest. In addition, exploring the effects on probabilities for different type cases or under different counterfactual distributions of independent variables can give a rich and thorough understanding of the substantive meaning of nonlinear relationships, something that is often hard to convey with a single OR or LnOR.

I do not suggest that OR and LnOR should never be reported. For some research questions they may be appropriate, but in these cases it is crucial not to apply the logic of interpretation from linear regression. For example, a larger LnOR or OR for a given variable in one group than in another may reflect a larger effect of this variable but also a larger effect of some other variable in the model. Clearly, this is in many cases an undesirable feature of an effect estimate, and in cases where OR

are deemed appropriate one should be clear about what drives any differences in OR across groups.

Finally, the fact that it is extremely common to misinterpret OR as relative risks (ratios of probabilities) means that anyone reporting OR must take great care to safeguard against this misunderstanding. I am inclined to agree with Schwartz, Woloshin, and Welch (1999) that OR are "bound to be interpreted as risk ratios", because this misinterpretation is so prevalent also among experienced researchers. Reporting of OR as relative risks (RR) (e.g., using wordings such as "times more likely") is very common in several disciplines[10], in spite of recurring critique of this practice (e.g. Davies, Crombie, and Tavakoli, 1998; Osborne, 2006; Deeks, 1998; Sackett, Deeks, and Altman, 1996; Sinclair and Bracken, 1999; Schwartz, Woloshin, and Welch, 1999) and in spite of most textbooks clearly pointing out the distinction between OR and RR.[11] This is a sign that even if OR have nice mathematical properties, they are probably too hard to handle for the general empirical researcher.

9 Appendix

The test of whether differences in LnOR across model and groups are due to differences in unexplained variance can be straightforwardly done in a regression framework similar to the setup in Allison (1999), by adding a constant weight to all the independent variables. In this case, however, we would not let the model estimate this weight as Allison did, but define it in terms of the R^2 (Gini's).

Consider a set of nested logistic models:

(1) log-odds of tenure $= \alpha + \beta_1 articles_i$ (giving R_1^2)
(2) log-odds of tenure $= \alpha + \beta_1 articles_i + \beta_2 prestige_i$ (giving R_2^2)
(3) log-odds of tenure $= \alpha + \beta_1 articles_i + \beta_2 prestige_i + \beta_3 select_i$ (giving R_3^2)

To adjust for differences in unexplained variance, we would construct the following weights:

$w1 = (1 - R_3^2)/(1 - R_1^2)$
$w2 = (1 - R_3^2)/(1 - R_2^2)$

And then re-run the models with these weights:

(1) log-odds of tenure $= \alpha + \beta_1(articles_i \times w1)$
(2) log-odds of tenure $= \alpha + \beta_1(articles_i \times w2) + \beta_2(prestige_i \times w2)$
(3) log-odds of tenure $= \alpha + \beta_1 articles_i + \beta_2 prestige_i + \beta_3 select_i$

[10] Some recent examples of this practice in high-profile papers is Fryer (2016) [Economics], Sandefur (2015) [Sociology], Trinquart, Johns, and Galea (2016) [Epidemiology] and O'Brien and Klein (2017) [Psychology]. These are just random examples that I have stumbled upon, and my reference to them does not mean that the problem is more severe in these papers than in others, and I do not judge the overall merit of these papers

[11] To argue that "times more likely" can refer to odds as well as probabilities (e.g., DeMaris 1993) is a mere play with words encouraging deceptive reporting that inflates the perceived effect sizes. OR and RR are not equivalent and it is extremely unlikely that anyone would interpret "times more likely" as ratios of odds rather than ratios of risks, so a responsible researcher should shun such terminology.

Now consider comparing the same model for men and women:

log-odds of tenure$_w$ $= \alpha + \beta_1 articles_i + \beta_2 prestige_i + \beta_3 select_i$ (giving R^2_w)

log-odds of tenure$_m$ $= \alpha + \beta_1 articles_i + \beta_2 prestige_i + \beta_3 select_i$ (giving R^2_m)

We construct a weight for women: $w_w = (1 - R^2_m)/(1 - R^2_w)$

And we re-run the model for women with weights:

log-odds of tenure$_w$ $= \alpha + \beta_1(articles_i \times w_w) + \beta_2(prestige_i \times w_w) + \beta_3(select_i \times w_w)$

We can also estimate a common model for men and women, showing us the statistical significance of the difference in LnOR for men and women:

log-odds of tenure$_w$ $= \alpha + \beta_{1m} articles_i + \beta_{2m} prestige_i + \beta_{3m} select_i + \beta_{1w}(articles_i \times w_w) + \beta_{2w}(prestige_i \times w_w) + \beta_3(select_i \times w_w) + \beta_4 female$

Acknowledgements

I thank Jan O Jonsson, Georg Treuter, Martin Hällsten, Per Engzell and Richard Breen for helpful comments.

References

Aldrich, John H and Forrest D Nelson (1984). *Linear probability, logit, and probit models*. Vol. 45. Sage.

Allison, Paul D. (1999). "Comparing logit and probit coefficients across groups". *Sociological Methods and Research* 28, pp. 186–208.

Angrist, J. D. (2001). "Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors: Simple Strategies for Empirical Practice". *Journal of Business Economic Statistics* 19, pp. 2–28.

Berkson, Joseph (1944). "Application of the logistic function to bio-assay". *Journal of the American Statistical Association* 39, pp. 357–365.

Cramer, J. S. (2007). "Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-specified Disturbances". *Oxford Bulletin of Economics and Statistics* 69, pp. 545–555.

Davies, HTO, IK Crombie, and M Tavakoli (1998). "When can odds ratios mislead?" *British Medical Journal* 316.7136, pp. 989–991.

Deeks, JJ (1998). "When can odds ratios mislead?" *British Medical Journal* 317, pp. 1155–1156.

Fryer, Roland G. Jr (2016). "An Empirical Analysis of Racial Differences in Police Use of Forc?" *NBER Working Paper No. 22399*.

Greenland, Sander (1986). "Interpretation and choice of effect measures in epidemiologic analyses". *American Journal of Epidemiology* 125, pp. 761–768.

Greenland, Sander, James M Robins, and Judea Pearl (1999). "Confounding and Collapsibility in Causal Inference". *Statistical Science* 14, pp. 29–46.

Hellevik, Ottar (2009). "Linear versus logistic regression when the dependent variable is a dichotomy". *Quality Quantity* 43.1, pp. 59–74.

Karlson, Kristian B., Anders Holm, and Richard Breen (2012). "Comparing Regression Coefficients Between Same-Sample Nested Models using Logit and Probit: A New Method". *Sociological Methodology* 42, pp. 286–313.

Long, J. Scott (2009). *Group Comparisons in Logit and Probit Using Predicted Probabilities. Working paper*.

Mittlböck, Martina and Michael Schemper (1996). "Explained Variation for Logistic Regression". *Statistics in Medicine* 15, pp. 1987–1997.

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, pp. 67–82.

O'Brien, Ed and Nadav Klein (2017). "The tipping point of perceived change: Asymmetric thresholds in diagnosing improvement versus decline". *Journal of Personality and Social Psychology: Attitudes and Social Cognition* 112, pp. 161–185.

Osborne, JW (2006). "Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses". *Practical Assessment Research Evaluation* 11, pp. 1–7.

Sackett, DL, JJ Deeks, and DG Altman (1996). "Down with odds ratios!" *Evidence-based Medicine* 1, pp. 164–166.

Sandefur, Rebecca L. (2015). "Elements of Professional Expertise: Understanding Relational and Substantive Expertise through Lawyers Impact". *American Sociological Review* 80.5, pp. 909–933.

Schwartz, Lisa M, Steven Woloshin, and H Gilbert Welch (1999). "Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization". *The New England Journal of Medicine* 341.4, pp. 279–283.

Sinclair, JC and M Bracken (1999). "Clinically useful measures of effect in binary analyses of randomized trials". *J Clin Epidemiol* 47, pp. 881–889.

Trinquart, Ludovic, David Merritt Johns, and Sandro Galea (2016). "Why do we think we know what we know? A metaknowledge analysis of the salt controversy". *International Journal of Epidemiology* 45.1, p. 251.

Williams, R. W. (2009). "Using Heterogeneous Choice Models To Compare Logit and Probit Coefficients Across Groups". *Sociological Methods Research* 37, pp. 531–559.