European Sociological Review VOLUME 26 | NUMBER 1 | 2010 67–82
DOI:10.1093/esr/jcp006, available online at www.esr.oxfordjournals.org
Online publication 9 March 2009

Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It

Carina Mood

Logistic regression estimates do not behave like linear regression estimates in one important respect: They are affected by omitted variables, even when these variables are unrelated to the independent variables in the model. This fact has important implications that have gone largely unnoticed by sociologists. Importantly, we cannot straightforwardly interpret log-odds ratios or odds ratios as effect measures, because they also reflect the degree of unobserved heterogeneity in the model. In addition, we cannot compare log-odds ratios or odds ratios for similar models across groups, samples, or time points, or across models with different independent variables in a sample. This article discusses these problems and possible ways of overcoming them.

Introduction

The use of logistic regression is routine in the social sciences when studying outcomes that are naturally or necessarily represented by binary variables. Examples are many in stratification research (educational transitions, promotion), demographic research (divorce, childbirth, nest-leaving), social medicine (diagnosis, mortality), research into social exclusion (unemployment, benefit take-up), and research about political behaviour (voting, participation in collective action). When facing a dichotomous dependent variable, sociologists almost automatically turn to logistic regression, and this practice is generally recommended in textbooks in quantitative methodology. However, our common ways of interpreting results from logistic regression have some important problems.[1]

The problems stem from *unobservables*, or the fact that we can seldom include in a model all variables that affect an outcome. Unobserved heterogeneity is the variation in the dependent variable that is caused by variables that are not observed (i.e. omitted variables).[2] Many sociologists are familiar with the problems of bias in effect estimates that arise if omitted variables are correlated with the observed independent variables, as this is the case in ordinary least squares (OLS) regression. However, few recognize that in logistic regression omitted variables affect coefficients also through another mechanism, which operates regardless of whether omitted variables are correlated to the independent variables or not. This article sheds light on the problem of unobserved heterogeneity in logistic regression and highlights three important but overlooked consequences:

(i) It is problematic to interpret log-odds ratios (LnOR) or odds ratios (OR) as substantive effects, because they also reflect unobserved heterogeneity.
(ii) It is problematic to compare LnOR or OR across models with different independent

variables, because the unobserved heterogeneity is likely to vary across models.

(iii) It is problematic to compare LnOR or OR 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.

Though some alarms have been raised about these problems (Allison, 1999), these insights have not penetrated our research practice and the interrelations between the problems have not been fully appreciated. Moreover, these problems are ignored or even misreported in commonly used methodology books. After introducing and exemplifying the behaviour of logistic regression effect estimates in light of unobserved heterogeneity, I outline these three problems. Then, I describe and discuss possible strategies of overcoming them, and I conclude with some recommendations for ordinary users.

The Root of The Problems: Unobserved Heterogeneity

In this section, I demonstrate the logic behind the effect of unobserved heterogeneity on logistic regression coefficients. I describe the problem in terms of an underlying latent variable, and I use a simple example with one dependent variable (y) and two independent variables (x_1 and x_2) to show that the LnOR or OR of x_1 will be affected in two different ways by excluding x_2 from the model: First, it will be biased *upwards or downwards* by a factor determined by (a) the correlation between x_1 and x_2 and (b) the correlation between x_2 and y when controlling for x_1. Second, it will be biased *downwards* by a factor determined by the difference in the residual variance between the model including x_2 and the model excluding it. For readers less comfortable with algebra, verbal and graphical explanations of the problem and extensive examples are given in the following sections.

One can think of logistic regression as a way of modelling the dichotomous outcome (y) as the observed effect of an unobserved propensity (or latent variable) (y^*), so that when $y^* > 0$, $y = 1$, and when $y^* < 0$, $y = 0$. The latent variable is in turn linearly related to the independent variables in the model (Long, 1997, section 3.2). For example, if our observed binary outcome is participation in some collective action (yes/no), we can imagine that among those who participate, there are some who are highly enthusiastic about doing so while others are less enthusiastic, and that among those not participating there are some who can easily tip over to participation and some who will never participate. Thus, the latent variable can in this case be perceived as a 'taste for participation' that underlies the choice of participation.

For simplicity, we assume only one independent variable. The latent variable model can then be written as:

$$y_i^* = \alpha + x_{1i}\beta_1 + \varepsilon_i \quad (1)$$

where y_i^* is the unobserved individual propensity, x_{1i} is the independent variable observed for individual i, α and β_1 are parameters, and the errors ε_i are unobserved but assumed to be independent of x_1. To estimate this model, we must also assume a certain distribution of ε_i, and in logistic regression we assume a standard logistic distribution with a fixed variance of 3.29. This distribution is convenient because it results in predictions of the *logit*, which can be interpreted as the natural logarithm of the odds of having $y = 1$ versus $y = 0$. Thus, logistic regression assumes the logit to be linearly related to the independent variables:

$$\mathrm{Ln}\left[\frac{P}{1-P}\right] = a + x_{1i}b_1 \quad (2)$$

where P is the probability that $y = 1$. For purposes of interpretation, the logit may easily be transformed to odds or probabilities. The odds that $y_i = 1$ is obtained by exp(logit), and the probability by exp(logit)/[1 + exp(logit)]. The logit can vary from $-\infty$ to $+\infty$, but it always translates to probabilities above 0 and below 1. Transforming the logit reveals that we assume that $y_i = 1$ if the odds is above 1 or, which makes intuitive sense, the probability is above 0.5. Results from logistic regression are commonly presented in terms of log-odds ratios (LnOR) or odds ratios (OR). Log-odds ratios correspond to b_1 in Equation 2, and odds ratios are obtained by e^{b1}, so LnOR gives the additive effect on the logit, while OR gives the multiplicative effect on the odds of $y = 1$ over $y = 0$. In other words, LnOR tells us how much the logit increases if x_1 increases by one unit, and OR tells us how many times higher the odds of $y = 1$ is if x_1 increases by one unit. While the effect on probabilities depends on the values of other independent variables in a model, LnOR and OR hold for any values of the other independent variables.

The total variance in y^* in Equation 1 consists of explained and unexplained variance. When we use Equation 2 to estimate this underlying latent variable model we force the unexplained (residual) part of the variance to be fixed. This means that any increases in

the explained variance forces the total variance of the dependent variable, and hence its scale, to increase. When the scale of the dependent variable increases, b_1 must also increase since it now expresses the change in the dependent variable in another metric. So because the residual variance is fixed, the coefficients in (2) estimate the effect on the dependent variable on a scale that is not fixed, but depends on the degree of unobserved heterogeneity. This means that the size of b_1 reflects not only the effect of x_1 but also the degree of unobserved heterogeneity in the model. This can be seen if we rewrite Equation 1 in the following way:

$$y_i^* = \alpha + x_{1i}\beta_1 + \sigma\,\varepsilon_i \tag{3}$$

where everything is as in Equation 1, except for the fact that the variance of ε_i is now fixed and the factor σ adjusts ε_i to reflect its true variance. σ is the ratio of the *true* standard deviation of the errors to the *assumed* standard deviation of the errors. Because we cannot observe σ, and because we force ε_i to have a fixed variance, b_1 in the logistic regression model (Equation 2) estimates β_1/σ and not β_1 (cf. Gail, Wieand and Piantadosi, 1984; Wooldridge, 2002: pp. 470–472).[3] In other words, we standardise the true coefficients β_1 so that the residuals can have the variance of the standard logistic distribution (3.29).

To further clarify the consequences of unobserved heterogeneity in logistic regression, consider omitting the variable x_2 when the *true* underlying model is

$$y_i^* = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \sigma\,\varepsilon_i \tag{4}$$

where ε_i is logistically distributed with a *true* variance of 3.29 (which means that $\sigma = 1$ in this case), and the relation between x_2 and x_1 is

$$x_{2i} = \gamma_0 + \gamma_1 x_{1i} + \nu_i \tag{5}$$

where γ_0 and γ_1 are parameters to be estimated and ν_i is the error term (which is uncorrelated to ε_i in Equation 4). The omission of x_2 from Equation 4 leads to two problems, one that is familiar from the linear regression case and one that is not. First, just as in linear regression, the effect of x_1 is confounded with the effect of x_2, so that when (5) is substituted into (4), the effect of x_1 becomes $\beta_1 + \beta_2\gamma_1$. That is, to the degree that x_1 and x_2 are correlated β_1 captures the effect of x_2.

The second problem, which does not occur in linear regression, is that the residual variance increases. In Equation 4, $\sigma = \sqrt{3.29}/\sqrt{3.29}$, that is, the assumed variance equals the true variance and β_1/σ is just β_1. When omitting x_2, the true residual variance becomes $\mathrm{var}(\varepsilon) + \beta_2^2\mathrm{var}(\nu)$, and as a consequence σ changes to $\sqrt{3.29 + \beta_2^2\mathrm{var}(\nu)}/\sqrt{3.29}$.

Taken together, these two problems imply that if we exclude x_2 from Equation 4, instead of β_1 we estimate:[4]

$$b_1 = (\beta_1 + \beta_2\gamma_1)\,\frac{\sqrt{3.29}}{\sqrt{3.29 + \beta_2^2\mathrm{var}(\nu)}} \tag{6}$$

If x_1 and x_2 are uncorrelated, Equation 6 collapses to

$$b_1 = \beta_1\,\frac{\sqrt{3.29}}{\sqrt{3.29 + \beta_2^2\mathrm{var}(x_2)}} \tag{7}$$

Therefore, the size of the unobserved heterogeneity depends on the variances of omitted variables and their effects on y, and LnOR and OR from logistic regression are affected by unobserved heterogeneity even when it is unrelated to the included independent variables. This is a fact that is very often overlooked. For example, one of the standard sociological references on logistic regression, Menard (1995, p. 59), states that 'Omitting relevant variables from the equation in logistic regression results in biased coefficients for the independent variables, to the extent that the omitted variable is correlated with the independent variables', and goes on to say that the direction and size of bias follows the same rules as in linear regression. This is clearly misleading, as the coefficients for the independent variables will as a matter of fact change when including other variables that are correlated with the dependent variable, even when these are unrelated to the original independent variables.

An Example Using Probabilities

The previous section has explained how logistic regression coefficients depend on unobserved heterogeneity. The logic behind this is perhaps most easily grasped when we express it from the perspective of the estimated probabilities of $y = 1$. Imagine that we are interested in the effect of x_1 on y and that y is also affected by a dummy variable x_2 defining two equal-sized groups. For concreteness, assume that x_1 is intelligence (as measured by some IQ test), y is the transition to university studies, and x_2 is sex (boy/girl). I construct a set of artificial data[5] where (a) IQ is normally distributed and strongly related to the transition to university, (b) girls are much more likely to enter university than are boys, and (c) sex and IQ are uncorrelated (all these relations are realities in many developed countries). I estimate the

following logistic models and the results are shown in Table 1:

Model 1: $y_i^* = \alpha + x_{1i}\beta_1 + \varepsilon_i$
Model 2: $y_i^* = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$
$R(x_1, x_2) = 0;\ P(y=1) = 0.5$

In Table 1, we can see that the LnOR and OR for IQ increase when we control for sex—*even though IQ and sex are not correlated*. This happens because we can explain the transition to university better with both IQ and sex than with only IQ, so in Model 2 the unobserved heterogeneity is less than in Model 1. The stronger the correlation between sex and the transition to university, the larger is the difference in the coefficient for IQ between Models 1 and 2.

The estimated odds of $y=1$ versus $y=0$ can be translated to probabilities (P) through the formula $P(y=1) = \text{odds}/(1+\text{odds})$. Figure 1 shows these predicted probabilities from Model 1 and Model 2 in Table 1, for different values of IQ. The figure contains the following curves: (1) the predicted transition probability at different values of IQ from Model 1; (2) the predicted transition probability at different values of IQ from Model 2, setting sex to its mean; (3) and (4) the predicted transition probability from Model 2 for boys and girls separately, and (5) the average predicted transition probability from Model 2 for small intervals of the IQ variable.

Table 1 Logistic regression estimates

	Model 1		Model 2	
	LnOR	OR	LnOR	OR
IQ	0.80	2.24	0.99	2.69
Sex			2.00	7.36
Constant	−0.01		−1.01	

In Figure 1, we can see that curves (1) and (2) look quite different. In addition, we can see that even though there is no interaction effect in Model 2, the predicted transition probability at different levels of IQ differs between boys and girls [curves (3) and (4)]. The reason that curves (3) and (4) differ is that boys and girls have different average transition probabilities (i.e. different intercepts in the logistic regression). Importantly, curves (2), (3), and (4) are parallel in the sense that for any point on the y-axis, they have the same slope, which means that *the curves represent the same OR and LnOR* (recall that the estimated LnOR and OR is constant for all values of IQ even though the effect on the transition probability varies). However, curve (1) is more stretched out, and thus represents a smaller OR and LnOR. Why is this?

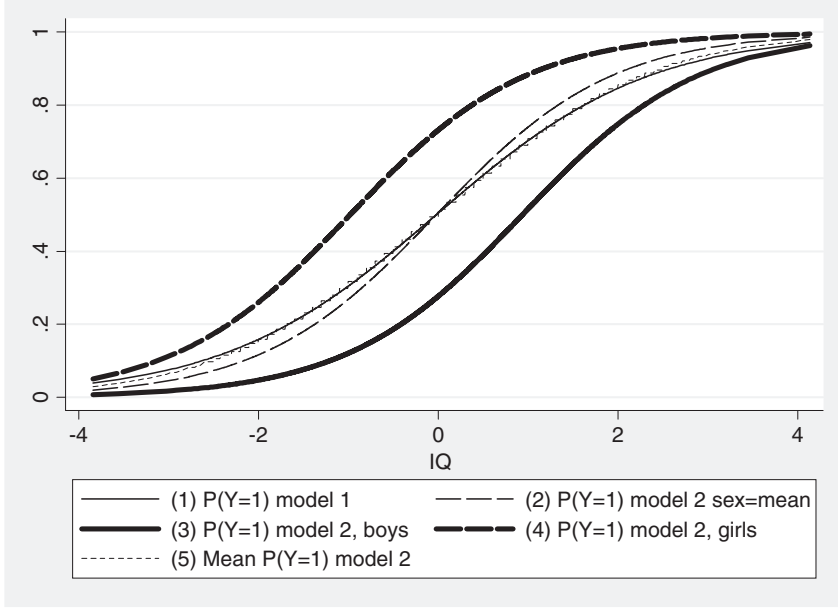

Figure 1 Predicted probabilities from Model 1 and Model 2, Table 1

Recall that both Model 1 and Model 2 are estimated using the same observed combinations of IQ and transitions. Even though boys and girls have different transition probabilities at different values of IQ, the *average* predicted transition probability among all students at different values of IQ should be approximately the same in Model 1 as in Model 2 (since IQ and sex are uncorrelated). This is confirmed by curve (5), which is the average for small intervals of IQ of the predicted transition probabilities that are described by curves (3) and (4).[6] So curve (1) approximately represents the average of curves (3) and (4) *on the probability scale*. However, the curve describing the average of the probabilities for boys and girls does *not* represent the average of the *OR* for boys and girls [curves (3) and (4)]. Curve (2), on the other hand, correctly represents the average of the OR of curves (3) and (4) (or rather, it represents the same OR as these curves). So Model 1 and curve (1) estimate transition probabilities at different values of IQ and the corresponding OR *for the population*, but for any single individual in the population the effects of IQ on the transition probability is better captured by Model 2 and curves (3) and (4). If expressed on the probability scale, the average of these curves [as represented by curve (1)] makes intuitive sense as an average in that it underestimates the transition probability for girls and overestimates it for boys. However, because curve (1) is more stretched out than curves (3) and (4), its OR is smaller, and this OR will therefore underestimate the OR *for all individuals*, i.e. for girls as well as for boys.[7]

Simulation of the Impact of Unobserved Heterogeneity

A fundamental question is how large effects we can expect from a given level of unobserved heterogeneity on the estimates in logistic regression. In order to show the impact of unobserved heterogeneity, I carry out 1,000 simulations of a sequence of logistic regressions. Each simulation constructs an artificial dataset ($n = 10{,}000$) where variables x_1 and x_2 are continuous variables drawn from normal distributions (command *drawnorm* in Stata 10.0). Both have a mean of 0 and a standard deviation of 1, and they are uncorrelated. y^* is determined by Equation 4, where α is invariantly set to 0, β_1 to 1, and ε_i follows a logistic distribution with a variance of 3.29. The size of the unobserved heterogeneity is varied by setting β_2 (the LnOR for x_2) to 0.5, 1.0, or 2.0, which are within the range of commonly observed effect sizes. Take note, however, that in this simulation, the unobserved heterogeneity is represented by a single variable, while in most real cases it will consist of a range of variables.

Table 2 Average estimate of β_1 for different values of β_2 and calculated estimate of β_1

	$\beta_2 = 0.5$	$\beta_2 = 1$	$\beta_2 = 2$
β_1, Model 1	0.95	0.84	0.61
	(0.03)	(0.03)	(0.03)
β_1, Model 2	1.00	1.00	1.00
	(0.03)	(0.03)	(0.03)
β_1, Equation 7	0.97	0.88	0.67

Standard deviations in parentheses. 1,000 replications, n = 10,000.

Table 2 reports the average of β_1 from these simulations, where Model 1 is the model excluding x_2 and Model 2 is the true model including x_2. The last row gives the attenuated estimate of β_1 as calculated by Equation 7.

We can see in Table 2 that the estimate of β_1 is clearly biased towards zero when not accounting for x_2, and this bias increases with the size of β_2. The table demonstrates that bias occurs already at relatively modest, and common, effect sizes and may become a considerable obstacle to interpretation of effects when important predictors are left out of the model. Recall, however, that the bias actually depends not only on β_2 but also on the variance in x_2 (which is invariantly set to one in this example). Table 2 also reveals that β_1 in this example is even more biased than Equation 7 would suggest. This is in line with the results in Cramer (2003), and the discrepancy can be explained by the misspecification of the shape of the error distribution in Model 1 (cf. Note 6).

Interpretation of Log-Odds Ratios and Odds Ratios as Effect Estimates

In Figure 1, we could see that curves (2), (3), and (4) had a common slope (i.e. their OR and LnOR were the same), but that curve (1) looked different. That is, the LnOR and OR describing the relation between x_1 and $P(y = 1)$ with x_2 set at fixed values [curves (2), (3), and (4)] are different from the LnOR and OR for the same relation averaged over the distribution of x_2

[curve (1)]. None of these curves is inherently 'wrong', but they estimate different underlying quantities. Curve (1) represents the *population-averaged* effect of x_1 on $P(y=1)$, while curves (2), (3), and (4) represent the effect of x_1 on $P(y=1)$ *conditional* on having a certain value on x_2. Hence, in the logistic regression *without* x_2 we obtain the LnOR or OR corresponding to a population-averaged probability curve, while the logistic regression *with* x_2 gives us the LnOR or OR corresponding to a conditional probability curve.

We can think of the conditional estimates as moving from the aggregate level in direction towards the individual-level effects, i.e. the effects that would occur for individuals upon a change in the independent variable. In terms of the above example, each individual must be either a boy or a girl and hence the OR or LnOR for IQ conditional on sex comes closer to the individual-level effect than the OR or LnOR for IQ from the bivariate model. Controlling for any other variable that improves our prediction of the university transition moves the OR for IQ even closer to the individual-level effect. As a consequence, estimates of a variable's LnOR or OR from a model averaging $P(y=1)$ over the distribution of some characteristic may be poor approximations of the same variable's LnOR or OR from a model conditioning on the characteristic, even if this characteristic is unrelated to the independent variable.

As a consequence of the above, when using logistic regression, we should be even more cautious to interpret our estimates as causal effects than we are when we use linear regression. It is difficult to control for all factors related to both independent and dependent variables, but it is of course even more difficult to control for all variables that are important for explaining the dependent variable. Considering the small degree of variance that we can usually explain, unobserved heterogeneity is almost always present. The problem need not be serious in all applications, but the risk that it is should be a cause for caution when interpreting results.

Recall that even if we do not know the *size* of the impact of unobserved heterogeneity unrelated to the independent variables, we always know the *direction* of the impact: it can only lead to an *underestimation* of the effect that one would estimate accounting for the unobserved heterogeneity. In addition, unobserved heterogeneity that is unrelated to the independent variables does not affect our possibilities to draw conclusions about the direction of an effect and the relative effect of different variables within a model (i.e. which variable has the largest effect) (Wooldridge 2002, p. 470).

Comparing Log-Odds Ratios or Odds Ratios Across Models in a Sample

In Table 1 above, we saw that the LnOR and OR for one variable (x_1) can change when we control for another variable (x_2), even though x_1 and x_2 are not correlated. This has the consequence that the common practice of 'accounting for' effects by including new independent variables in a model may lead researchers astray. In OLS regression we can start with a bivariate model of, e.g. the sex wage gap, and control for a range of variables to see to which extent these variables account for the wage gap. If we swap the dependent variable from wages to promotion and use logistic regression, we cannot use the same procedure, as the changes in coefficients across models can depend also on changes in unobserved heterogeneity.

Using another example, say that we estimate regional differences in unemployment, starting with a bivariate model, and in the second step controlling for education. If we do this on data from a country where there are no differences in average level of education between regions, and if education is relevant in explaining unemployment, the LnOR or OR for the region variable will suggest a larger effect of region when controlling for education. In a linear regression, this would be interpreted as education being a suppressor variable, meaning that the regional difference was partially suppressed by a higher level of education in regions with more unemployment. In logistic regression, we cannot draw such a conclusion.

Now imagine the same analysis on data from a country where there *are* regional differences in education. The change in the estimated LnOR or OR for the region variable will then depend *both* on the relation between region and education *and* on the change of scale induced by the reduction of unobserved heterogeneity when including education in the model. Because less-unobserved heterogeneity leads to larger effect terms, a downward change in the LnOR (towards 0) or in the OR (towards 1) for the region variable when education is added to the model cannot be due to a change of scale. This means that if the LnOR or OR for the region variable *decreases* in strength when adding education to the model, we can conclude that this is due to the relation between region and education. However, the shrinkage of the coefficient can be partly offset by an increase in explained variance, i.e. the decrease would have been larger in the absence of a change of scale. And if the coefficient

is unchanged, we may mistakenly conclude that regional differences in education are of no relevance for regional differences in unemployment.

Comparing Log-Odds Ratios or Odds Ratios Over Samples, Groups, or Time-Points

The fact that LnOR and OR reflect effects of the independent variables as well as the size of the unobserved heterogeneity does not only affect our possibility to draw conclusions about substantive effects and to compare coefficients over models. It also means that we cannot compare LnOR and OR across samples, across groups within samples (Allison, 1999), or over time, without assuming that the unobserved heterogeneity is the same across the compared samples, groups, or points in time. In most cases, this is a very strong assumption. Even if the models include the same variables, they need not predict the outcome equally well in all the compared categories, so different ORs or LnORs in groups, samples, or points in time can reflect differences in effects, but also differences in unobserved heterogeneity. This is an important point because sociologists frequently compare effects across, e.g. sexes, ethnic groups, nations, surveys, or years. The problem extends to comparisons across groups, time-points, etc. *within* one logistic regression model in the form of interaction effects. For example, in a model where we want to study the difference in the effect of one variable between different ethnic groups, between men and women, or between different years, the estimate will depend on the extent to which the model predicts the outcome differently in the different categories.

For example, we might be interested in how the effect of school grades on the probability of transition to university, controlling for the student's sex, varies over time or between countries. A weakening of LnOR or OR for grades between two points in time can mean that the effect of grades is diminishing over time, but it can also mean that the importance of sex for educational choice has decreased over time. Similarly, if the LnOR or OR for grades is higher in one country is higher than in another country, it can mean that the effect of grades is stronger, but it can also mean that sex is more important in explaining the educational choice in that country.

Proposed Solutions

A look in any sociological journal that publishes quantitative research confirms that the problem of unobserved heterogeneity has escaped the attention of the large majority of users of logistic regression. LnOR and OR are interpreted as substantive effects, and it is common practice to compare coefficients across models within samples, and across samples, groups etc. just as in linear regression. This is obviously problematic, so what should we do instead?

The easiest solution to the problem discussed here—or rather a way of avoiding it—is to replace the latent continuous variable with an observed continuous one; if a reasonable such variable exists. For example, if the outcome in question is a binary measure of poverty (coded 1 if poor and 0 if non-poor) one might consider using a continuous measure of economic hardship instead. This, however, is not always a feasible option, because we may be interested in dependent variables that are fundamentally qualitative, and for which no alternative continuous variables exist, such as mortality, divorce or educational transitions. Even so, there are ways around the problems, at least in some cases. These strategies could usefully be divided into those that concern odds ratios and log-odds ratios, and those that concern probability changes.

Solutions Using Odds Ratios and Log-Odds Ratios

The problem of comparing coefficients *across models* for the same sample but with different independent variables is discussed briefly by Winship and Mare (1984), who suggest that coefficients can be made comparable across models by dividing them with the estimated standard deviation of the latent variable (sdY*) for each model (*y*-standardization).[8] The total estimated sdY* is the sum of (1) the standard deviation of the predicted logits, and (2) the assumed standard deviation of the error term (which is always 1.81, i.e. $\sqrt{3.29}$). Because (2) is assumed to be fixed, all variation in the estimated sdY* across models must depend on (1), which in turn can depend only on the included independent variables. As explained above, the standard deviation of the logit (and hence its scale) increases when we include variables that improve our predictions of *y*. Because *y*-standardization divides coefficients by the estimated standard deviation, it neutralizes this increase and rescales coefficients to express the *standard-deviation-unit change in y** for a one-unit change in the independent variable. Note that

this method does in no way retrieve the 'true' scale of y^*: The size of the unobserved heterogeneity is still unknown, and the only thing we achieve is the possibility to compare a variable's coefficient between models estimated on the same sample with different control variables.

y-standardization works for comparisons across models estimated on the same sample because we know the size of the *difference* in unobserved heterogeneity across models. However, when comparing results for different groups, samples, points in time, etc., we do not know the differences between these in unobserved heterogeneity. What we do know is that unobserved heterogeneity that is unrelated to the independent variables affects all coefficients within a group etc. in the same way. Allison (1999) uses this fact to develop a procedure to test whether differences in coefficients across groups are due to unobserved heterogeneity, and his article contains a detailed description of this procedure using various programs. His procedure involves both a test of whether *at least one* coefficient differs across groups, and a test of whether a *specific* coefficient differs. As shown by Williams (2006a), the first test runs into problems if the effects of several variables in one group deviate in the same direction from the effects in the other group. In these cases, Allison's procedure can falsely ascribe group differences in effects to group differences in residual variation. In addition, the test of whether a specific coefficient differs requires an assumption that at least one of the underlying 'true' coefficients is the same in the compared groups. Of course, this assumption may be hard to justify, but if a researcher has strong theoretical reasons to make such an assumption, or have external evidence that supports it, and only wants to know whether an effect differs significantly across groups and in which direction, this test is likely to do the job.

The procedure proposed by Allison for group comparisons is criticized by Williams (2006a). He shows that it can be seen as one model in a larger family of so-called heterogeneous choice models, and argues that it will not be the right model for many situations. Instead, he proposes a more flexible use of the family of heterogeneous choice models to compare logit and probit coefficients across groups. Such models can be estimated in common statistical software, such as Stata [using the *oglm* command (Williams 2006b)], SPSS and Limdep, and are based on the idea of estimating one 'choice equation' that models the effect of a range of variables on the outcome, and one 'variance equation' that models the effect of a range of variables on the variance in the outcome. Because the mean and variance of categorical variables cannot be separately identified, models of this kind cannot really solve the problems discussed here. Rather, they can give alternative estimates under what are essentially different assumptions about the functional form of the relation (Keele and Park, 2006). For ordinal dependent variables, however, these models are more likely to be identified (Keele and Park, 2006), so if it is possible to express the outcome in ordinal rather than nominal terms this can be a good option.

As noted by Allison (1999), we can get a rough indication of whether differences in LnOR and OR depend on unobserved heterogeneity by simple inspection: If coefficients are consistently higher in one group, sample, model etc. than in another, it is an indication that there is less unobserved heterogeneity. However, because the true effects can also differ, this is not a foolproof test.

Solutions Using Probability Changes

Though the proposed solutions described above may allow comparability across models and in some cases across groups, the question remains as to how we should interpret the substantive effects. Angrist (2001) argues that problems with effect estimation from nonlinear limited dependent variable models (including logistic regression models) are primarily a consequence of a misplaced focus on the underlying latent variables. Instead, he thinks that we should care about the effects on probabilities (see also Wooldridge, 2002, p. 458). But how do estimates in probability terms fare in the context of unobserved heterogeneity?

The link between the logit and the probability of an outcome is the logistic cumulative distribution function (CDF):

$$F(\beta x_i) = \frac{\exp(\beta x_i)}{1 + \exp(\beta x_i)} \tag{8}$$

where βx_i is the value of the logit (i.e. the linear combination of values on variables x and their estimated coefficients β) for observation i. The slope of the logistic CDF is the logistic probability distribution function (PDF), which is given by:

$$f(\beta x_i) = \frac{\exp(\beta x_i)}{[1 + \exp(\beta x_i)]^2} \tag{9}$$

The CDF gives the $P(y_i = 1)$, and the PDF at a given $P(y_i = 1)$ equals $P(y_i = 1) \times [1 - P(y_i = 1)]$. Unlike the LnOR and OR, the effect of an independent variable on $P(y_i = 1)$ is not an additive or multiplicative constant, so there is no self-evident way of reporting results in probability terms. There are various measures

of changes in probability (Long, 1997, pp. 64–78), some of which are based on *derivatives*, which measure the slope at a particular point of the CDF, and others that measure the partial change in $P(y=1)$ for a *discrete change* in an independent variable from one specified value to another (e.g. a one-unit-change). As demonstrated by Petersen (1985), the derivative does not equal the change in $P(y=1)$ for a one-unit change in x, but it is often a decent approximation.

Marginal effects (MFX) at specified values (commonly means) of all independent variables take the value of the logistic PDF corresponding to the predicted logit for these specific values and multiply it by the estimated coefficient for x_1. MFX thus expresses the effect of x_1 on $P(y=1)$ *conditional on having the specified characteristics*. Another possibility is to evaluate the marginal effect at the logit corresponding to the average $P(y=1)$, i.e. conditional on having an average $P(y=1)$.[9] To get a better sense for the nonlinearity of the effect one can report marginal effects at different levels of independent variables or at different $P(y=1)$. When the specific values chosen for the point of evaluation are the means of independent variables, MFX is given by:

$$\beta_{x_1} f(\beta\bar{x}) \tag{10}$$

where $\beta\bar{x}$ is the value of the logit when all variables x have average values.

Another common measure based on derivatives is the *average marginal effect* (AME), which is given by

$$\frac{1}{n}\sum_{i=1}^{n} \beta_{x_1} f(\beta x_i) \tag{11}$$

where β_{x_1} is the estimated LnOR for variable x_1, βx_i is the value of the logit (i.e. the linear combination of values on variables x and their estimated coefficients β) for the i-th observation, and $f(\beta x_i)$ is the PDF of the logistic distribution with regard to βx_i. In words, AME expresses the average effect of x_1 on $P(y=1)$. It does so by taking the logistic PDF at each observation's estimated logit, multiplying this by the coefficient for x_1, and averaging this product over all observations.

Average partial effects (APE) differ from AME by averaging the marginal effects $\beta_{x_1} f(\beta x_i)$ across the distribution of other variables *at different given values of* x_1. While the AME gives one single estimate, APE varies by x_1 and thus acknowledges the nonlinear shape of the relation. APE differs from MFX in that MFX gives the marginal effect at given values of *all variables* in the equation, while APE gives the *average* effect at a given value of x_1 (or in a given interval of x_1). APE is thus also given by Equation (11), but it is calculated for subgroups with a specific value, or within a specific range of values, on x_1.

Wooldridge (2002, p. 471) shows for probit that if other variables are normally distributed and uncorrelated to x_1, the expectation of APE at different values of x_1 equals the MFX from the probit *without* these other variables. Similarly, under the assumption that omitted variables are unrelated to the included ones, the MFX at specified values of *several* included variables equals the APE (averaged over other unobserved or observed variables) among individuals who have this *combination* of values. For logistic regression Wooldridge offers no similar proof, but the intuition is the same: If other variables are uncorrelated to x_1, the APE of x_1 considering these variables and the MFX of x_1 disregarding these variables will be similar, and any difference depends on how much the probability distribution that the APE corresponds to deviates from a logistic distribution (cf. Figure 1 and note 6: The slope of curve (5) approximates APE, while the slope of curve (1) is the MFX from the model without x_2. Because curve (5) does not follow a logistic distribution, the curves are not identical).

Estimated changes in $P(y=1)$ for discrete changes in an independent variable (ΔP) (Petersen, 1995; Long, 1997) are normally evaluated taking the point of departure at a specific value of the logit, often a value corresponding to the averages of independent variables. In contrast to MFX, ΔP measures the change in $P(y=1)$ for a substantively meaningful change in x_1. For example, for a one-unit change in x_1 the ΔP would commonly be estimated as:

$$F(\beta\bar{x} + \beta_{x_1}) - F(\beta\bar{x}) \tag{12}$$

where $F(\beta\bar{x} + \beta_{x_1})$ and $F(\beta\bar{x})$ is the CDF of the logistic distribution with regards to $(\beta\bar{x} + \beta_{x_1})$ and $(\beta\bar{x})$, respectively. As with MFX, one can evaluate ΔP for logits corresponding to different values of the independent variables to better understand the nonlinearity of the relation.

To get a better sense of these different measures, Table 3 shows results from two logistic regressions on a fictitious data set.[10] AME is simply the mean of all individual derivatives $\beta_{x1} f(\beta x_i)$, MFX is evaluated at the means of independent variables (using the *mfx* command in Stata), APE is calculated as the mean of $\beta_{x1} f(\beta x_i)$ for the 153 (out of 20,000) observations that are in an interval of 0.01 standard deviations around the mean of x_1, and ΔP measures the change in the $P(y=1)$ for a one unit change in x_1 from ½ unit below average to ½ unit above average (using the *prchange*

command in the *Spost* routine in Stata). The following models are estimated:

Model 1: $y_i^* = \alpha + x_{1i}\beta_1 + \varepsilon_i$
Model 2: $y_i^* = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$
$R(x_1, x_2) = 0;\ P(y=1) = 0.5$

As can be seen in Table 3, MFX and ΔP follow the same pattern as the LnOR and OR: When controlling for x_2, these estimates suggest stronger effects of x_1 on y. AME and APE, however, are not affected more than marginally by controlling for x_2. The reason is that MFX and ΔP are conditional on specific values of the observed variables, while AME and APE represent averages of the conditional effects. Thus AME and APE are roughly invariant to the exclusion of independent variables unrelated to the independent variables already in the model [cf. Figure 1, in which curve (5)—that approximates APE—is close to curve (1)]. Recall that AME and APE average the *conditional* effects, which means that they are not invariant to the exclusion of independent variables that are correlated to the independent variables in the model.

Table 3 Comparison of estimates of β_1 with control for x_2 (Model 2) and without (Model 1) [$P(y=1)=0.50$]

	LnOR	OR	AME	APE	MFX	ΔP
Model 1	0.626	1.870	0.143	0.157	0.157	0.155
Model 2	1.005	2.732	0.142	0.151	0.251	0.246

The change in MFX and ΔP between Models 1 and 2 is a more complex issue than it may seem. It need not always be the case that these measures change in a way similar to the LnOR and OR. Recall that MFX at the mean of the independent variables is given by $\beta_{x_1} f(\beta\bar{x})$, and ΔP by $\mathrm{F}(\beta\bar{x} + \beta_{x_1}) - \mathrm{F}(\beta\bar{x})$, which means that the magnitude and direction of the change in these measures for x_1 when controlling for x_2 depends not only on the change in β_{x_1} (the LnOR for the effect of x_1 on y) but also on the size of $f(\beta\bar{x})$, i.e. the value of the logistic PDF at the point of evaluation (in this case, this point is at the average of the independent variables). This might appear clearer in Figure 2, which shows (1a) the predicted $P(y=1)$ for different values of x_1, not controlling for x_2 (the bold solid curve), (1b) the corresponding marginal effects at different values of x_1 (the bold dashed curve), (2a) the predicted $P(y=1)$ for different values of x_1, with x_2 held constant at 0, which is its mean value (the thin solid curve), and (2b) the corresponding marginal effects at different values of x_1 (the thin dashed curve).

MFX for x_1 without controls for x_2 (Model 1), the bold dashed curve, corresponds to the slope of the bold solid curve, while MFX for x_1 controlling for x_2

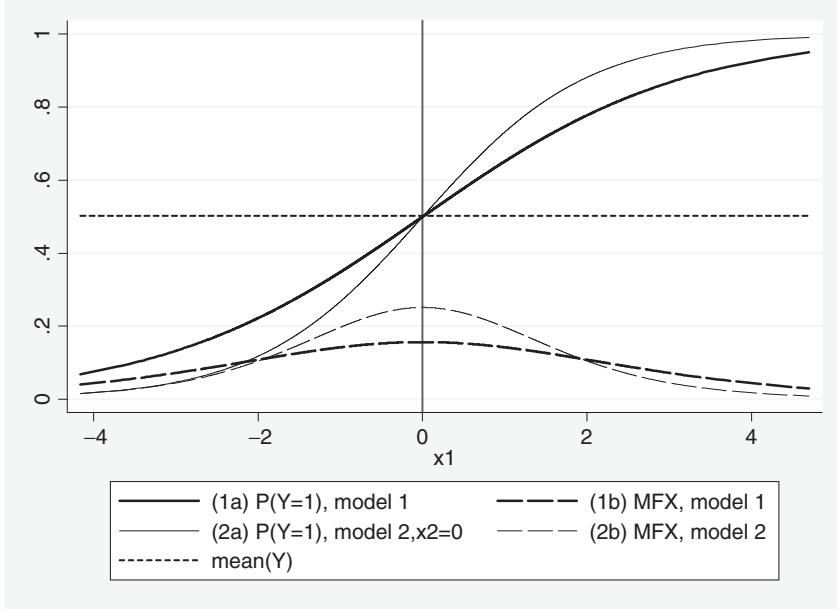

Figure 2 Predicted probability of $y=1$ by x_1, and corresponding marginal effects. Average probability of $y=1$ is 0.5

(i.e. Model 2), the thin dashed curve, corresponds to the slope of the thin solid curve. Because x_1 and x_2 are uncorrelated, the bold dashed curve also approximates APE. The solid curve changes shape when including x_2, because the more variables we control for, the better we can predict the outcome. In Table 3, MFX was evaluated at the average of x_1, which is 0. As we can see in Figure 2 (at the vertical line), this is where the slopes—the MFX—differ most. At other values of x_1, however, the differences between the MFX are smaller, and for very high or very low values of x_1 the MFX are actually larger when not controlling for x_2.

To understand how this happens, recall that the logistic PDF (i.e. the derivative of the CDF), has its peak when $P(y=1)$ is 0.5, which means that this is where the effect of an independent variable on $P(y=1)$ is strongest. The more unobserved heterogeneity we have, the more the scale of the logit shrinks towards zero. This means two things: (1) the coefficients shrink, and (2) the predicted logits shrink. Because a logit of 0 equals a probability of 0.5, the predicted probabilities move closer to 0.5 when the predicted logits shrink towards zero. As this is a move towards larger $f(\beta x)$, unobserved heterogeneity simultaneously shrinks β_{x1} and increases $f(\beta x)$. Though the shrinkage in β_{x1} is the same for all levels of x_1, the increase in $f(\beta x)$ varies with x_1, so more unobserved heterogeneity can for some values of x_1 lead to higher estimated MFX (Wooldridge, 2002; p. 471).

To further clarify the nature of MFX and ΔP, Table 4 gives results from a similar logistic regression of x_1 and x_2 on y, where $P(y=1)$ is 0.14 instead of 0.5.[11] The following models are estimated:

Model 1: $y_i^* = \alpha + x_{1i}\beta_1 + \varepsilon_i$
Model 2: $y_i^* = \alpha + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$
$R(x_1, x_2) = 0$; $P(y=1) = 0.14$

In Table 4 we can see that MFX and ΔP decrease when controlling for x_2. This is because these measures are now evaluated where the logistic PDF is further from its peak. For the models in Table 4, Figure 3 shows the same quantities as in Figure 2, that is: (1a) the predicted $P(y=1)$ for different values of x_1, not controlling for x_2 (the thick solid curve), (1b) the corresponding marginal effects at different values of x_1

Table 4 Comparison of estimates of β_1 with control for x_2 (Model 2) and without (Model 1) [$P(y=1)=0.14$]

	LnOR	OR	AME	APE	MFX	ΔP
Model 1	0.692	1.999	0.080	0.076	0.076	0.076
Model 2	1.002	2.723	0.078	0.077	0.044	0.045

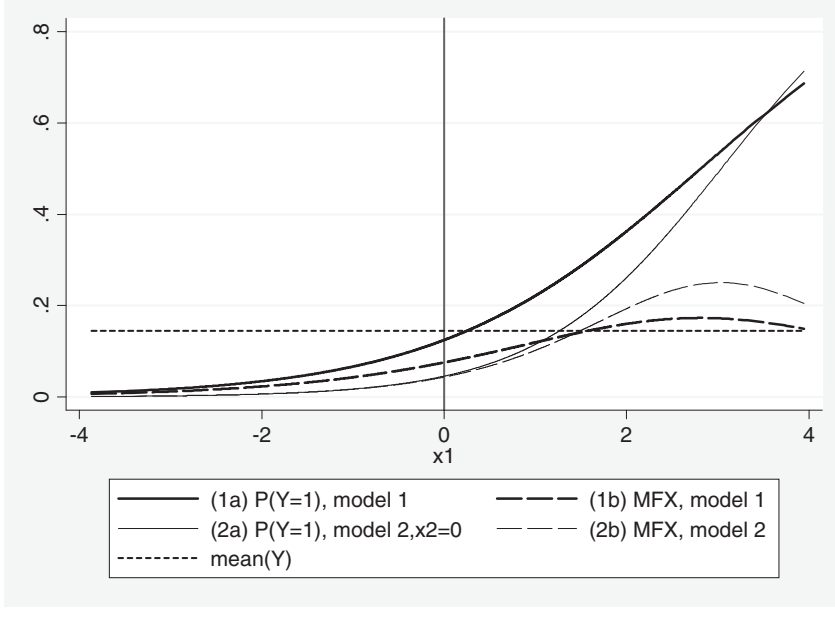

Figure 3 Predicted probability of $y=1$ by x_1, and corresponding marginal effects. Average probability of $y=1$ is 0.14

(the thick dashed curve), (2a) the predicted $P(y=1)$ for different values of x_1, with x_2 held constant at 0, which is its mean value (the thin solid curve), and (2b) the corresponding marginal effects at different values of x_1 (the thin dashed curve). As in Table 4, we see that at the point where MFX is evaluated (at $x_1=0$), the MFX (which is the slope of the solid curve predicting the probability) is *lower* when we control for x_2.

To conclude, AME and APE are not (at least not more than marginally) affected by unobserved heterogeneity that is unrelated to the independent variables in the model, and can thus be compared across models, groups, samples, years etc. However, these measures are population-averaged and represent the *average of the conditional effects* of x_1 on $P(y=1)$. AME gives the overall average of the conditional slopes for x_1, while APE corresponds to the average of the conditional slopes for different values of x_1. Though population-averaged estimates in probability terms are more intuitive in that they measure the average change in $P(y=1)$ and do not uniformly underestimate conditional effects as LnOR or OR do, they still estimate an average effect. If one is interested in the effect on an aggregate level, AME or APE is normally sufficient. Nonetheless, many are interested in the change in probability that would occur for individuals upon a change in the independent variable, and then MFX or ΔP evaluated at different values of independent variables are more appropriate because they come closer to this—especially ΔP as it measures the effect of a meaningful change in the independent variable. As noted above, these measures *are* affected by unobserved heterogeneity, and cannot be straightforwardly compared. In addition, they can suggest very different effects depending on the values chosen for the independent variable and the conditioning variables. Thus, a good option is often to report one estimate of probability change that is population-averaged, and one that is conditional.[12]

Linear Probability Models

Linear probability models (LPM), i.e. linear regression used with binary dependent variables, also yield results in terms of probability changes. In linear regression, we estimate the effects on the observed dependent variable, so coefficients are comparable over models, groups, etc. Using LPM is almost unthinkable in sociology, while it is common in economics. Generally, three problems are pointed out for LPMs:

(1) The possibility of predicted probabilities higher than 1 or lower than 0, i.e. out of range.
(2) Heteroscedastic and non-normal residuals, leading to inefficiency and invalid standard errors, and
(3) Misspecified functional form.

As noted by Long (1997), the occurrence of unrealistic predicted values as in (1) is also common in linear regression with non-binary dependent variables. This is not a serious problem unless many predicted values fall below 0 or above 1; and (2) can easily be corrected for.[13] This leaves (3) as the critical issue. It is often theoretically plausible that binary outcome variables are related to the independent variables in a non-linear fashion with smaller increments in the probability of the outcome at the extreme ends of the distribution.

As long as the misspecification of functional form does not alter (more than marginally) the substantive conclusions that are relevant to the questions asked, it is reasonable to choose LPM over logistic regression. Though LnOR or OR from logistic regression are valid for all values of the independent variable and acknowledge the non-linearity of the relation, these advantages are not exploited if substantive effects are reported only at a certain point of the distribution. For example, if we are only interested in sign and significance of an effect, or of an average effect estimate (such as AME) and not in the non-linearity of the relation *per se*, a LPM is entirely appropriate, and deriving AME from logistic regression is just a complicated detour.

In fact, the LPM effect estimates are unbiased and consistent estimates of a variable's average effect on $P(y=1)$ (e.g. Wooldridge, 2002; p. 454). For the data in Tables 3 and 4, the LPM coefficients are 0.143 and 0.080, respectively, which is identical to the AME for the same models. To get a more general picture of the equality of AME and LPM coefficients, I ran 1,000 simulations that fitted logistic regression and LPM to fictitious datasets ($n=5{,}000$) where x_1 and x_2 are continuous variables drawn from normal distributions, and y^* is generated by $\alpha+0.5x_{1i}+2x_{2i}+\varepsilon_i$, where ε_i follow a logistic distribution with a variance of 3.29, and α is set to vary between 0 and 6. Varying α in this way means that we obtain data with $P(y=1)$ ranging between 0.5 and 0.98 (because the logistic curve is symmetric, there is no need to consider $P(y=1)<0.5$). As before, Model 1 includes only x_1 and Model 2 includes both x_1 and x_2.

The results (see Table 5) are strikingly easy to summarize: Regardless of the overall $P(y=1)$, AME and LPM coefficients are identical or as good as identical. So if one seeks an estimate of the average effect, there appears to be no need to use logistic

Table 5 Average coefficients from LPM and AME from logistic regression

$P(y=1)$	$\alpha=0$ 0.50	$\alpha=1$ 0.65	$\alpha=2$ 0.77	$\alpha=3$ 0.87	$\alpha=4$ 0.93	$\alpha=5$ 0.97	$\alpha=6$ 0.98
Model 1							
AME x_1	0.074	0.069	0.056	0.039	0.024	0.013	0.006
	(0.007)	(0.006)	(0.006)	(0.005)	(0.004)	(0.003)	(0.002)
LPM x_1	0.074	0.069	0.056	0.039	0.024	0.013	0.006
	(0.007)	(0.006)	(0.006)	(0.005)	(0.004)	(0.003)	(0.002)
Model 2							
AME x_1	0.074	0.069	0.056	0.039	0.024	0.013	0.006
	(0.006)	(0.006)	(0.005)	(0.004)	(0.003)	(0.003)	(0.002)
LPM x_1	0.074	0.069	0.056	0.039	0.024	0.013	0.006
	(0.006)	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)	(0.002)
AME x_2	0.297	0.276	0.223	0.157	0.096	0.053	0.026
	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)
LPM x_2	0.298	0.277	0.223	0.157	0.096	0.053	0.026
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.003)

Standard deviations in parentheses. 1,000 replications, $n=5,000$.

regression. However, a problem with reporting one overall effect estimate is that it gives an impression that the relationship is linear. If the independent variables mainly vary over a range where the curve is roughly linear, this is of course not a problem, but if nonlinearity is substantively important, it can be misleading. If coefficients can be easily compared and appear easy to understand, but give a false impression of linearity, we have only exchanged one problem for another.

One argument sometimes levelled against LPMs is that their results as concerns interactions and non-linear transformations of the independent variables differ from those of logit and probit models for the same variables. This, however, should not be seen as a defect of LPMs, but as a simple reflection of the fact that interactive and non-linear transformations say something substantively different in a model where a nonlinear functional form is already inherently built into the model than in a model that assumes a linear relation. In other words, nonlinearities that are captured by the logistic functional form in the Logit model can be captured by non-linear transformations of the independent variables in a LPM.

Conclusions

Logistic regression is more complex than sociologists usually think. Standard textbooks in quantitative methods do not correctly reflect this complexity, and researchers continuously misunderstand and misreport effect estimates derived from logistic regression. Because coefficients depend both on effect sizes and the magnitude of unobserved heterogeneity, we cannot straightforwardly interpret and compare coefficients as we do in linear regression. Although these problems are mentioned in some econometric textbooks (e.g. Wooldridge, 2002) and may be known by sociologists specialized in quantitative methodology, the knowledge has hardly spread at all to practicing sociologists. If we do not consider these issues, we run great risk of drawing unwarranted conclusions from our analyses, and even to give wrong advice to policy-makers. The aim of this article is to present and discuss these problems in the context of logistic regression, but the problem is similar in other applications, such as probit models (Wooldridge, 2002, pp. 470–472) or proportional hazard models (Gail, Wieand and Piantadosi, 1984), methods also commonly used by sociologists.

To minimize problems at the stage of analysis and reporting, it is important to be aware of these problems already at the study planning and data collection stage. First, one should avoid collection of data in terms of dichotomies and qualitative variables if continuous (or at least ordinal) alternatives exist. Second, if the intention is to use logistic regression or some similar model using a non-linear link function, one must be careful to collect information on variables that are likely to be important for the outcome, even if these are likely to be only weakly, or not at all, related to the independent variables of interest.

There are no simple all-purpose solutions to the problems of interpretability and comparison of effect estimates from logistic regression. The situation is

Table 6 Characteristics of estimated effects on binary dependent variables

	Capture nonlinearity	Comparable across groups, samples etc.	Comparable across models	Conditional effect estimate[a]
Measures based on odds and log-odds				
Odds ratio	Yes	No	No	Yes
Log-odds ratio	Yes	No	No	Yes
y-standardization	Yes	No	Yes	No
Allison's procedure	Yes	Yes[b]	No	Yes
Heterogeneous choice models	Yes	Yes[c]	No	Yes
Measures based on percentages				
Average marginal effect	No	Yes	Yes	No
Average partial effect	Yes[d]	Yes	Yes	No
Marginal effect	Yes[d]	No	No	Yes
ΔP	Yes[d]	No	No	Yes
Linear probability model	No	Yes	Yes	No[e]

[a] *In a multivariate model.*
[b] *If assumption that one variable has same effect in groups etc. is correct.*
[c] *If assumption about the functional form of the relationship is correct.*
[d] *If estimated at several places in the distribution.*
[e] *If the true relationship is nonlinear.*

complicated by the fact that we often want estimates that simultaneously (i) capture the non-linearity of the relation, (ii) are comparable over groups, samples etc., (iii) are comparable over models, and (iv) indicate conditional effects. Because one estimate can normally not fulfil all these criteria, we need to carefully consider what is most relevant for our purposes and what we can estimate with available data. And, crucially, if our estimates do not fulfil all these criteria, we must report our results accordingly.

Table 6 provides a summary of the characteristics of the different effect estimates that have been discussed. Because different estimates fulfil different criteria, it is often advisable to report results using more than one type of estimate. I would also like to add a fifth criterion to the four above, namely that the estimates we report should be understandable to the reader. Many find log-odds ratios hard to grasp, and odds ratios are frequently misunderstood as relative risks, so it is often a good choice to present at least one effect estimate in terms of effects on probabilities.

Notes

1. For an accessible introduction to logistic regression, see Long (1997). I here discuss these problems in the context of dichotomous dependent variables, but they hold also for multinomial or ordinal logistic regression.
2. Unobserved heterogeneity is sometimes defined as unobserved differences between certain categories (e.g. men/women, treated/non-treated) or unobserved individual characteristics that are stable over time. In this article, I consider all variation that is caused by unobserved variables to be unobserved heterogeneity, and the problems that I discuss occur regardless of whether the unobserved variables are group-specific and/or stable over time.
3. The problems discussed also apply to coefficients from probit and most other models using non-linear link functions (Wooldridge, 2002, pp. 470–472; Gail, Wieand and Piantadosi, 1984). I concentrate on logistic regression here because it is used very frequently in sociology.
4. Equations 6 and 7 are only approximately true, because b_1 will be affected not only by the size but also by the shape of the residual variation, and this shape cannot be logistic both when including x_2 and when excluding it (cf. note 6).
5. The dataset ($n = 10{,}000$) is constructed in Stata 10.0. IQ (x_1) is a continuous variable (mean 0,

sd 1) drawn from a normal distribution (command *drawnorm*) and sex (x_2) is a dummy variable (mean 0.5, sd 0.5). y_i^* is generated by $-1 + 1x_{1i} + 2x_{2i} + \varepsilon_i$, where ε_i follows a logistic distribution with a variance of 3.29.

6. As can be seen in Figure 1, the probability curve from the bivariate logistic model (curve 1) does not perfectly represent the average of the conditional probability curves (3) and (4), but deviates systematically at very high and very low values of x_1. This is because the average of two logistic curves is not a logistic curve itself, but Model 1 is restricted by the assumption of the logistic distribution so that the predictions from it (curve 1) must take a logistic shape. This exemplifies an additional problem with logistic regression estimates: they can be affected not only by the size of the error but also by misspecification of the shape of the distribution. However, this problem appears minor relative to the one discussed in this article (cf. Cramer, 2003).
7. For the case of log-linear models and cross-tabular analysis the conflict between averages on the probability and on the OR scale has been discussed in terms of the collapsibility of OR over partial tables (Whittemore, 1978; Ducharme and Lepage, 1986).
8. *y*-standardized coefficients can easily be obtained in Stata using the *Spost* package (Long and Freese, 2005).
9. This is in fact a convenient way to roughly gauge the substantive meaning of results reported in terms of LnOR. Because the logistic PDF is simply $P(1-P)$, one can calculate the marginal effect at average $P(y=1)$ by taking a variable's LnOR $\times$ average $P(1-$average $P)$.
10. The dataset ($n=20{,}000$) is constructed in Stata 10.0. x_1 and x_2 are continuous variables (mean 0, sd 1) that are uncorrelated and drawn from a normal distribution (command *drawnorm*). y_i^* is generated by $x_{1i} + 2x_{2i} + \varepsilon_i$, where ε_i follows a logistic distribution with a variance of 3.29.
11. The dataset ($n=20{,}000$) is constructed in Stata 10.0. x_1 and x_2 are uncorrelated continuous variables (mean 0, sd 1) drawn from a normal distribution (command *drawnorm*). y_i^* is generated by $-3 + x_{1i} + 2x_{2i} + \varepsilon_i$, where ε_i follows a logistic distribution with a variance of 3.29.
12. As discussed in the section about LnOR and OR, the distinction between population-averaged and conditional estimates is really a matter of degrees. Estimates that are conditional on some variable are still averaged over the distribution of other variables. Thus, the MFX for x_1 in Model 2 in Tables 3 and 4 are conditional on x_2, but averaged over the distribution of variables that remain unobserved.
13. A simple way is to use heteroscedasticity-robust standard errors. A more efficient alternative is to estimate LPM by weighted least squares (WLS), where the observations with smaller residuals are given more weight. However, because the smaller residuals are for predicted values close to 0 or 1, which are the predictions most likely to be misleading in the LPM if the true relationship is nonlinear, WLS can be misleading when the true relationship is nonlinear and many predicted probabilities lie close to 0 or 1.

Acknowledgements

I thank Jan O. Jonsson, Martin Hällsten, Richard Breen, Frida Rudolphi, Robert Erikson, and two anonymous referees for helpful comments and advice. Financial support from the Swedish Council for Working Life and Social Research (FAS grant no 2006-0532) is gratefully acknowledged.

References

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research*, **28**, 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**, 2–28.

Cramer, J. S. (2003). *Logit Models from Economics and Other Fields*. Cambridge: Cambridge University Press.

Ducharme, G. R. and Lepage, Y. (1986). Testing collapsibility in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **48**, 197–205.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized

experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.

Keele, L., and Park, D. K. (2006). *Difficult Choices: An Evaluation of Heterogeneous Choice Models*. Paper for the 2004 Meeting of the American Political Science Association.

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

Long, J. S. and Freese, J. (2005). *Regression Models for Categorical Outcomes Using Stata*. TX: Stata Press.

Menard, S. (1995). *Applied Logistic Regression Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences 07-106. Thousand Oaks, CA: Sage.

Petersen, T. (1985). A Comment on presenting results from logit and probit models. *American Sociological Review*, **50**, 130–131.

Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40**, 328–340.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, **49**, 512–525.

Williams, R.W. (2006a). *Using Heterogeneous Choice Models To Compare Logit and Probit Coefficients Across Groups*. Working Paper, Notre Dame University.

Williams, R.W. (2006b). *OGLM: Stata Module to Estimate Ordinal Generalized Linear Models*. Working Paper, Notre Dame University.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

Author's Address

Swedish Institute for Social Research (SOFI), Stockholm University, SE-10691 Stockholm, Sweden. Email: carina.mood@sofi.su.se