Annual Review of Sociology

Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models

Richard Breen,[1] Kristian Bernt Karlson,[2] and Anders Holm[3]

[1]Nuffield College and Department of Sociology, University of Oxford, OX1 1NF, Oxford, United Kingdom; email: richard.breen@nuffield.ox.ac.uk

[2]Department of Sociology, University of Copenhagen, DK-1353 Copenhagen, Denmark

[3]Department of Sociology, University of Western Ontario, London, Ontario N6A 5C2, Canada

Annu. Rev. Sociol. 2018. 44:4.1–4.16

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-073117-041429

Keywords

logit, probit, KHB method, *Y*-standardization, marginal effects, linear probability model, mediation

Abstract

Methods textbooks in sociology and other social sciences routinely recommend the use of the logit or probit model when an outcome variable is binary, an ordered logit or ordered probit when it is ordinal, and a multinomial logit when it has more than two categories. But these methodological guidelines take little or no account of a body of work that, over the past 30 years, has pointed to problematic aspects of these nonlinear probability models and, particularly, to difficulties in interpreting their parameters. In this review, we draw on that literature to explain the problems, show how they manifest themselves in research, discuss the strengths and weaknesses of alternatives that have been suggested, and point to lines of further analysis.

INTRODUCTION

Sociologists and other social scientists often use the logit or probit model when an outcome variable is binary, an ordered logit or ordered probit when it is ordinal, and a multinomial logit when it has more than two categories. However, empirical applications of these nonlinear probability models (NLPMs) seldom take account of a body of work that, over the past three decades, has pointed to their problematic aspects and, particularly, to difficulties in interpreting their parameters.

The problems stem from the fact that, unlike in linear regression models, in logits, probits, and other NLPMs, the mean and variance of the dependent variable are not separately identified. This has two immediate implications: First, when comparing the coefficients from the same NLPM fitted to two or more groups, we need to be cautious in how we interpret them, and we need to be aware of how the assumptions on which the model is based will affect these interpretations. Second, the same holds for comparisons of the coefficients of the same variable in two or more differently specified models fitted to the same sample. An example of the first case—comparisons of the coefficients from the same model fitted to different groups—would arise if we wanted to know whether a particular variable had a stronger effect on an outcome among men rather than women, or Asians rather than Whites. The second case—comparisons of the coefficients from different models fitted to the same sample—could occur if we wanted to know how the coefficient for a particular variable changed when we added possible confounders or mediators to our model. Comparisons of both sorts continue to be made routinely, apparently with little thought for whether the conclusions drawn are warranted, despite the cautions expressed in some recent papers by sociologists (Allison 1999, Mood 2010, Karlson et al. 2012) that, to a large extent, echo warnings made as far back as the early 1980s by sociologists, economists, and statisticians (Lee 1982; Winship & Mare 1984; Gail et al. 1984; Yatchew & Griliches 1985; Gail 1986; Hauck et al. 1991; Neuhaus et al. 1991; Robinson & Jewell 1991; Swait & Louviere 1993; Cameron & Heckman 1998; Agresti 2002; Wooldridge 2002; Ai & Norton 2003; Cramer 2003, 2007; Mare 2006; Breen et al. 2013, 2014; Breen & Karlson 2013). Beyond these two simple examples, the problems we mention complicate the path decomposition of effects into direct and indirect components and severely complicate the interpretation of causal effects in both experimental and observational settings.

In this review, we draw on the literature of the past 30 years to explain the problems of NLPMs, show how these problems manifest themselves in a range of empirical research settings, discuss the strengths and weaknesses of alternatives that have been suggested, and point to what we believe are fruitful lines of further analysis. We begin by explaining the relevant differences between linear models and nonlinear probability models.

NONLINEAR PROBABILITY MODELS

We use the term NLPM to refer to the class of regression models for discrete, dependent variables that make a nonlinear transformation to obtain a model that is linear in its parameters. Among the best known is the logistic response (logit) model, which specifies the conditional mean of a discrete outcome variable as a logistic function of covariates. The probit model is similar but uses the cumulative normal instead of the logistic.

NLPMs can be derived from two different perspectives that reflect a famous controversy in statistics involving Karl Pearson and his former student, George Udny Yule (Agresti 2002). The first perspective (championed by Yule) assumes that the outcome is genuinely categorical or truly discrete: In this case, the probability of an observation being in a particular category is expressed as a nonlinear function of a set of predictors. This is sometimes referred to as the transformational approach (Powers & Xie 2008). The second perspective, that of Pearson, assumes that the discrete

outcome variable is a partially observed continuous latent variable. This is sometimes referred to as the latent variable approach. To illustrate the two perspectives, consider the binary outcome of completing a four-year college degree. The first perspective would regard completing college as a truly discrete event: A person either completes college or they do not. The second perspective might instead argue that individuals have a latent propensity to acquire education, and some have a propensity great enough to complete college while others do not.

Although the two perspectives differ, they are empirically indistinguishable. We begin by using the latent variable approach to explain the difference between these models and linear regression models, and we focus initially on models in which the observed outcome is binary rather than ordered or multinomial. We return later to the transformational approach.

Assume that we have a continuous outcome variable, Y^*, and one or more predictor variables, X. In a linear regression, we would write the relationship between these in the form

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i. \quad 1.$$

In Equation 1, ε is the error term. Using an estimator such as ordinary least squares (OLS), we would obtain estimates of the βs and the residual variance σ_ε^2. Now, assume we observe not Y^* but a binary variable, Y, that takes the value 1 if Y^* exceeds a threshold (say, τ) and 0 otherwise. We might then fit a logit or probit model to estimate the relationship between X and Y. Our choice implies a distribution for the error term of the original equation, Equation 1. For the probit model, we are implicitly assuming that ε follows a Normal distribution; for the logit, we are assuming that it follows a logistic distribution.

In either case, we would fit the model

$$b(\Pr(Y_i = 1)) = b_0 + b_1 X_i, \quad 2.$$

where h( ) indicates either the probit or logit transformation. The relationship between the parameters of the underlying linear model for Y^* and the parameters of the NLPM for Y is straightforward: $b = \frac{\beta}{s}$. The logit or probit coefficient is equal to the corresponding linear regression coefficient divided by s, a scale factor. The scale factor is defined as $s = \sigma_\varepsilon/\omega$, where σ_ε is the true standard deviation of the underlying linear model's error term and ω is an assumed standard deviation (1 in the Normal case and $\pi/\sqrt{3}$ for the logistic).[1] The scale factor thus expresses, in multiplicative form, the degree to which the true standard deviation of the underlying linear model differs from the standard deviation of the assumed standard distribution.

We first point out that in NLPMs, unlike in the linear model, there are not separate parameters for the coefficients and the residual variance; the coefficients themselves (the βs) are not identified separately from the residual variation (as captured by the scale factor, s). Sometimes the coefficients are said to be "identified only up to scale" (Cameron & Heckman 1998, p. 281). Second, we point out that interpretation of the bs rests entirely on assumptions about the true standard deviation, σ_ε, and these are untestable because in real applications, the only data we have are the binary Y and the predictor X (in other words, we do not know the true standard deviation of the error term in the underlying linear model).[2] The recent sociological literature on NLPMs (Allison 1999, Mood 2010, Karlson et al. 2012, Breen et al. 2014) has focused on the problems that this property

[1] Throughout this article, we use the term "true" to refer to the data-generating process, that is, the real-world process assumed to have generated the data we observe.

[2] To some extent this property has been hidden from view in books that introduce these models by the assumption of a scale parameter of unity, $s = 1$, which implies that $\sigma_\varepsilon = 1$ (for the probit) or $\sigma_\varepsilon = \pi/\sqrt{3}$ for the logit. For example, in his well-known book, Maddala (1983, p. 23) wrote of the probit model: "It can be easily seen ... that we can estimate only β/σ and not β and σ separately. Hence, we might as well assume $\sigma_\varepsilon = 1$ to start with."

of the models causes for comparisons of coefficients across different models. These coefficients can differ because the relationship between X and Y^* differs and/or because the residual standard deviations differ. When we compare the coefficients of different models fitted to the same data, it is inevitable that the models will differ in their residual standard deviation. When we compare estimates from the same model fitted to different groups, there is always the possibility that any differences we see are driven by differences between them in their residual variances. We deal in detail with both these cases later.

TRULY BINARY OUTCOMES

We have outlined the problems in interpreting the coefficients of NLPMs using the latent variable approach, but we can also think of the outcome, Y, as truly binary and not as the observed realization of a latent variable. In this case, the logit and probit transformations are just convenient ways of modeling the relationship between the outcome and the predictor variables. In the authors' experience, there is a widespread belief that, because the two perspectives make different assumptions about the nature of the dependent variable, they entail different consequences for applied research. In what follows, we demonstrate why we consider this belief unfounded (see Allison 1999 for a similar argument).

In the transformational approach, the equations for the logit or probit models have no explicit error term. Therefore, when we motivate their use through this approach, we must think of residual variation in these models in terms of omitted covariates. As before, Y denotes a binary, dependent variable and X denotes a predictor, but now we let U denote an omitted covariate independent of X. Under a probit model (with h being the probit transformation), we can write:

$$\Pr(Y = 1|X, U) = h^{-1}(\mu + \beta X + \gamma U). \quad 3.$$

We take this model to be the one that has generated the data. Because we do not observe U, the model we estimate is

$$\Pr(Y = 1|X) = h^{-1}(a + bX). \quad 4.$$

To see the relationship between b and β, we integrate the first equation over U. Zeger et al. (1988) showed that, under the assumption that γU follows a Normal distribution,[3] the relationship between the regression coefficients of X in Equations 3 and 4 is

$$b = \frac{\beta}{\sqrt{1 + \gamma^2 \text{var}(U)}}. \quad 5.$$

For the logit, a closed-form expression is not available, but Zeger et al. (1988) derive the approximate relationship between the estimated and true coefficients:

$$b \approx \beta \sqrt{\frac{\pi^2/3}{1.14\gamma^2 \text{var}(U) + \pi^2/3}} = \frac{\beta}{\sqrt{1 + 0.35\gamma^2 \text{var}(U)}}. \quad 6.$$

The denominator in Equations 5 and 6 is always larger than 1 (except in the instances where the omitted covariate has no impact on the outcome or is degenerate), and so b is a downwardly biased estimate of β (which we have assumed to be the true effect of X on Y). The degree of bias depends on the impact of the omitted covariate, γ, and the dispersion in the omitted variable, var(U). U is assumed to be independent of X, and so its omission cannot confound the estimate of β in

[3]The normality assumption is innocuous and serves only expository purposes. Derivations based on less restrictive assumption yield similar results (see Neuhaus et al. 1991).

the conventional sense. Consequently, omitting covariates that are independent of the predictor variable of interest will downwardly bias the estimated logit or probit coefficient. This is not the case in ordinary linear regression, where the inclusion of another predictor that is independent of the main predictor will have no effect on the latter's coefficient.

For completeness, we show briefly how the omitted variable approach works in the latent variable case. The data-generating process is assumed to be as follows:

$$\begin{aligned} Y^* &= \beta_0 + \beta_1 X + \gamma U + \varepsilon \\ Y &= 1 \text{ if } Y^* > \tau, \\ Y &= 0 \text{ otherwise.} \end{aligned}$$

As before, U is unobserved and independent of X and of ε, which we now assume to have either a standard Normal or standard logistic distribution. As before, the logit or probit model that we can estimate is

$$b(\Pr(Y_i = 1)) = b_0 + b_1 X_i,$$

where $b = \frac{\beta}{s}$, whereas the logit or probit model corresponding to the data-generating model is

$$b(\Pr(Y_i = 1)) = b'_0 + b'_1 X_i + b'_2 U_i,$$

where $b' = \frac{\beta'}{s'}$. The scale factors are the ratio of the residual standard deviation to the assumed standard deviation of the error, so we can approximate the relationship between b and b' as

$$b = \frac{b'}{\sqrt{1 + b'^2_2 \text{ var}(U)}} \qquad 7.$$

for the probit (Yatchew & Griliches 1985) and

$$b = \frac{b'}{\sqrt{1 + 0.304 \cdot b'^2_2 \text{ var}(U)}} \qquad 8.$$

for the logit (Breen & Karlson 2013). b is a downwardly biased estimate of b', and therefore, as before, also of β. If we use a probit model and assume the scale parameter of the original model equals unity, $s' = 1$, Equation 7 reduces to the bias for the transformational approach in Equation 5. Comparing Equations 6 and 8, we find that this relationship holds only approximately for the logit, although the difference between the two is substantively unimportant. Whether we use the latent variable approach or the transformational approach, we always recover a downwardly biased estimate of the effect of the variable of interest in the case of omitted variables, and the bias occurs even when the omitted covariates are unrelated to the predictor of interest.

In the remainder of this article, we give a nontechnical overview of the issues that arise in interpreting coefficients from NLPMs, providing references to publications containing relevant derivations, proofs, and simulations. For the sake of clarity, we focus on the logit and probit models for binary outcomes, but we stress that all the issues and problems we identify also apply to NLPMs for ordered outcomes (the ordered logit and ordered probit) and to the multinomial and conditional logit models for more than two categorical outcomes. Breen et al. (2014) and Breen & Karlson (2013) show the extensions to both cases.

COMPARISONS OF COEFFICIENTS ACROSS SAME-SAMPLE NESTED MODELS

Many sociological papers include a table that reports the results of a set of nested regressions, in the first of which the outcome, Y, is regressed on a single predictor variable of interest, X.

Subsequent models add control variables that are thought to be either confounders or mediators of the *X*-*Y* relationship. Attention focuses on how the addition of these controls changes the size of the coefficient of *X*. Often the goal is to discover how much this coefficient can be reduced and which control variables are chiefly responsible. This is a valid strategy when the models are linear, but when NLPMs are used, changes in the coefficient of *X* across nested models do not straightforwardly reflect mediation or confounding.

Regardless of whether we adopt the latent variable or transformational approach, adding variables to the model will have two effects: The residual variance will decline in magnitude (provided that the new variables have additional explanatory power) and the relationship between the outcome and the predictor variables already in the model will change (unless the new variables are orthogonal to them). In a linear model, these two effects are captured separately, through reduction in the variance of the residuals and change in the coefficients of the variables, but in NLPMs they are confounded. In, say, a logit model regressing a college completion dummy on parental socioeconomic status (SES), the addition of control variables will reduce the residual variance and thus also the scale factor, and this will make the SES coefficient grow larger. This will happen even if the controls are independent of SES. But suppose we include a measure of cognitive ability as one of our controls. Ability would be expected to affect the outcome and be correlated with SES. Although we would expect the inclusion of ability to reduce the SES coefficient (it might serve as both a mediator and a confounder), the decline in the scale factor (due to the reduction in the residual variance) will have the opposite effect. Thus, rather than a decline in the coefficient for SES, we might observe no change or even an increase. In general, estimates of change in a logit or probit coefficient caused by the introduction of control variables will be biased toward zero, and so one cannot directly compare coefficients from NLPMs across nested models fitted to the same sample. One manifestation of this will occur when the coefficient of the predictor of interest appears to be remarkably robust to the introduction of controls that, at first glance, should have caused it to decline.

Examples of such comparisons are easily found in most leading social science journals. One comes from Breen & Goldthorpe (1999). They used data on respondents to the British National Child Development Study and fitted a multinomial logit model with a person's social class position at age 33 as the outcome variable and dummy variables for social class origins as the predictors of interest. They then added controls for ability, measured at age 10, and effort, measured at age 16. Comparing the initial estimates for class origins with the conditional estimates, they write, "while there is some reduction in the . . . (class origin) parameters . . . this is rather variable and could not in any instance be described as more than modest. In other words, even when we control for both ability and effort . . . substantial inequalities in class mobility chances are still clearly in evidence" (Breen & Goldthorpe 1999, p. 14, parentheses added). But, because ability and effort are correlated with class origins and associated with class destinations, the conditional estimates on which Breen & Goldthorpe focused are certainly upwardly biased, leading them to understate the degree to which mobility chances are mediated by ability and effort (though we do not know how great this understatement is).

COMPARISONS OF COEFFICIENTS ACROSS GROUPS

The second problem on which attention has focused concerns comparisons of the coefficients of the same model fitted to different groups (this is equivalent to the problem of interpreting interaction effects in NLPMs; see Ai & Norton 2003). Allison (1999) lucidly explained this problem to sociologists. It arises because the residual variation, and thus the scale factor, may differ between groups. Suppose we fit our logit model for college completion to Whites and Asians separately,

with the aim of comparing the effect of SES. A naïve approach would look at how the logit of college completion changed, given a one-unit change in SES, in each group—in other words, compare b_{White} with b_{Asian} (b being the logit coefficient for SES). But, using the latent variable formulation of the logit and letting s denote the scale factor, we have

$$b_{\text{White}} = \frac{\beta_{\text{White}}}{s_{\text{White}}}$$
$$b_{\text{Asian}} = \frac{\beta_{\text{Asian}}}{s_{\text{Asian}}},$$

and so any difference that we observe may be due to differences in the real relationship between SES and college completion (captured by the βs) or in the scale factors.

The residual variance could differ between Whites and Asians for many reasons. For example, if omitted variables unrelated to SES are more important for completing college among Whites than among Asians, then the scale factors would differ. Whether this is true cannot be known from the available data (these data comprise the dummy Y, our measure of SES, and a dummy for group membership). So, whereas in the first problem (of comparing coefficients between different models) we know that the scale factors differ and we know the direction of the bias, in this case we do not. We do not know whether comparisons across groups are biased or not by differences in residual variation, and if they are biased, we do not know the extent or even the direction of the bias.

We can see the same problem using the transformational approach (Allison 1999, p. 190). Here, residual variation is captured by the effect of an omitted, orthogonal covariate. If we fit a probit model to the two groups to examine whether the effect of SES is stronger among one group or the other, then according to Equation 5,

$$b_{\text{White}} = \frac{\beta_{\text{White}}}{\sqrt{1 + \gamma^2_{\text{White}} \text{var}(U_{White})}}.$$

and

$$b_{\text{Asian}} = \frac{\beta_{\text{Asian}}}{\sqrt{1 + \gamma^2_{\text{Asian}} \text{var}(U_{\text{Asian}})}},$$

so any difference that we observe in the estimates could be due to differences between Whites and Asians in the impact, γ^2, or variance, var(U), of the omitted covariates, or both. The data do not allow us to separately identify these two sources of the difference.

Instances of coefficient comparisons between groups are very easy to find: Comparisons between men and women, Blacks and Whites, and different countries are routine in sociological research. For example, Breen et al. (2009) used an ordered logit model to investigate change in the relationship between class origins and educational attainment across five birth cohorts born during the twentieth century in eight European countries. They focus on within-country comparisons, so their estimates are open to bias from differences in residual variation in each cohort. In this case, however, reanalyses using methods that are robust to differences in residual variation (these are discussed below) lend support to their conclusion that the relationship between social class origins and educational attainment declined over the twentieth century (Breen et al. 2014).

INTERPRETING COEFFICIENTS

The literature we have drawn on generally assumes that the goal is to estimate the regression parameters of the assumed underlying model for the latent outcome, Y^*. In many cases, this

assumption is warranted, most obviously when using an ordered logit or ordered probit model. When we model the responses to a Likert-type attitudinal item, for example, we inevitably think in terms of an underlying continuous attitude. But there are many applications in which we are not concerned with any underling continuum. This is not to say that in such cases the binary outcome did not arise from a continuous latent variable (this is perhaps as much a philosophical as a sociological question). Rather, it is a matter of what quantity we want to estimate: Is it a parameter from an assumed latent variable regression or a measure of the probability of being in one outcome category rather than another? Does this choice have consequences for any of the problems we have reviewed?

Imagine we have data from a randomized controlled trial (RCT), which takes the form of a binary outcome Y and a treatment indicator X. A natural measure of the average treatment effect (ATE) is the odds ratio— in this case the ratio, between the treated and untreated, of a favorable versus an unfavorable outcome. The difficulties of interpreting odds ratios from RCTs have long been recognized (Gail et al. 1984, Gail 1986, Robinson & Jewell 1991, Hauck et al. 1991). Randomization ensures that estimates of the ATE are not confounded in linear models, but it is usually thought desirable to control for measured covariates to increase the precision of the estimate (Fisher 1935, Cox 1958). Including such covariates in a logit model, however, will increase the magnitude of the estimated odds ratio. Gail et al. (1984, p. 443) show that, for certain nonlinear models (including the logit), "randomization does not always lead to asymptotically unbiased estimates of treatment effects when needed covariates are omitted." Hauck et al. (1991, p. 77) called these variables "mavericks" because their omission biases the estimate of the odds ratio in an RCT, yet they are not conventional confounders. However, these covariates may not have been measured in the study. Put differently, estimates of the ATE in terms of odds ratios could differ between two populations simply because of differences in their composition not captured by the included covariates.

Statisticians draw a distinction between subject-specific (SS) and population-averaged (PA) effects: "The principal distinction between SS and PA models is whether the regression coefficients describe an individual's or the average population response to changing [the predictor] x" (Zeger et al. 1988, p. 1050). The distinction is also sometimes referred to as being between marginal (PA) and conditional (SS) models. The odds ratio from an RCT is a PA effect and, as Hauck et al. (1998, p. 253) point out, "When population-averaged models average, the results are necessarily averaged over the distribution of the omitted covariates in the trial. Two trials with the same treatment, outcome, and included covariates can have different measures of treatment effect solely because of differing distributions of the omitted covariates. It would be more scientifically sound to compare results for differing treatments if estimated treatment effects did not depend on the differing covariate distributions of the trials." Agresti (2002) makes the same point,[4] and Allison (1999, p. 191) expresses a widely held view about the usefulness of SS and PA estimates: "For purely descriptive purposes, comparison of population-averaged coefficients may be acceptable. But if the goal is to make inferences about causal relationships, a focus on subject-specific coefficients seems more appropriate."

PA and SS effects may be useful to researchers in different ways (Rodríguez 2008, 2015). To illustrate this, we return to the example of comparing the SES effect on college completion between populations—in this case, between two birth cohorts for which the availability of college differs as a result of reforms in tertiary education. We make four assumptions. First, the population in each cohort can be divided into two mutually exclusive groups of individuals: The first group has a

[4]Lee & Nelder (2004) similarly regard the conditional or SS model to be more fundamental (see also Allison 2009).

low preference for going to college, the second has a high preference, and we assume that college completion is more likely for individuals in the high preference group. Second, college preference is unrelated to SES in both cohorts. Third, the fraction in the high preference group is larger in the second cohort, perhaps as a result of educational reforms.[5] Fourth, the underlying SES effect expressed as an odds ratio for each preference group (the SS odds ratio) does not change over cohorts.

Under these four assumptions, the SS odds ratios are the same in the two cohorts, whereas the PA odds ratios (the unconditional odds ratio in each cohort) differ because of the changing distribution of the college preference. In this situation, changes in PA effects depend on factors unrelated to SES. The change in the PA odds ratios is descriptively true: The average college completion response to a unit change in SES would indeed have declined over time, but the effect of SES on college completion would nevertheless have stayed unchanged. The usefulness of the PA and SS odds ratios depends on the question that is being addressed, but it is important not to confuse the two.

SOLUTIONS: I. INTERPRETING COEFFICIENTS

Because NLPM coefficients are always attenuated (they are lower bounds to the true or underlying coefficients unless all relevant covariates are included), it is difficult to interpret their magnitude in substantive terms. We can, however, be confident of their sign, and their statistical significance is also unaffected by the attenuation bias (Breen & Karlson 2013). Furthermore, the sizes of coefficients of different covariates within the same NLPM can be compared because the attenuation bias is the same for all of them (Train 2009). For the same reason, ratios of coefficients from the same model will accurately reflect the relationship between the true or underlying coefficients.

Y-standardization has long been a popular approach among sociologists (McKelvey & Zavoina 1975, Winship & Mare 1984, Long 1997, Karlson 2015). *Y*-standardization refers to standardizing NLPM coefficients to make them interpretable in the same way as OLS regression coefficients that have been standardized on the outcome variable. The standardization in the NLPM is on the latent outcome, Y^*, not the observed, binary outcome:[6]

$$b^{\text{YSTD}} = \frac{\beta}{\text{sd}(Y^*)}.$$

We can recover this coefficient from the logit or probit model. Using Equation 1, we can write the variance of Y^* as

$$\sigma^2_{Y^*} = \beta_1^2 \text{ var}(X) + \text{var}(\varepsilon).$$

Bearing in mind that $\beta_1 = b_1 s$ and that $\varepsilon = s\omega$, the variance of Y^* is

$$\sigma^2_{Y^*} = s^2 b_1^2 \text{ var}(X) + s^2 \text{ var}(\omega) = s^2 \left[b_1^2 \text{ var}(X) + \text{var}(\omega) \right],$$

we can then recover the *Y*-standardized coefficient:

$$b_1^{\text{YSTD}} = \frac{\beta_1}{\text{sd}(Y^*)} = \frac{b_1 s}{\sqrt{s^2 [b_1^2 \text{ var}(X) + \text{var}(\omega)]}} = \frac{b_1}{\sqrt{b_1^2 \text{ var}(X) + \text{var}(\omega)}}. \quad 9.$$

[5] The reforms might make college completion less demanding and this may shift the preferences for some individuals toward college completion.

[6] The "*Y*" in *Y*-standardization refers to the latent propensity, which we denote Y^*. In our terminology, the name for the technique would be Y^*-standardization although we continue to refer to it as Y-standardization for consistency.

The *Y*-standardized coefficient involves only estimable or known quantities. It has the same interpretation as standardized coefficients in linear regression, and when the predictor variable of interest, *X*, is binary, the *Y*-standardized coefficient equals the effect size measure known as Cohen's *D* (Breen & Karlson 2013).

More recently, marginal effects have become popular, not least because they express relationships on the probability scale and so are easy to interpret. The marginal effect refers to the expected change in the probability of success, $\Pr(Y = 1)$, for a unit change in *X*. In the logit model, the individual marginal effect is

$$b_i^{\mathrm{ME}} = b_{\mathrm{logit}} \hat{p}_i \left(1 - \hat{p}_i\right), \qquad 10.$$

where b_{logit} is the coefficient estimated from a logit model and $\hat{p}_i$ is the predicted probability for the individual unit (indexed by i). From these individual effects, two marginal effects can be derived.[7] One is the marginal effect at the mean of the other covariates in the model. This is calculated by using the means of the covariates to estimate predicted probabilities and then using the formula in Equation 10 to calculate the marginal effect. The other is the average marginal effect or average partial effect (Wooldridge 2002). This is the average of the individual marginal effects (shown in Equation 10). The two measures often yield substantively similar results, but while the marginal effect at the mean is the effect for the possibly hypothetical person whose covariates all take the average values, the average marginal effect is the marginal effect on average, i.e., for a person picked at random. Thus, the marginal effect at the mean is conditional on holding covariates at their mean, while the average marginal effect is sensitive to the particular distribution of covariates in the population. Because marginal effects are unaffected by omitted covariates independent of the predictor of interest (Cramer 2007), they do not suffer from the attenuation bias described earlier.[8]

Both *Y*-standardized coefficients and average marginal effects provide readily interpretable effect estimates. Returning to the example of the effect of SES on college completion, a *Y*-standardized coefficient of 0.50 would tell us that, for a standard deviation change in SES, the expected change in the underlying college propensity is half a standard deviation. An average marginal effect of 0.1 would tell us that a standard deviation increase in SES increases the probability of completing college by 10 percentage points on average.

SOLUTIONS: II. COMPARING COEFFICIENTS ACROSS MODELS

Karlson et al. (2012) proposed the KHB (Karlson, Holm, and Breen) method as a way of assessing the degree to which the addition of controls changes the impact of a predictor of interest. The logic of the method is straightforward. We have two NLPMs: The first, or reduced, model includes only *X* as a predictor, and the second, or full, model adds another predictor, *Z*, presumed to be a confounder or mediator of the *X*-*Y* relationship. Its addition could change the coefficient for *X* through both confounding and rescaling, but we can identify the effect of rescaling by fitting a further model in which the predictors are *X* and the residuals from a linear regression of *Z* on *X*. By construction, this residualized *Z* will be orthogonal to *X* and therefore cannot confound the *X*-*Y* relationship. However, it will have the same conditional relationship with *Y* as the original *Z*, and so the model will have the same error term and thus same scale factor as the full model. After

[7] A third option is to draw the predicted probabilities as a function of covariates (Long 1997). This allows researchers to inspect possible nonlinearities that are neglected in the two marginal effect measures.

[8] Cramer (2007) also shows that marginal effects computed from the logit model are robust to deviations from the assumption of a logistically distributed latent error term.

identifying the effect of rescaling in this way, it then becomes possible to calculate the impact of confounding, net of rescaling. Breen et al. (2013) show how this can be used to carry out a path-analytic decomposition, into direct and indirect effects, of systems of equations estimated using NLPMs. The method is implemented in a user-written Stata routine called KHB (Kohler et al. 2011).

The KHB method, *Y*-standardization, and average marginal effects are all unaffected by rescaling, so they can be used to measure the change in the effect of the predictor of interest when controls are added. However, as Karlson et al. (2012) point out, this is not the only problem. If we assume that the latent error term of, say, the full model, has a logistic distribution, then the error of the reduced model cannot be logistic; its distribution will depend on the true distribution of the error in the full model and the distribution of the variable(s) omitted from the reduced model. In their simulations, Karlson et al. (2012, p. 300) find that their own method is unaffected by this (as they explain, "our model compares the full model with a reparameterization of the full model, thereby holding the error distribution constant across the full and reduced models"), whereas average marginal effects, *Y*-standardization, and the linear probability model (LPM) are affected. This is unlikely to be substantively important unless the difference in error distributions between the full and reduced models is large, but because these distributions are unobservable, we cannot know how great the difference is.

SOLUTIONS: III. COMPARING COEFFICIENTS BETWEEN GROUPS

Solutions to the problems in comparing NLPM coefficients across groups have been reviewed and assessed by Breen et al. (2014); they include the use of predicted probabilities (Long 2009) and several approaches already discussed (marginal effects, *Y*-standardization).

The naïve comparison of NLPM coefficients from different groups rests on the untestable assumption of a common residual variance. However, as Allison (1999) points out, other cross-group constraints will also serve to make the coefficients comparable: For example, if one coefficient was known to be equal in all groups, this would identify group differences in the other coefficients. To solve the cross-group comparison problem, Williams (2009) argues for using location-scale models (McCullagh 1980) in which the residual variances themselves are modeled. However, these models constrain the thresholds (identifying where the latent continuum is broken into distinct categories) to be equal across groups, and without strong theory, there is no reason to impose any such constraint. Thus, comparisons of NLPM coefficients between groups rely on assumptions that cannot be tested against the data.

Breen et al. (2014) show that one can recover the correlations between the predictors and the latent Y^* from the parameters of NLPMs, and so one can also recover the standardized regression coefficients. They discuss the circumstances under which these quantities might be useful in making comparisons between groups or populations.

THE LINEAR PROBABILITY MODEL

When logit and probit models were introduced to sociologists, they were often argued to be preferable to the more straightforward LPM because the LPM is heteroskedastic (though this is easily fixed using sandwich estimators) and the predicted probabilities are not bounded between 0 and 1. However, in recent years, because of its straightforward interpretation and also because of the problems with NLPMs we have discussed, the LPM has made a strong comeback, especially among economists (see, for example, Angrist & Pischke 2009). Because the LPM is a linear model, it does not suffer from any of the interpretational problems that the scale factor introduces in logit

and probit models. It models the probability of the outcome as a linear function of covariates and therefore consistently recovers the conditional expectation of the outcome (Greene 2011). Perhaps most importantly, when we use the LPM to analyze an RCT, regressing the dummy outcome on the dummy treatment indicator will return an unbiased estimate of the ATE measured on the probability scale.

LPM coefficients are closely related to average marginal effects derived from logit or probit models. In models including only a binary predictor, the two will be identical.[9] In models including multiple predictors or continuous covariates, they will differ, but often not by very much, because they use slightly different weighting schemes (Holm et al. 2015).

The many advantages of the LPM have made it an increasingly popular choice for modeling binary outcomes. In an RCT with a binary outcome, the LPM coefficient measures the difference in the probability of a positive outcome between treated and controls: It is an absolute effect measure. However, effects can also be stated in relative terms. The odds ratio is one example.[10] In contrast to LPM coefficients, the odds ratio depends on the baseline probabilities, and for changing baselines, these effect measures will also change even when these changes are unrelated to the treatment dummy. This will not be the case for the LPM and therefore, more generally, differences in LPM coefficients between populations or groups may not correspond to differences using odds ratios.

PRACTICAL SUGGESTIONS

Interpreting coefficients from NLPMs presents researchers with some serious challenges that do not exist for linear models. These challenges are not specific to whether one adopts a transformational or latent variable approach to NLPMs; in both cases, the magnitudes of their coefficients depend on omitted variables even when these variables are independent of the predictor variable of interest. For this reason, caution is called for in the interpretation of parameters from these models. Based on our review of the recent literature, we suggest the following six items of practical advice.

1. Researchers can always compare the magnitude of NLPM coefficients within the same model because the effects of omitted covariates operate the same way on all coefficients (that is, they are affected by a common scale factor).
2. The absolute magnitude of NLPM coefficients cannot be interpreted meaningfully because it depends arbitrarily on residual variation or omitted covariates. However, because the coefficients are always attenuated, their sign and statistical significance can be interpreted. If, for example, an estimated coefficient is positive and statistically significant, then we know that the true or underlying coefficient will also be positive and statistically significant.
3. For comparing NLPM coefficients across models fitted to the same sample, we recommend using the method of Karlson et al. (2012), which provides a general way of controlling for the effects of rescaling. The formulae and methods in Breen et al. (2013) can be used to

[9]More generally, in fully saturated models including only discrete covariates, the LPM and the logit model will yield the same estimates of the predicted probabilities because they both describe the data perfectly.

[10]Yet another relative measure is the risk ratio or relative risk, which is defined as the ratio between the conditional means of the binary outcome for treated and controls. In contrast to odds ratios, risk ratios are not affected by controlling for orthogonal confounders (Greenland et al. 1999). Odds ratios and risk ratios will be very close to each other for rare outcomes, and risk ratios can be recovered from odds ratios for common outcomes (see Greenland 2004). Risk ratios, however, are not symmetrical, meaning that their values will differ depending on whether one evaluates the conditional probability of success or of failure (Cummings 2009).

decompose total effects into direct and indirect effects. Another option would be to abandon NLPM coefficients altogether and instead use the LPM to obtain effects expressed on the probability scale, as these are unaffected by the rescaling issue and have a very straightforward interpretation.[11]

4. If we are concerned with empirical description, NLPM coefficients from the same model fitted to different samples can be compared: They are population averaged statistics. But if we care about effects, then the problem of comparing NLPM coefficients across models fitted to different samples has no satisfactory solution. Generalized NLPMs—known as location-scale or heterogeneous choice models—that allow the scale factor to vary between groups depend on untestable assumptions. For certain comparisons, the methods based on *Y*-standardization discussed in Breen et al. (2014) may be useful. LPMs can be used for group comparisons because the magnitude of LPM coefficients does not depend on orthogonal omitted covariates.
5. For RCTs with binary outcomes, *Y*-standardized coefficients will yield an effect size measure known as Cohen's *D*. The LPM will give effects on the probability scale.
6. When reporting and interpreting odds ratios, it is important to be aware of the distinction between marginal and conditional odds ratios. There is not one odds ratio—rather, there are many, depending on the covariates included in the model. Researchers should be explicit about whether their research question is better addressed with conditional (SS) or marginal (PA) odds ratios. For descriptive purposes, the latter may suffice, but when our interest is in underlying mechanisms, the SS odds ratios will be appropriate.

FURTHER RESEARCH

In recent years, nonparametric regression techniques have become a popular alternative to conventional regression techniques, particularly in economics. Nonparametric regression techniques model the conditional expectation of the outcome at local points in the distribution of the predictor variable, and visual displays are often used to show the functional relationship between two variables. These models can be extended to binary outcomes, in which case they model the conditional probability at local values of the predictor variable. Because these techniques make very few distributional assumptions, they might be well suited to make robust comparisons across groups. These methods may also potentially be combined with the KHB approach to yield robust analyses of mediation and confounding.

The distinction between PA and SS models originates in the statistical literature on NLPMs for repeated responses, that is, random effects or multilevel models (Neuhaus et al. 1991; Rodríguez 2008, 2015). These models and their applications in sociological research are reviewed in Guo & Zhao (2000). The models exploit the clustering of the data by individuals or any other larger unit to model the distribution of omitted covariates orthogonal to the predictor of interest. They therefore provide estimates of SS effects (where subject in SS refers to the clustering unit), whereas other approaches yield the PA effects (so-called generalized estimating equations). Thus, with clustered data, researchers can evaluate the difference between the two in a concrete, empirical study. This may be a fruitful research agenda in some sociological applications.

NLPMs appear in several techniques commonly employed by sociologists. In some cases, this is innocuous. The first stage of the widely used Heckman approach to address sample selection bias

[11]The method of Karlson et al. (2012) extends to average partial effects, meaning that the method can also yield effects on the probability scale.

usually involves estimating a probit model. Here, the scaling issue does not cause problems because the method uses the predicted probabilities, rather than the coefficient estimates, to correct for nonrandom selection. In contrast, conclusions from discrete time event history models are based on the interpretation of NLPM coefficients. Here, we are concerned with the parameters of the model for the underlying latent variable (in this case, time) but we estimate a set of (usually) logit models, one for each discrete time period, with the dependent variable being whether or not the transition in question occurred in that period, given that the transition had not already been made. It is widely appreciated that the successive loss of cases from the data can lead to bias in the coefficient estimates (Vaupel & Yashin 1985, Cameron & Heckman 1998, Rodríguez 2015) but it seems to be less appreciated that, for the same reason, the logit models in each period will differ in their residual variance (see Holm & Jaeger 2011). One implication is that comparing coefficients over time in discrete time event history models suffers from the same difficulties as comparisons of NLPM coefficients between different groups. This is an area that would reward further study.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to thank the *Annual Review of Sociology* Editors and reviewer for helpful comments on an earlier draft.

LITERATURE CITED

Agresti A. 2002. *Categorical Data Analysis*. New York: Wiley

Ai C, Norton EC. 2003. Interaction terms in logit and probit models. *Econ. Lett.* 80:123–29

Allison PD. 1999. Comparing logit and probit coefficients across groups. *Sociol. Methods Res.* 28:186–208

Allison PD. 2009. *Fixed Effects Regression Models*. Los Angeles, CA: Sage

Angrist J, Pischke J-S. 2009. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton Univ. Press

Breen R, Goldthorpe JH. 1999. Class inequality and meritocracy: a critique of Saunders and an alternative analysis. *Br. J. Sociol.* 50:1–27

Breen R, Holm A, Karlson KB. 2014. Correlations and non-linear probability models. *Sociol. Methods Res.* 43:571–605

Breen R, Karlson KB. 2013. Counterfactual causal analysis and non-linear probability models. In *Handbook of Causal Analysis for Social Research*, ed. SL Morgan, pp. 167–87. Dordrecht, Neth.: Springer

Breen R, Karlson KB, Holm A. 2013. Total, direct, and indirect effects in logit and probit models. *Sociol. Methods Res.* 42:164–91

Breen R, Luijkx R, Müller W, Pollak R. 2009. Nonpersistent inequality in educational attainment: evidence from eight European countries. *Am. J. Sociol.* 114:1475–521

Cameron SV, Heckman JJ. 1998. Life cycle schooling and dynamic selection bias: models and evidence for five cohorts of American males. *J. Political Econ.* 106:262–333

Cox DR. 1958. *Planning of Experiments*. New York: Wiley

Cramer JS. 2003. *Logit Models from Economics and Other Fields*. Cambridge, UK: Cambridge Univ. Press

Cramer JS. 2007. Robustness of logit analysis: unobserved heterogeneity and mis-specified disturbances. *Oxf. Bull. Econ. Stat.* 69:545–55

Cummings P. 2009. The relative merits of risk ratios and odds ratios. *Arch. Pediatr. Adol. Med.* 163:438–45

Fisher RD. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd

Gail MH. 1986. Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology*, ed. SH Moolgavkar, RL Prentice, pp. 3–18. New York: Wiley

Gail MH, Wieand S, Piantdosi S. 1984. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71:431–44

Greene WH. 2011. *Econometric Analysis*. Upper Saddle River: Prentice Hall. 7th ed.

Greenland S. 2004. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *Am. J. Epidemiol.* 160:301–5

Greenland S, Robins JM, Pearl J. 1999. Confounding and collapsibility in causal inference. *Stat. Sci.* 14:29–46

Guo G, Zhao H. 2000. Multilevel modeling for binary data. *Annu. Rev. Sociol.* 26:441–62

Hauck WW, Anderson SD, Marcus SM. 1998. Should we adjust for covariates in nonlinear regression analysis of randomized trials? *Control. Clin. Trials* 19:249–56

Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. 1991. A consequence of omitted covariates when estimating odds ratios. *J. Clin. Epidemiol.* 44:77–81

Holm A, Ejrnaes M, Karlson KB. 2015. Comparing linear probability models across groups. *Q. Quantity* 49:1823–34

Holm A, Jaeger MM. 2011. Dealing with selection bias in educational transition models: the bivariate probit selection model. *Res. Soc. Stratif. Mobil.* 29:311–22

Karlson KB. 2015. Another look at the method of *y*-standardization in logit and probit models. *J. Math. Sociol.* 39:29–38

Karlson KB, Holm A, Breen R. 2012. Comparing regression coefficients between same-sample nested models using logit and probit: a new method. *Sociol. Methodol.* 42:274–301

Kohler U, Karlson KB, Holm A. 2011. Comparing coefficients of nested nonlinear probability models. *Stata J.* 11:420–38

Lee L-F 1982. Specification error in multinomial logit models: analysis of the omitted variable bias. *J. Econom.* 20:197–209

Lee Y, Nelder JA. 2004. Conditional and marginal models: another view. *Stat. Sci.* 19:219–38

Long JS. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage

Long JS. 2009. *Group comparisons in logit and probit using predicted probabilities*. Work. Pap., Indiana Univ. **http://www.indiana.edu/~jslsoc/files_research/groupdif/groupwithprobabilities/groups-with-prob-2009-06-25.pdf**

Maddala GS. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge Univ. Press

Mare RD. 2006. Response: statistical models of educational stratification: Hauser and Andrew's models for school transitions. *Sociol. Methodol.* 36:27–37

McCullagh P. 1980. Regression models for ordinal data. *J. R. Stat. Soc. Ser. B* 42:109–42

McKelvey RD, Zavoina W. 1975. A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* 4:103–20

Mood C. 2010. Logistic regression: why we cannot do what we think we can do, and what we can do about it. *Eur. Sociol. Rev.* 26:67–82

Neuhaus JM, Kalbfleisch JD, Hauck WW. 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* 59:25–35

Powers DA, Xie Y. 2008. *Statistical Methods for Categorical Data Analysis*. Bingley, UK: Emerald. 2nd ed.

Robinson LD, Jewell NP. 1991. Some surprising results about covariate adjustment in logistic regression models. *Int. Stat. Rev.* 58:227–40

Rodríguez G. 2008. Multilevel generalized linear models. In *Handbook of Multilevel Analysis*, ed. Jan de Leeuw, Erik Meijer, pp. 335–76. New York: Springer

Rodríguez G. 2015. Multilevel models in demography. In *International Encyclopedia of the Social and Behavioral Sciences*, ed. JD Wright, pp. 48–56. Oxford: Elsevier. 2nd ed.

Swait J, Louviere J. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. *J. Mark. Res.* 30:305–14

Train K. 2009. *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge Univ. Press

Vaupel JW, Yashin AI. 1985. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *Am. Stat.* 39:176–85

Williams R. 2009. Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociol. Methods Res.* 37:531–59

Winship C, Mare RD. 1984. Regression models with ordinal variables. *Am. Sociol. Rev.* 49:512–25

Wooldridge JM. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press

Yatchew A, Griliches Z. 1985. Specification error in probit models. *Rev. Econ. Stat.* 67:134–39

Zeger SL, Liang K-Y, Albert PS. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44:1049–60