*In logit and probit regression analysis, a common practice is to estimate separate models for two or more groups and then compare coefficients across groups. An equivalent method is to test for interactions between particular predictors and dummy (indicator) variables representing the groups. Both methods may lead to invalid conclusions if residual variation differs across groups. New tests are proposed that adjust for unequal residual variation.*

# Comparing Logit and Probit Coefficients Across Groups

PAUL D. ALLISON
*University of Pennsylvania*

For binary dependent variables, logit (logistic) and probit regression have become standard methods of analysis. As with ordinary linear regression, researchers often estimate separate binary regression models for two or more groups of individuals and then compare coefficients across groups. Ideally, such comparisons are accompanied by statistical tests for the significance of the differences. An alternative procedure is to estimate a single model for all groups combined, with interactions between dummy (indicator) variables for groups and the variables of interest. Significant interactions indicate significant differences in coefficients across groups. If the model includes interactions for all the explanatory variables crossed with all the group dummies, the interaction method is equivalent to running separate regressions. (For some recent examples applying these methods, see Baxter [1994], Kalmijn [1994], Wright and Jacobs [1994], and Sekulic, Massey, and Hodson [1994].)

Unfortunately, there is a potential pitfall in cross-group comparisons of logit or probit coefficients that has largely gone unnoticed. Unlike linear regression coefficients, coefficients in these binary regression models are confounded with residual variation (unobserved hetero-

AUTHOR'S NOTE: *I am indebted to J. Scott Long, William Greene, Tim Futing Liao, S. Philip Morgan, Herbert Smith, Kazuo Yamaguchi, and several anonymous reviewers.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 2, November 1999 186-208

geneity). Differences in the degree of residual variation across groups can produce apparent differences in coefficients that are not indicative of true differences in causal effects. I will develop these ideas in some detail below. I will also propose a method for comparing logit or probit coefficients across groups while removing the confounding effects of residual variation.[1]

## *EXAMPLE: PROMOTIONS TO ASSOCIATE PROFESSOR*

To make these issues more concrete, I begin with an example. In Table 1, we see the results of logit regressions predicting the probability of promotion to associate professor for samples of 301 male and 177 female biochemists. These scientists received their doctorates in the late 1950s and early 1960s and were assistant professors at graduate departments in U.S. universities at some time during their careers. (For a detailed description of the data and its sources, see Long, Allison, and McGinnis [1993].)

For the regressions reported in Table 1, the units of analysis were person-years rather than persons, with 1,741 person-years for men and 1,056 person-years for women. As shown in Allison (1982), the likelihood function for this sort of data factors in such a way that the multiple observations per person are effectively independent. Hence, it is entirely appropriate to use ordinary logistic regression without any correction for dependence.

The explanatory variables used in these regressions are a greatly reduced subset of the variables considered in Long et al. (1993), and the results here differ somewhat from those in the original article. No substantive conclusions should be drawn from Table 1, or any of the other analyses reported here. In Table 1, duration is the number of years since the beginning of the assistant professorship, undergraduate selectivity is a measure of the selectivity of the college where scientists received their bachelor’s degrees (ranges from 1 to 7), number of articles is the cumulative number of articles published by the end of each person-year, and job prestige is a measure of prestige of the department in which scientists were employed (ranges from 0.65 to 4.60). For men, all the coefficients are statistically significant in the expected direction. The coefficients for women all have the same sign

**TABLE 1: Results of Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists**

| | *Men* | | *Women* | | | |
|---|---|---|---|---|---|---|
| *Variable* | *Coefficient* | SE | *Coefficient* | SE | *Ratio of Coefficients* | *Chi-Square for Difference* |
| Intercept | –7.6802*** | .6814 | –5.8420*** | .8659 | .76 | 2.78 |
| Duration | 1.9089*** | .2141 | 1.4078*** | .2573 | .74 | 2.24 |
| Duration squared | –0.1432*** | .0186 | –0.0956*** | .0219 | .67 | 2.74 |
| Undergraduate selectivity | 0.2158*** | .0614 | 0.0551 | .0717 | .25 | 2.90 |
| Number of articles | 0.0737*** | .0116 | 0.0340** | .0126 | .46 | 5.37* |
| Job prestige | –0.4312*** | .1088 | –0.3708* | .1560 | .86 | 0.10 |
| Log likelihood | –526.54 | | –306.19 | | | |

*$p$ < .05. **$p$ < .01. ***$p$ < .001.

as those for men, but one of them (undergraduate selectivity) was not significant at the .05 level.

As shown in the penultimate column of Table 1, the ratios of the coefficients for females to males are all substantially less than one. The last column reports the Wald chi-square statistic for testing the difference between coefficients for men and women. The formula for this statistic is

$$\frac{(b_M - b_W)^2}{[s.e.(b_M)]^2 + [s.e.(b_W)]^2}, \tag{1}$$

where $b_M$ is the coefficient for men, $b_W$ is the coefficient for women, and $s.e.(.)$ is the estimated standard error. Each statistic has 1 degree of freedom. The only variable whose coefficients are significantly different at the .05 level is number of articles. Apparently, the effect of number of articles on the log odds of being promoted is about twice as great for males as it is for females. Using the transformation $100(e^{\beta} – 1)$, we can say that each additional article yields an increase in the odds of promotion of about 8 percent for men and about 4 percent for women. If accurate, this difference suggests that men get a greater payoff from their published work than do females, a conclusion that many would find troubling.

## *MODELS FOR UNEQUAL RESIDUAL VARIATION*

I now argue that the difference in the two coefficients for article counts may be an artifact of differences in the degree of residual variation (unobserved heterogeneity) in the models for men and women. There are two ways of approaching this issue, both of which lead to the same conclusion. First, suppose that the observed dichotomy—promoted or not promoted—is wholly determined by whether an unobserved, continuous variable $y$ is above or below some threshold value $\mu$. Let $z = 1$ if $y > \mu$ and let $z = 0$ if $y \leq \mu$. We can think of $y$ as the latent propensity for promotion. Assume further that $y$ is generated by the linear model

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_J x_{iJ} + \sigma\varepsilon_i \tag{2}$$

for $i = 1, \ldots, n$ cases. In this equation, $\varepsilon_i$ is a random disturbance that is assumed to be independent of the $x$ variables and has a fixed variance. The parameter $\sigma$ allows the disturbance variance to be adjusted upward or downward. If we also assume that $\varepsilon_i$ has a standard logistic distribution, it follows that the observed dichotomy $z$ is governed by the logit model

$$g[\Pr(z_i = 1)] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_J x_{iJ}, \tag{3}$$

where $g(p) = \log[p/(1 - p)]$, the logit "link" function. The $\beta$ coefficients in (3) are related to the $\alpha$ coefficients in (2) by

$$\beta_0 = (\alpha_0 - \mu) / \sigma \tag{4}$$

$$\beta_j = \alpha_j / \sigma \qquad\qquad j = 1, \ldots, J. \tag{5}$$

These results are well known (e.g., Amemiya 1985:269). The same results apply if we assume that $\varepsilon$ has a standard normal distribution, except that $g$ becomes the probit link function—the inverse of the cumulative distribution function for a standard normal variable.

For drawing inferences about the slope coefficients, since $\beta_j = 0$ implies that $\alpha_j = 0$, the usual chi-square statistics provide valid tests for whether $x_j$ has an effect on $y$. On the other hand, comparisons of coefficients across groups will be problematic if $\sigma$ differs across groups.

Unless we are willing to assume that the disturbance variance is constant across groups, the standard tests for cross-group differences in the $\beta$ coefficients tell us nothing about differences in the $\alpha$ coefficients.

In most cases, I think there is insufficient justification for that assumption. In the case of assistant professors, for example, there is reason to believe that women have more heterogeneous career patterns than men (Zuckerman, Cole, and Bruer 1991; Long and Fox 1995), especially in the period covered by the data used here. Hence, unmeasured variables affecting the chances of promotion may be more important for women than for men. That difference could explain why the coefficients in Table 1 are larger for men than for women.

If the latent variable formulation seems too hypothetical, there is also a more direct approach to the problem. Suppose we take a standard logit or probit model and introduce a variable $\varepsilon$ to represent unmeasured factors that affect the probability of a promotion:

$$g[\Pr(z_i = 1|\varepsilon_i)] = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_J x_{iJ} + \sigma\varepsilon_i. \tag{6}$$

As before, $g$ can be either the logit link function or the probit link function. Note that the probability on the left-hand side is now a conditional probability. But since $\varepsilon$ is not observed, what we really need is the unconditional probability. If $g$ is the probit function and $\varepsilon$ has a standard normal distribution, it can be shown that the unconditional probability is also given by a probit model (Finney 1971:196-97):

$$g[\Pr(z_i = 1)] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_J x_{iJ}, \tag{7}$$

with

$$\beta_j = \frac{\alpha_j}{\sqrt{1+\sigma^2}}, \qquad j = 0,\ldots,J. \tag{8}$$

If $g$ is the logit function and $\varepsilon$ has a standard logistic distribution, the result is *not* a logit model but something that can be very closely approximated by a logit model (Allison 1987). Again, all the coefficients are scaled downward by the factor $1/(1 + \sigma^2)^{1/2}$. (For a more general treatment of this issue, see Gail, Wieand, and Piantadosi [1984].)

So we reach the same conclusion as with the latent variable models, that comparisons of logit or probit coefficients across groups are potentially confounded with differences in residual variation. This situation is quite similar to the well-known problem of comparing standardized coefficients for linear models across groups (Kim and Ferree 1981). Most researchers now recognize that such comparisons are potentially invalidated by differences in the standard deviations across groups. Instead, they compare unstandardized coefficients. The problem with logit and probit coefficients, however, is that they are inherently standardized because they depend on the magnitude of the disturbance variance. The coefficients representing the true causal effects cannot be directly estimated.

The issue is also related to the distinction between population-averaged and subject-specific effects for generalized linear models (Agresti and Lang 1993). My equation (6) corresponds to subject-specific effects that describe the impact that a variable would have on the outcome for a particular individual. Equation (7), on the other hand, gives population-averaged effects representing the net impact of a unit change in a variable for the entire population. 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.

## *TESTS THAT ADJUST FOR UNEQUAL RESIDUAL VARIATION*

I now propose some methods for comparing logit coefficients across groups while adjusting for possible differences in the disturbance variances. I will work with the latent variable formulation of equations (2) and (3), but the same methods apply to the direct formulation of equations (6) and (7).

The major difficulty is that a model that allows both the underlying coefficients and the disturbance variances to differ across groups is not identified. The key to a solution is that the disturbance variance affects all the coefficients in the same way. If we are willing to assume that one or more $x$ variables have the same underlying $\alpha$ coefficients across groups, the ratios of those coefficients across groups are equal to ratios of the disturbance variances. Suppose, for example, that we assume that $\alpha_{2M} = \alpha_{2W}$; that is, the underlying coefficient for $x_2$ is the same for men and women. It follows that

$$\frac{\beta_{2M}}{\beta_{2W}} = \frac{\alpha_{2M} / \sigma_M}{\alpha_{2W} / \sigma_W} = \frac{\sigma_W}{\sigma_M}. \qquad (9)$$

In principle, all the other coefficients for women could be multiplied by that ratio to make them comparable to the coefficients for men. So we have identification. In fact, if more than one coefficient is assumed to be the same across groups, we have overidentification. What we need, then, are methods that take sampling error and overidentification into account. I will consider two likelihood ratio tests:

- A test of the null hypothesis that all the underlying $\alpha$ coefficients are the same across groups versus the alternative that at least one of them differs.
- A test of the null hypothesis that all the $\alpha$ coefficients are the same across groups versus the alternative that a specified coefficient differs.

The second test has a precedent in the work of Sobel and Arminger (1992), who developed a nonlinear simultaneous probit model for household fertility decisions.

I will work with the two-group case (with the groups labeled "men" and "women"). The extension to three or more groups is straightforward and will be outlined later. I will also focus on the logit model, but the results apply equally to the probit model.

Let $G_i$ be a variable with a value of one for women and zero for men. Under the null hypothesis that all coefficients are the same across groups, we can write the underlying model as a single equation,

$$y_i = \alpha_0 + \alpha_1 G_i + \sum_{j>1} \alpha_j x_{ij} + \sigma_i \varepsilon_i, \qquad (10)$$

where $\varepsilon_i$ has a standard logistic distribution and is independent of $x$. To make the disturbance variance differ by groups, we specify

$$\sigma_i = \frac{1}{1 + \delta G_i}, \qquad (11)$$

with $\delta > -1$. Equation (11) implies that $\sigma = 1$ for men and $\sigma = 1/(1 + \delta)$ for women. Thus, if $\delta$ is positive, the disturbance variance is smaller for women than for men, whereas if $\delta$ is negative, the disturbance variance is larger for women. Moreover, $100\delta$ is the percentage by which the disturbance standard deviation for men is greater or less than the standard deviation for women. (There are other functional forms that allow the standard deviations to differ, but equation (11) is particularly convenient for the algebra below.) The choice of $\sigma = 1.0$ for men is an arbitrary normalization, reflecting the fact that we can identify only ratios of $\alpha$ coefficients, not the original coefficients themselves. Note that $G$ is also an explanatory variable in the model. That allows the intercepts to differ in the male/female equations, even when the slopes are the same.

Together with equations (2)-(5), equations (10) and (11) imply that

$$\log\left[\frac{p_i}{1-p_i}\right] = \left(\alpha_0^* + \alpha_1 G_i + \sum_{j>1} \alpha_j x_{ij}\right)(1+\delta G_i), \tag{12}$$

where $p_i = \Pr(z_i = 1)$ and $\alpha_0^* = \alpha_0 - \mu$. After some algebraic manipulation, we get

$$\log\left[\frac{p_i}{1-p_i}\right] = \alpha_0^* + (\alpha_1 + \alpha_0^*\delta + \alpha_1\delta)G_i + \sum_{j>1} \alpha_j x_{ij} + \sum_{j>i} \alpha_j \delta G_i x_{ij}. \tag{13}$$

Thus, we have a standard logit equation with all the “main effects” of the $x$ variables and all their interactions with the group dummy $G$. But the interactions have a special form: The ratio of each interaction coefficient to its corresponding main effect is $\delta$. If we had a logit regression program that could impose nonlinear constraints on the coefficients, we could readily estimate (13). I know of no commercial program that will do this, however.

Equation (13) does assure us that the model is identified because the unconstrained model (with all group-by-variable interactions) is known to be identified. (Imposition of constraints on an identified model does not ordinarily destroy identification.) For estimation, it is easier to work with equation (12), although, again, no commercial software will do this automatically. However, many commercial packages now have general optimization procedures that can be easily adapted to equation (12).

The log likelihood for a single individual can be written as

$$z_i u_i - \log[1 + \exp(u_i)], \tag{14}$$

where

$$u_i = \left( \alpha_0 + \alpha_1 G_i + \sum_{j>1} \alpha_j x_{ij} \right)(1 + \delta G_i). \tag{15}$$

I used the NLIN procedure in SAS (SAS Institute 1990) to maximize the likelihood function, as described in the appendix. Other possible candidates (with sample code in the appendix) include the LE program in BMDP (Dixon 1992), the MINIMIZE command in LIMDEP (Greene 1992), and the ML command in Stata (Gould and Sribney 1999). The appendix also shows how to fit the model with a standard logit program by doing a simple line search on $\delta$.

The first two columns of Table 2 give the results for fitting the model under the null hypothesis that all the $\alpha$ coefficients are the same for men and women but $\sigma$ differs. The coefficients can be interpreted as the estimated effects of the variables when $\sigma$ is constrained to be 1.0. $\hat{\delta}$ is the estimate of how much the disturbance standard deviation changes by sex. The value –.26 tells us that the standard deviation of the disturbance variance for men is 26 percent lower than the standard deviation for women. This estimate is significantly different from zero by a Wald chi-square test (the squared ratio of the estimate to its standard error) and also by a likelihood ratio test. The latter is calculated by taking twice the positive difference between the log likelihood for this model and the log likelihood (–838.53) for an ordinary logit model (which asserts that $\delta = 1$ for both groups).

To test the hypothesis that at least one of the $\alpha$ coefficients differs by sex, we would ideally compare the constrained model in Table 2 with a model that allows both the $\alpha$ coefficients and the $\sigma$ parameter to differ by sex. As already noted, this unconstrained model is not identified. That is not a problem, however, because its likelihood is always

**TABLE 2: Logit Regressions Predicting Promotion to Associate Professor for Male and Female Biochemists, Disturbance Variances Unconstrained**

| | *All Coefficients Equal* | | *Articles Coefficient Unconstrained* | |
|---|---|---|---|---|
| *Variable* | *Coefficient* | SE | *Coefficient* | SE |
| Intercept | 7.4913*** | .6845 | –7.3655*** | .6818 |
| Female | –0.93918** | .3624 | –0.37819 | .4833 |
| Duration | 1.9097*** | .2147 | 1.8384*** | .2143 |
| Duration squared | –0.13970*** | .0173 | –0.13429*** | .01749 |
| Undergraduate selectivity | 0.18195** | .0615 | 0.16997*** | .04959 |
| Number of articles | 0.06354*** | .0117 | 0.07199*** | .01079 |
| Job prestige | –0.4460*** | .1098 | –0.42046*** | .09007 |
| $\hat{\delta}$ | –0.26084* | .1116 | –0.16262 | .1505 |
| Articles × Female | | | –0.03064 | .0173 |
| Log likelihood | –836.28 | | –835.13 | |

*$p$ < .05. **$p$ < .01. ***$p$ < .001.

identical to the likelihood for a model allowing the α coefficients to vary by sex and constraining the σ parameter to be the same. We can get this likelihood by estimating an ordinary logit model that contains the full set of unrestricted interactions with sex. Alternatively, we can add the two log likelihoods in Table 1 to get –832.73. Taking twice the positive differences between this number and the log likelihood in the first column of Table 2 yields 7.10. This has 4 degrees of freedom (the difference in the number of estimated parameters in the constrained model and the unconstrained model), implying a *p* value of .13. Because this is greater than the conventional .05 level, we cannot reject the null hypothesis that the αs are the same, suggesting that the apparent differences between male and female coefficients are the result of a difference in disturbance variances.

If there were no a priori reasons for expecting a particular coefficient to differ by sex, a strong case could be made for stopping at this point. Further tests of individual coefficients would only capitalize on chance. If we *had* found a *p* value less than .05, we could then proceed to test individual coefficients, ideally with some sort of correction for multiple comparisons. In our promotion example, let us see how to test for a sex difference in the effect of article counts, the one variable that showed a significant difference under standard methods. Here, the strategy is to compare the constrained model with a model that allows an unconstrained interaction between sex and article counts. If article counts is denoted by $x_2$, we can modify equation (15) to read

$$u_i = \left( \alpha_0 + \alpha_1 G_i + \sum_{j>1} \alpha_j x_{ij} + \lambda x_{i2} G_i \right)(1 + \delta G_i). \tag{16}$$

The last two columns of Table 2 show the results of fitting this model. Testing whether $\lambda = 0$ is equivalent to testing whether the effect of $x_2$ differs by sex while allowing for a difference in the disturbance variance. The Wald chi-square of $3.14 = (–.03064/.0173)^2$ is not significant at the .05 level. The same conclusion is reached from the likelihood ratio chi-square, found by taking twice the positive difference between the log likelihoods for the two models in Table 2, which yields a value of 2.30. (Of course, a Bonferroni correction for multiple comparisons would make it even harder to reject the null hypothesis.) The apparent difference in the coefficients for article counts in Table 1 does not necessarily reflect a real difference in causal effects. It can be readily explained by differences in the degree of residual variation between men and women.

If we had found a significant effect for $\lambda$, the interpretation of the coefficients would be just like a standard analysis with interactions: $\hat{\alpha}_2 = .072$ is the estimated effect of article counts for males, and $\hat{\alpha}_2 + \hat{\lambda} = .072 – .031 = .041$ is the estimated effect of article counts for females. While these estimates are uncontaminated by differences in the disturbance variance, they do assume that $\sigma$ is fixed at 1 for males.

## *ADDITIONAL ISSUES*

Consider some possibilities that are not covered by this example. Suppose that $\delta$ is not significantly different from zero, by either the Wald or the likelihood ratio tests discussed above. That tells us that there is insufficient evidence for concluding that the disturbance variances differ across groups. In that event, it is probably safe to go ahead with standard methods for testing for differences in the coefficients. (While there are subtle and controversial issues with regard to the use of one statistical test to determine the form of another statistical test, this practice is widely adopted for many different applications.)

What if the estimate of δ is less than –1? While this makes no sense for the model under consideration, it can occasionally happen in practice. Specifically, it may happen if most of the coefficients for one group are positive and most of the coefficients for the other group are negative. Changes in sign cannot be explained by a difference in the disturbance variances. Accordingly, an estimate for δ that is significantly less than –1 is itself an indication that there are at least some real differences in coefficients across groups. This can be tested with the Wald chi-square,

$$\left(\frac{\hat{\delta}+1}{s.e.(\hat{\delta})}\right)^2, \tag{17}$$

which has 1 degree of freedom.

A more common occurrence is that the estimate for δ is positive but that two specific coefficients to be compared are opposite in sign. Because a difference in disturbance variances cannot account for a difference in sign, it is natural to ask whether the specialized tests proposed here are necessary. The answer is yes. Although a difference in disturbance variances cannot fully account for a change in sign, it can certainly affect the magnitude of the observed difference. And because the sign of an estimated coefficient can easily change due to sampling error, it is important to remove all confounding effects before making the comparison.

If the new test shows that a particular predictor variable has coefficients that differ across groups, it is not clear how to proceed in testing for differences for other variables. The test described in the previous section for a given predictor assumed that the coefficients for all the other variables were equal across groups. However, if there is clear evidence that one set of coefficients differs across groups, then it makes sense to relax the constraint for that variable when testing equality of coefficients for other variables. As noted earlier, however, there is a limit to how far one can go with this strategy. At least one set of coefficients must be constrained to be equal across groups.

## *THREE OR MORE GROUPS*

The extension to three or more groups is straightforward. Suppose there are $K$ groups with a running index $k = 1, \ldots, K$. Let $G_{ik}$ have a value of 1 if individual $i$ is in group $k$, and 0 otherwise except that $G_{iK}$ is always 0. The extended version of equation (12) is

$$\log\left(\frac{p_i}{1-p_i}\right) = \left(\alpha_0^* + \sum_j \alpha_j x_{ij} + \sum_k \gamma_k G_{ik}\right)\left(1 + \sum_k \delta_k G_{ik}\right), \tag{18}$$

where $\delta_k > -1$ for all $k$, and $G_{ik} = 0$ for $k = K$. The log likelihood for a single individual can be written as

$$z_i u_i - \log[1 + \exp(u_{i)}],$$

where

$$u_i = \left(\alpha_0^* + \sum_j \alpha_j x_{ij} + \sum_k \gamma_k G_{ik}\right)\left(1 + \sum_k \delta_k G_{ik}\right). \tag{19}$$

When the effect of a single variable, say $x_J$, is allowed to vary across groups, the last equation is modified to read

$$u_i = \left(\alpha_0^* + \sum_j \alpha_j x_{ij} + \sum_k \gamma_k G_{ik} + \sum_k \lambda_k x_{iJ} G_{ik}\right)\left(1 + \sum_k \gamma_k G_{ik}\right). \tag{20}$$

## *DISCUSSION*

This article has (1) shown that there is a potential source of invalidity in comparisons of logit and probit coefficients across groups, (2) proposed a method for removing that invalidity, and (3) presented an example in which the new method yields results that are qualitatively different from those of the standard method. Several questions remain to be answered.

### *HOW COMMON IS THE PROBLEM?*

Heterogeneity of disturbance variances is often present in applications of linear models, but the problem is less serious in that setting because there is no bias in the coefficient estimates. Because the logit model can be derived from a dichotomized linear model, it is natural to expect that heterogeneity in disturbance variances will be just as common in the logit case.

Keep in mind, however, that the severity of the problem may depend on the set of variables included in the model. Models with additional covariates may be less problematic because of reductions in residual heterogeneity. For the example of promotions to associate professor, the much larger models reported in Long et al. (1993) may be less prone to unequal residual heterogeneity than the models examined here. In fact, the models with more covariates show no systematic tendency for male coefficients to be larger than female coefficients.

Binary regression models differ in this respect from ordinary linear models. With linear models, it is essential to include variables that affect the dependent variable and are correlated with those variables already in the model. With binary regression models, it is important to include *any* variables that affect the dependent variable, regardless of whether they are correlated with the current set of variables.

### *IS THERE ANY WAY TO GAUGE THE SERIOUSNESS OF THE PROBLEM WITHOUT IMPLEMENTING THE NEW METHOD?*

I suggest examining the ratios of the coefficients in the different groups as I did in Table 1. If one group has coefficients that are consistently higher or lower than those in another group, it is a good indication of a potential problem that is amenable to solution.

### *HOW GOOD IS THE PROPOSED METHOD?*

While the method has some weaknesses, it is hard to see how it could be improved. To evaluate conventional procedures, I elaborated a model that is widely used to justify logit or probit analysis. Under that model, the proposed tests have the usual optimality properties of likelihood ratio statistics. One possible flaw is that the tests cannot detect departures from the null hypothesis (of no difference between groups) if *all* the true coefficients differ by a constant multiple across groups. That is because such uniform differences are attributed to unequal disturbance variances rather than to real differences in the coefficients. For example, if the true coefficients for women are all 40 percent lower than the corresponding coefficients for men, the tests are unlikely to detect any differences, regardless of sample size. Unfortunately, there is no way to circumvent this problem because it stems from underidentification.

### *DOES THE METHOD REST ON ANY PROBLEMATIC ASSUMPTIONS?*

Except for the caveat just mentioned, the test of the hypothesis that *at least* one set of coefficients differs across groups is relatively unproblematic. The same cannot be said for the test of the hypothesis that the coefficients for a *specific* variable differ across groups. That test requires that the true coefficients for other variables be constrained equal across groups, and those constraints may not be appropriate. Under conventional procedures, there is a choice between introducing variable-by-group interactions one at a time into a model or estimating them all at once. The advantage of the latter (which is equivalent to estimating separate models for each group) is that fewer constraints are imposed and the estimate of each interaction "controls" for the others. Under the method proposed here, however, at least one set of true coefficients must be constrained across groups in order to identify the δ parameter. And to get good estimates of δ, it is desirable to constrain as many coefficients as possible. Unfortunately, the observed differences between the coefficients provide no foolproof way of determining which ones should be constrained equal.

### *WILL THE NEW METHOD MAKE IT MORE DIFFICULT TO DETECT CROSS-GROUP DIFFERENCES?*

In most cases, yes. For the null hypothesis that at least one coefficient differs across groups, the test proposed here will always yield a chi-square that is smaller than a conventional test. That is because the null model in the new test is less restrictive than the null model in the conventional test, but the alternative model is the same. Hence, differences in log likelihoods between the null and alternative hypotheses must be smaller in the new test. When testing for differences for a specific covariate, the new test can be either smaller or larger than the conventional test. If the difference between a pair of estimated coefficients is in the same direction as the prevailing differences between the rest of the coefficients, the new test will have a smaller chi-square than the conventional test. On the other hand, if the difference between a pair of coefficients is in the opposite direction to the prevailing differences, the chi-square could be larger. It could also be smaller, however, because the estimation of the $\delta$ parameter introduces additional sampling error.

### *HOW DIFFICULT IS THE METHOD TO IMPLEMENT?*

The method is certainly more difficult than conventional tests and may intimidate those with modest statistical and computational skills. Once one has done it a couple of times, however, the method is relatively straightforward. When the NLIN procedure in SAS was applied to the academic promotion example, convergence occurred in 13 seconds (on a Power Macintosh 7100/80). By comparison, fitting an ordinary logit model using the SAS procedure LOGISTIC took 5 seconds.

### *SHOULD THE NEW METHOD BE ROUTINELY USED IN MAKING COMPARISONS?*

Given the strong possibility of invalid inferences with conventional methods and the relative ease of the new method, routine use seems advisable. On the other hand, if the conventional tests and the new tests yield the same conclusions, it may not be necessary to report the new tests (except, perhaps, in a footnote).

## APPENDIX
## Computer Methods for Fitting the New Models

*SAS*

I used the NLIN procedure in SAS to get the estimates reported in Table 2. NLIN is designed for weighted least squares estimation of nonlinear models. It can also produce maximum likelihood estimates of generalized linear models by the method of iteratively reweighted least squares. For the promotion data, I used the following statements to fit the constrained model:

```
data promo;
  infile ‘c:promo.dat’;
  input prom female dur undgrd arts prest;
  dur2=dur*dur;
run;
proc nlin nohalve sigsq=1;
 parms del=0 b0=0 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0;
  int=1+del*female;
  pred=b0+b1*dur+b2*dur2+b3*female+b4*undgrd+b5*arts
    +b6*prest;
    u=pred*int;
    w=exp(pred*int);
    _weight_=(1+w)**2/w;
    _loss_=(-prom*u+log(1+exp(u)))/_weight_;
  model prom=w/(1+w);
run;
```

The `PARMS` statement assigns names and starting values to the eight parameters. It is followed by a series of programming statements that define the linear function of the explanatory variables and the logistic transform. The `_WEIGHT_` variable weights each observation by the inverse of the variance. `_LOSS_` defines the negative log likelihood as the criterion function (divided by `_WEIGHT_` to cancel the weight given to each observation). In the output file, the negative log likelihood is labeled “sum of loss.”

To fit the model that allows the effect of `ARTS` to vary across groups, I simply added `B7=0` to the `PARMS` statement and `+ B7*ARTS*FEMALE` to the statement defining the `PRED` variable.

I have also written an SAS macro called `GLOGIT` that automates this process and generalizes it to three or more groups. For the promotion data, the macro is called by the statement

```
%glogit(data=promo,response=prom,vars=dur dur2 female undgrd
   arts prest,groupvar=female)
```

A copy of the macro is available at http://www.ssc.upenn.edu/~allison.

### *LIMDEP*

The LIMDEP command MINIMIZE may be used to maximize the likelihood functions given in equations (14), (15), and (16). The following commands were used to read in the data and estimate the constrained model:

```
read; file=‘c:promo.dat’; nvar=6;
names=prom,female,dur,undgrd,arts,prest $
create; dur2=dur*dur $
namelist; z=one,female,dur,dur2,undgrd,arts,prest $
minimize; labels=b0,b1,b2,b3,b4,b5,b6,del;
start=0,0,0,0,0,0,0,0;
fcn=-prom*dot[z]*(1+del*female)+
   log(1+exp(dot[z]*1+del*female)) $
```

The NAMELIST command assigns all the explanatory variables to a single label Z. The ONE in the name list refers to the intercept. The LABELS statement assigns names to all the parameters to be estimated. START assigns starting values to those parameters. FCN defines the negative log likelihood function for a single individual. In that statement, DOT[Z] denotes the weighted sum of the variables in the list Z, the weights being the initial eight parameters.

To estimate the partially constrained model (with the coefficient for articles allowed to vary freely across groups), the MINIMIZE command was modified to read

```
minimize; labels=b0,b1,b2,b3,b4,b5,b6,b7,del;
start=0,0,0,0,0,0,0,0;
fcn=-prom*(dot[z]+b8*arts*female)*(1+del*female)
   +log(1+exp((dot[z]+b8*arts*female)*(1+del*female))) $
```

While MINIMIZE does the job, my experience is that it converges very slowly.

### *STATA*

The ML command in Stata will maximize the likelihood functions given in equations (14), (15), and (16). The following commands were used to read the raw data and estimate the constrained model:

```
infile prom female dur undgrd arts prest using c:promo.dat
generate dur2=dur*dur
program define glogit
```

```
  version 6
  args lnf theta delta
  quietly replace ‘lnf’ =
   $ml_y1*‘theta’*(1+‘delta’)-
    ln(1+exp(‘theta’*(1+`delta’)))
end
ml model lf glogit (prom = undgrd arts dur dur2 prest female)
   (delta: female,nocons)
ml maximize
```

The commands bracketed by PROGRAM and END define the estimation problem in a general way. These commands can be saved in a DO file and used for any data set or particular set of variables in the model. In the ML MODEL command, LF specifies the optimization algorithm and GLOGIT refers to the name of the program defined earlier. To estimate the model with the articles coefficient unconstrained, it is only necessary to define the interaction between FEMALE and ARTS in a GENERATE statement and include the new variable as one of the independent variables in the model specification.

### *BMDP*

The LE program in the BMDP package will also estimate the new models. Here is the code for the constrained model:

```
/ input file=‘c:promo.dat’. variables=6. format=free.
/ variable names=prom,female,dur,undgrd,arts,prest.
/ transform dur2=dur*dur.
/ estimate parameters=8.
/ density
u=exp((p1+p2*female+p3*dur+p4*dur2+p5*undgrd
  +p6*arts+p7*prest)*(1+p8*female)).
   if (prom eq 1) then f=u/(1+u).
   if (prom eq 0) then f=1/(1+u).
/end
```

To estimate the partially restricted model (with the coefficient for articles allowed to vary freely across groups), the DENSITY paragraph was modified to read

```
/ density
U=exp((p1+p2*female+p3*dur+p4*dur2+p5*undgrd+
   p6*arts+p7*prest+p9*arts*female)*(1+p8*female)).
   if (prom eq 1) then f=u/(1+u).
   if (prom eq 0) then f=1/(1+u).
/ end
```

## *STANDARD LOGIT PROGRAMS*

It is relatively straightforward, although tedious, to estimate the models and calculate the chi-square tests with an ordinary logit program. The basic approach is to estimate the models in equations (13) or (16) multiple times, treating $\delta$ as a fixed parameter at each estimation. By trying out different values of $\delta$, we can find the value that maximizes the likelihood.

To estimate the constrained model, I used the following statements in SAS (SAS Institute 1990) at each iteration using the `LOGISTIC` procedure:

```
data promo;
   infile ‘c:promo.dat’;
   input prom female dur undgrd arts prest;
   delta=-.5;
   int=1+delta*female;
   undint=undgrd*int;
   prestint=prest*int;
   durint=dur*int;
   dur2int=dursq*int;
   femint=female*int;
   artsint=arts*int;
run;
proc logistic descending;
   model promo=int undint prestint dur2int dursqint
    artsint femint / noint;
run;
```

This program is run several times, with a different value of `DELTA` at each iteration. The line search algorithm (Press 1992) is essentially as follows:

1. Begin with three starting values for `DELTA` and find the log likelihood of the logit model for each value. One of these starting values can be `DELTA=0`, which is just an ordinary logit model estimated for the entire sample. Another reasonable guess can be found by averaging the ratios of the coefficients for the two groups and subtracting 1.0. For the ratios in Table 1, the geometric mean of the coefficients (excluding the intercept) was .55. Subtracting 1.0 yields –.45. I used –.50. For the third value, I used the midpoint of these two, –.25.
2. If the highest log likelihood is at one of the extreme values of `DELTA`, then choose another value that is even more extreme. Repeat until the maximum is at an interior value.
3. If the highest log likelihood is at an interior value, then choose two new values between the value with the highest log likelihood and the two adjacent values. Repeat until convergence.

The first two columns of Table A1 show the results of applying this algorithm to the promotion example. Values of `DELTA` are listed in the order they were chosen. The likelihood reaches a maximum at approximately `DELTA = –.26`. The choice of new values between the current maximum and the adjacent values was not systematic. For optimal rules for choosing values, see Press (1992). Note that the log likelihood function is rather flat between `DELTA = –.15` and `DELTA = –.30`. As a result, the chi-square test for the null hypothesis of no group differences is relatively insensitive to the precise estimate of DELTA. There would be little point in iterating to three significant digits.

To fit the model in which the coefficient for articles is allowed to vary freely between males and females, the program is modified as follows (changes in bold):

```
data rank2;
   set rank;
   delta=-.5;
   int=1+delta*female;
   undint=undgrd*int;
   prestint=prest*int;
   durint=dur*int;
   dur2int=dursq*int;
   femint=female*int;
   artsint=arts*int;
   artsfem=arts*female*int;
run;
   proc logistic descending;
   model promo=undint prestint dur2int dursqint
    artsint artsfem femint;
run;
```

Again, this code is executed multiple times using the line search algorithm to choose successive values of `DELTA`. Results are shown in the right-hand column of Table A1.

**TABLE A1: DELTA and Log Likelihoods for Line Search Algorithm Applied to Promotion Data**

| | *Fully Restricted* | | *Partially Restricted* | |
|---|---|---|---|---|
| *Iteration* | DELTA | *–2 ◈ Log Likelihood* | DELTA | *–2 ◈ Log Likelihood* |
| 1 | .00 | 1677.066 | .00 | 1671.4917 |
| 2 | –.25 | 1672.5746 | –.25 | 1670.7157 |
| 3 | –.50 | 1679.0048 | –.15 | 1670.2753 |
| 4 | –.15 | 1673.5031 | –.20 | 1670.3452 |
| 5 | –.35 | 1673.3155 | –.10 | 1670.4657 |
| 6 | –.20 | 1672.8621 | –.17 | 1670.2699 |
| 7 | –.30 | 1672.7018 | –.13 | 1670.3223 |
| 8 | –.22 | 1672.7015 | –.14 | 1670.2938 |
| 9 | –.27 | 1672.5720 | –.16 | 1670.2673 |
| 10 | –.24 | 1672.6011 | | |
| 11 | –.26 | 1672.5648 | | |

## NOTE

1. The issues examined here are confusingly similar to those discussed by Liao (1995), who proposed methods for comparing coefficients in generalized linear models with heterogeneity in dispersion parameters. Liao's methods apply when the data are naturally clustered, with multiple, nonindependent observations in each cluster. An example would be repeated, dichotomous observations for a single individual. In Liao's models, the disturbance variable is specific to the cluster, the disturbance variance is identified, and the problem is that standard errors are biased. By contrast, I deal with the more common situation in which all observations are independent and the disturbance variable is specific to each observation. In the models developed here, only ratios of the disturbance variances are identified, and the problem is that differences between coefficient estimates are biased.

## REFERENCES

Agresti, Alan, and Joseph B. Lang. 1993. "A Proportional Odds Model With Subject-Specific Effects for Repeated Ordered Categorical Responses." *Biometrika* 80:527-34.

Allison, Paul D. 1982. "Discrete-Time Methods for the Analysis of Event Histories." Pp. 61-98 in *Sociological Methodology 1982*, edited by Samuel Leinhardt. San Francisco: Jossey-Bass.

———. 1987. "Introducing a Disturbance Into Logit and Probit Regression Models." *Sociological Methods & Research* 15:355-74.

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Baxter, Janeen. 1994. "Is Husband's Class Enough? Class Location and Class Identity in the United States, Sweden, Norway, and Australia." *American Sociological Review* 59:220-35.

Dixon, W. J. 1992. *BMDP Statistical Software Manual*. Vol. 2. Berkeley: University of California Press.

Finney, D. J. 1971. *Probit Analysis*. 3d ed. Cambridge, UK: Cambridge University Press.

Gail, M. H., S. Wieand, and S. Piantadosi. 1984. "Biased Estimates of Treatment Effect in Randomized Experiments With Nonlinear Regression and Omitted Covariates." *Biometrika* 71:431-44.

Gould, William, and William Sribney. 1999. *Maximum Likelihood Estimation With Stata*. College Station, TX: Stata Press.

Greene, William H. 1992. *LIMDEP: User's Manual and Reference Guide, Version 6.0*. Bellport, NY: Econometric Software Inc.

Kalmijn, Matthijs. 1994. "Mother's Occupational Status and Children's Schooling." *American Sociological Review* 59:257-75.

Kim, Jae-On, and G. Donald Ferree. 1981. "Standardization in Causal Analysis." *Sociological Methods & Research* 10:187-210.

Liao, Tim Futing. 1995. "Testing Coefficient Equality and Adjusting for Dispersion Heterogeneity in Generalized Linear Models Between Two or More Groups." Paper prepared for presentation at the annual meeting of the American Sociological Association, Washington, DC, August.

Long, J. Scott, Paul D. Allison, and Robert McGinnis. 1993. "Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity." *American Sociological Review* 58:703-22.

Long, J. Scott, and Mary Frank Fox. 1995. "Scientific Careers—Universalism and Particularism." *Annual Review of Sociology* 21:45-71.

Press, William H. 1992. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge, UK: Cambridge University Press.

SAS Institute. 1990. *SAS/STAT User's Guide, Version 6*. 4th ed. Cary, NC: SAS Institute Inc.

Sekulic, Dusko, Garth Massey, and Randy Hodson. 1994. "Who Were the Yugoslavs—Failed Sources of a Common Identity in the Former Yugoslavia." *American Sociological Review* 59:83-97.

Sobel, Michael, and Gerhard Arminger. 1992. "Modeling Household Fertility Decisions: A Nonlinear Simultaneous Probit Model." *Journal of the American Statistical Association* 87:38-47.

Wright, Rosemary, and Jerry Jacobs. 1994. "Male Flight From Computer-Work—A New Look at Occupational Resegregation and Ghettoization." *American Sociological Review* 59:511-36.

Zuckerman, Harriet, Jonathan R. Cole, and John T. Bruer, eds. 1991. *The Outer Circle: Women in the Scientific Community.* New York: Norton.

*Paul D. Allison is a professor of sociology at the University of Pennsylvania. His recently published books include* Multiple Regression: A Primer *and* Logistic Regression Using the SAS System: Theory and Applications*. He is currently writing a book on missing data. Each summer, he teaches 5-day workshops on event history analysis and categorical data analysis.*