Logit and probit regression models for dichotomous data make no explicit allowance for heterogeneity induced by omitted explanatory variables or by random fluctuations. This article considers several alternative models for incorporating heterogeneity by the inclusion of a disturbance term. When all the observations are independent, the presence of the disturbance has few empirical consequences. In particular, the variance of the observed counts does not increase and conventional estimators are still appropriate. For some models, however, the disturbance variance may invalidate cross-population comparisons. Quite different implications arise with grouped data when there is a single realization of the disturbance for each group. The observed variance is *increased and conventional estimators are inefficient. Several alternative estimators are considered.*

Introducing a Disturbance into Logit and Probit Regression Models

PAUL D. ALLISON
University of Pennsylvania

Regression models for dichotomous dependent variables usually take the following form. Let Y be a random variable with possible values of 0 and 1, and let $\mathbf{x}$ be a $K \times 1$ vector of explanatory variables.[1] Then

$$\Pr(Y = 1) = F(\boldsymbol{\beta}\mathbf{x}) \qquad [1]$$

where $\boldsymbol{\beta}$ is a $1 \times K$ vector of unknown coefficients and F is some function bounded by 0 and 1. Since the cumulative distribution function (c.d.f.) for any random variable must satisfy these bounds, it is common to choose F as some c.d.f. For example, the logit (logistic) regression model (Cox, 1970) specifies F as the c.d.f. for the standard logistic distribution,

$$F(z) = 1/(1 + e^{-z}), \qquad [2]$$

AUTHOR'S NOTE: *For helpful comments on earlier versions of this article, I am indebted to Thomas Santner, Anne Shoemaker, Clifford Clogg, Ronald Schoeberg, J. Scott Long, Ralph Hakkert, Roger Avery, and several anonymous referees.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 15 No. 4, May 1987 355-374

while the probit regression model (Finney, 1971) specifies F as the c.d.f. for a standard normal variable.

While the model of equation 1 has proved useful in a wide variety of settings, it has been criticized for asserting that Pr(Y = 1) is completely determined by **x**, with no explicit allowance for heterogeneity induced by omitted explanatory variables or for random fluctuations of the probability about some expected value (Amemiya and Nold, 1975; Hanushek and Jackson, 1977). These criticisms have been motivated both by the inherent implausibility of this specification, and by the observation that, for grouped data, the variation in the observed frequencies often greatly exceeds that expected under the postulated model (Baker and Nelder, 1978: §6.4).

To remedy this shortcoming, Amemiya and Nold (1975) introduced an unobserved random disturbance into equation 1. They argued that the presence of such a disturbance term implies that the usual maximum likelihood (ML) or weighted least squares (WLS) estimators for β are inefficient, and that estimated standard errors are biased downward. To avoid these difficulties, they proposed a modification of the usual WLS estimator.

In this article I consider some alternative ways to introduce a disturbance into equation 1, and I conclude that the consequences differ markedly among the various specifications. I first demonstrate that, for individual-level data with independent observations, the disturbance term has few if any consequences for estimation but may invalidate cross-population comparisons. For grouped data, I show that the results of Amemiya and Nold are misleading because they failed to recognize that their model implies dependence among the dichotomous observations in each group. Finally, I survey alternative methods for models in which the disturbance term induces dependence among the observations.

ALTERNATIVE MODELS FOR INDIVIDUAL-LEVEL DATA

Amemiya and Nold (1975) proposed a model of the general form

$$\Pr(Y_i = 1 \mid \epsilon_i) = F(\boldsymbol{\beta}\mathbf{x}_i + \epsilon_i), \tag{3}$$

where the ϵ_i are unobserved random variables, independently distributed with $E(\epsilon_i) = 0$ for $i = 1, \ldots, n$. This is equivalent to the binomial trials model of Winship and Mare (1983). Amemiya and Nold further assumed that F is the logistic function (equation 2) and that $\text{var}(\epsilon_i) = \sigma^2$, a constant over i.

While equation 3 is certainly a plausible specification, another approach is to position ϵ_i outside the function F so that

$$\Pr(Y_i = 1 \mid \epsilon_i) = F(\boldsymbol{\beta}\mathbf{x}_i) + \epsilon_i, \tag{4}$$

again with $E(\epsilon_i) = 0$ for all i. To distinguish these alternatives, I shall refer to equation 3 as an *internal* model and to equation 4 as an *external* model. Note that the disturbance term in the external model must satisfy

$$-F(\boldsymbol{\beta}\mathbf{x}_i) \leq \epsilon_i \leq 1 - F(\boldsymbol{\beta}\mathbf{x}_i). \tag{5}$$

While this constraint may seem awkward, it is of course equivalent to assuming that $\Pr(Y_i = 1)$ is a random variable on the interval [0, 1] with a mean of $F(\boldsymbol{\beta}\mathbf{x}_i)$. Such an approach has a long tradition in Bayesian analysis (Wonnacott and Wonnacott, 1984: 524), and is also well known for a special case of equation 4, the beta-binomial model discussed below.

For ML estimation of either the internal or external models, it is desirable to obtain the marginal distribution of Y (not conditioned on ϵ_i). This may be difficult for the internal model but is simple for the external model. From equation 4 we have

$$\begin{aligned} \Pr(Y = 1) &= E[\Pr(Y = 1 \mid \epsilon)] \\ &= E[F(\boldsymbol{\beta}\mathbf{x}) + \epsilon] \\ &= F(\boldsymbol{\beta}\mathbf{x}) + E(\epsilon) = F(\boldsymbol{\beta}\mathbf{x}). \end{aligned} \tag{6}$$

Thus the marginal distribution of Y does not depend at all on the distribution of ϵ and, in fact, is the same as the original model of equation 1 that did not have a disturbance term. This implies that the likelihood function for observations on Y and **x** is identical for equations 1 and 4 and, hence, no special estimation procedures are needed. Note that this conclusion holds for any distribution function F. It is also unnecessary to assume that var(ϵ_i) is a constant; in fact, the distribution of ϵ_i could have an entirely different shape for each observation. Although this result simplifies the problem considerably, it does mean that the data contain no information about ϵ_i even though, in some situations, the distribution of ϵ_i might be of substantive interest.

If one begins with the internal model of equation 3, on the other hand, the situation is more difficult. The problem comes in evaluating the integral

$$\Pr(Y = 1) = E[\Pr(Y = 1 \mid \epsilon)] = \int_{-\infty}^{+\infty} F(\beta x + \epsilon)g(\epsilon)d\epsilon \qquad [7]$$

where g(.) is the density function for ϵ. To proceed further, it is necessary to specify the functions F and g.

Consider first the probit model in which $F = \Phi$ where Φ is the standard normal c.d.f. Assume further that ϵ has a normal distribution with a mean of zero and a constant variance σ^2. Under these conditions, it has been shown (Finney, 1971: 196-197; Muthén, 1979) that

$$\Pr(Y = 1) = \Phi(\boldsymbol{\beta}^* \mathbf{x}) \qquad [8]$$

where

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}/(\sigma^2 + 1)^{1/2} \qquad [9]$$

Thus introducing an internal, normal random disturbance into the probit model yields another probit model, but with the coefficient vector scaled downward by a factor that varies inversely with the variance of the disturbance term.

Although the presence of the disturbance term is consequential in this case, the fact that equation 8 has the same form as the standard probit model means that the two models cannot be empirically distinguished. More specifically, $\boldsymbol{\beta}$ and σ^2 are not separately identifiable since for any change in σ^2 there is a compensating change in $\boldsymbol{\beta}$ that leaves the likelihood unchanged. On the other hand, $\boldsymbol{\beta}^*$ can be estimated using standard maximum likelihood procedures with the usual asymptotic properties of efficiency and consistent standard error estimates.

Although we cannot directly estimate $\boldsymbol{\beta}$ under this model, some inferences about $\boldsymbol{\beta}$ for a single population are quite feasible. Since $\beta_k^* = 0$ implies that $\beta_k = 0$ (and vice versa) for any k, a test of the former hypothesis is also a test of the latter. Inferences about relative magnitudes of coefficients of different explanatory variables are also appropriate because $\beta_k^* / \beta_j^* = \beta_k / \beta_j$ for all k and j. On the other hand, comparisons of parameter estimates across populations or subpopulations will be meaningless unless the variance of ϵ is constant across the groups being compared.

For an internal logit model, evaluation of the integral in equation 7 is considerably more difficult. To be more specific, let us suppose that

$$\Pr(Y = 1 | \epsilon) = F(\boldsymbol{\beta}\mathbf{x} + \sigma\epsilon) \qquad [10]$$

where F is the logistic function (equation 2) and ϵ has a standard logistic distribution, that is, its c.d.f. is also given by equation 2. Although Amemiya and Nold (1975) correctly observed that there is no general closed-form solution to equation 7 in the logistic case, Hakkert (1978) has obtained results for certain values of σ. For $\sigma = 1$,

$$\Pr(Y = 1) = [1 - (\boldsymbol{\beta}\mathbf{x} + 1)e^{-\boldsymbol{\beta}\mathbf{x}}]/[1 - e^{-\boldsymbol{\beta}\mathbf{x}}]^2. \qquad [11]$$

For $\sigma = 2$,

$$\Pr(Y = 1) = \frac{e^{2\boldsymbol{\beta}\mathbf{x}} + e^{\boldsymbol{\beta}\mathbf{x}}(\boldsymbol{\beta}\mathbf{x} + 1) - \frac{1}{2}\pi e^{\frac{1}{2}\boldsymbol{\beta}\mathbf{x}}(e^{\boldsymbol{\beta}\mathbf{x}} - 1)}{(e^{\boldsymbol{\beta}\mathbf{x}} + 1)^2} \qquad [12]$$

Thus the disturbance term clearly changes the form of the dependence of $\Pr(Y = 1)$ on $\mathbf{x}$. The fact that the likelihood function, in this case, *does* depend on the magnitude of the disturbance variance suggests that it may be possible to draw inferences about σ. Nevertheless, for the general model of equation 3, Schoenberg

(1985) has shown that the expected information matrix for $\boldsymbol{\beta}$ and σ is singular, implying that $\boldsymbol{\beta}$ and σ are not separately identifiable.

Although equations 11 and 12 appear to be quite different from the simple logistic function in equation 2, both equations can actually be closely approximated by equation 2. Since the normal and logistic distributions are so similar in shape (Cox, 1970), it seems plausible that a result similar to equations 8 and 9 should also hold approximately for the internal logit model. Specifically, one might expect that under equation 10,

$$\Pr(Y = 1) \simeq F(\boldsymbol{\beta}^*\mathbf{x}) \qquad [13]$$

where F is the logistic function and $\boldsymbol{\beta}^*$ is given by equation 9. I have carried out two studies to investigate the adequacy of this approximation, and both support it strongly.[2] Moreover, the closeness of the approximation does not seem to deteriorate with increasing values of σ. It thus appears that when an internal logistic disturbance is introduced into a logit regression model, the practical consequence is that the coefficient estimates are scaled downward by the factor given in equation 9.

In sum, the presence of a random disturbance in logit and probit regression models does not vitiate the use of standard estimation procedures. Nevertheless, the ability to compare coefficient estimates across different populations (or the same population at different points in time) depends critically on whether an internal or external model applies. Under the external model, such comparisons are perfectly legitimate. Under the internal model, however, cross-population comparisons are valid only if the populations have the same disturbance variance.

Unfortunately, there seems to be little basis for choosing between these alternative approaches. Empirically, the internal and external probit models are indistinguishable. While the external logit model is, in principle, distinguishable from the internal logit model, the functional forms are so similar that it would take extremely large samples to discriminate between them; even if such samples were available, minor specification errors would surely confound the discrimination.

With regard to theoretical plausibility, there are some reasons to prefer the internal model. For one thing, it avoids the constraints of expression 5 that some may find artificial. For another, the internal model seems to capture better the notion of omitted explanatory variables. These are hardly compelling arguments, however.

ALTERNATIVE MODELS FOR GROUPED DATA

We turn now to the case in which there are multiple observations on Y for each value of **x**. This can happen in one of two ways: First, for a single unit of analysis with unchanging characteristics represented by **x**, we may observe Y on multiple occasions or for multiple components.[3] Second, for a sample of cases, we may group together those that have the same values on all the explanatory variables. Although both of these designs are commonly referred to as grouped data, we shall see that different models are needed to describe them. In fact, the main problem with the work of Amemiya and Nold (1975) was their failure to distinguish these two cases.

Suppose we have J groups ($j = 1, \ldots, J$), and for each group we have a $K \times 1$ vector of explanatory variables $\mathbf{x}_j$. For each j we also have n_j observations on Y_{ij} that may take on values of 0 or 1. Let $R_j = \Sigma_j Y_{ij}$, that is, the number of times that $Y_{ij} = 1$ within unit j. Now suppose the aim is to formulate an external model for such data. Assuming that each observation on Y has its own disturbance term, we can write

$$\Pr(Y_{ij} = 1 \mid \epsilon_{ij}) = F(\boldsymbol{\beta}\mathbf{x}_j) + \epsilon_{ij}, \qquad [14]$$

where the ϵ_{ij} are mutually independent with $E(\epsilon_{ij}) = 0$ for all i and j. This model is fully equivalent to equation 4 for individual-level data and, hence, the same conclusions apply. Specifically, the marginal distribution of Y_{ij} does not depend on the distribution of ϵ_{ij}, and reduces to equation 1. This implies that R_j is binomially distributed with parameter $\pi_j = F(\boldsymbol{\beta}\mathbf{x}_j)$. It follows that standard WLS or ML estimators for $\boldsymbol{\beta}$ are entirely appropriate.

The same reasoning applies when ϵ_{ij} is placed internal to the function F(.). In the case of the internal probit model, R_j is binomially distributed with $\pi_j = \Phi(\boldsymbol{\beta}^*\mathbf{x}_j)$ where $\boldsymbol{\beta}^*$ is given by equation 9. For the internal logit model, R_j is again binomially distributed with π_j given by the integral in equation 7 where F and g are the c.d.f. and density, respectively, for the standard logistic distribution. This can be approximated by $F(\boldsymbol{\beta}^*\mathbf{x}_j)$ where F is the logistic c.d.f. and $\boldsymbol{\beta}^*$ is given by equation 9. These are essentially the conclusions reached in the previous section for individual-level data.

Amemiya and Nold (1975), on the other hand, took an approach that is not equivalent to the models for individual-level data. They assumed that there is a single realization of the random disturbance ϵ_j for all the observations in each group. For an internal model, this can be expressed as

$$\Pr(Y_{ij} = 1 \mid \epsilon_j) = F(\boldsymbol{\beta}\mathbf{x}_j + \epsilon_j), \qquad [15]$$

where the ϵ_j's are mutually independent with means of 0. Under this model, R_j still has a binomial distribution *conditional* on ϵ_j, but unconditionally it is a mixture of binomials with a variance exceeding that of the binomial variance (Kleinman, 1973). As a result, conventional ML or WLS estimators for both logit and probit models will, in general, be inefficient. Amemiya and Nold also claimed that the estimated covariance matrix for the coefficients will be biased toward**0**.

The situation is virtually unchanged if we shift to an external positioning of the disturbance term:

$$\Pr(Y_{ij} = 1 \mid \epsilon_j) = F(\boldsymbol{\beta}\mathbf{x}_j) + \epsilon_j. \qquad [16]$$

Under this alternative specification, the distribution of R_j is still a mixture of binomials and the same conclusions apply: The variance of R_j will exceed that of a binomial variate, and conventional estimation procedures will be inefficient.

What Amemiya and Nold failed to note is that the models in equations 15 and 16 imply mutual dependence among the observations in each group. Conditional on ϵ_j, the Y_{ij} in each group are independent, but unconditionally they are not. In the case of the external model (equation 16), for example, it is easily shown that cov $(Y_{ij}, Y_{i'j}) = \text{var}(\epsilon_j)$ for $i \neq i'$. As discussed below, such dependence may be quite appropriate for data in which multiple observations on Y_{ij} are collected for a single physical or social unit. It is not appropriate, however, for most applications of logit or probit to grouped data. Typically, one starts with a sample of apparently independent observations, then groups them by equality or proximity of the values of the explanatory variables.

For example, Amemiya and Nold (1975) applied their model to a sample of 1,523 households for which the dependent variable was whether or not consumer durables were purchased in the previous year. The households were aggregated into 32 groups formed by the cross-classification of income (with 16 intervals) and whether or not a residence change occurred in the preceding two years. Now if the households were independent to begin with (and there was no reason to suspect otherwise), grouping them by the values of the explanatory variables should not change that fact. Parks (1977) also applied the model of Amemiya and Nold to a set of data in which there was no apparent reason why observations within each group should be dependent. For these and most other applications, a more appropriate model would be equation 14 that specifies a distinct disturbance variable for each observation, with independence among the disturbances. For this and similar models, conventional estimators should be quite satisfactory.[4]

There will, of course, be situations in which it is desirable to allow for dependence among the observations in each group. In much toxicological work, for example, the groups are animal litters and there has been considerable interest in modeling dependence due to litter effects (Haseman and Kupper, 1979). For psychological tests consisting of a set of dichotomous items, it is implausible that an individual's responses to these items would be independent (Lord and Novick, 1968). Heckman and Willis (1977) analyzed data in which the "groups" were individual women and the response variables were whether or not a woman was in the labor force in each year of a five-year interval. Clearly such observations should not be treated as independent.

The data in Table 1 provide another example. The groups are universities and the observations consist of all persons who received a Ph.D. in biochemistry at each university in 1957-1958 or 1962-1963. The outcome variable (Y_{ij}) was whether or not the person subsequently received postdoctoral training (at any university). The single explanatory variable is the total amount of NIH obligations to each university in 1964; the question to be answered is whether universities with greater federal funding produce Ph.Ds who are more likely to get postdoctoral training. For a more detailed description of the data, see McGinnis, Allison, and Long (1982).

For these data, it is reasonable to believe that there are unobserved university effects, which might be represented by a single random disturbance for each university. This, in turn, would produce dependence among the Y_{ij}'s. Evidence for this possibility is the fact that an ordinary logit model (fitted by maximum likelihood) gives a relatively poor fit. This chi-square goodness-of-fit statistic was 64.16 with only 38 degrees of freedom.

In the next two sections, I survey alternative methods for estimating the coefficients in equations 15 and 16 with grouped data. One notable feature of all these methods is that $\boldsymbol{\beta}$ and the disturbance variance are separately identified. With the exception of the WLS estimator for equation 16, most of the results are described elsewhere. Nevertheless, the range of alternatives and the connections among them have not been generally appreciated. Those methods that are computationally feasible will be applied to the data in Table 1.

INTERNAL MODELS FOR GROUPED, DEPENDENT DATA

In this section I examine estimation methods for two special cases of equation 15, which postulates a single, internally positioned disturbance variable for each group.

TABLE 1
NIH Funding, Doctorates in Biochemistry, and Proportions with Postdoctoral Training for 40 Universities

NIH Obligations[a] (x)	No. of Biochemistry Doctorates (n)	No. (out of n) with Postdoc. Training (R)	Proportion (R/n)
0.500	8	1	.125
0.500	9	3	.333
0.835	16	1	.063
0.998	13	6	.462
1.027	8	2	.250
2.036	9	2	.222
2.106	29	10	.345
2.329	5	2	.400
2.523	7	5	.714
2.524	8	4	.500
2.874	7	4	.571
3.898	7	5	.714
4.118	10	4	.400
4.130	5	1	.200
4.145	6	3	.500
4.242	7	2	.286
4.280	9	4	.444
4.524	6	1	.167
4.858	5	2	.400
4.893	7	2	.286
4.944	5	4	.800
5.279	5	1	.200
5.548	6	3	.500
5.974	5	4	.800
6.733	6	5	.833
7.000	12	5	.417
9.115	6	2	.333
9.684	5	3	.600
12.154	8	5	.625
13.059	5	3	.600
13.111	10	8	.800
13.197	7	4	.571
13.433	86	33	.384
13.749	12	7	.583
14.367	29	21	.724
14.698	19	5	.263
15.440	10	6	.600
17.417	10	8	.800
18.635	14	9	.642
21.524	18	16	.889

a. In millions of dollars, 1964.

A LOGIT MODEL

Suppose F(.) in equation 15 is the logistic function and assume that $\mathrm{var}(\epsilon_j) = \sigma^2$, a constant over j. For this model, Amemiya and Nold proposed a modified WLS estimator. Their working dependent variable is the empirical logit,[5] $U_j = \ln[R_j/(n_j - R_j)]$. The approximate variance of U_j is $\sigma^2 + 1/[n_jP_j(1 - P_j)]$ where P_j is the conditional probability defined by equation 15. To correct for heteroscedasticity, an estimate of this variance is needed. Amemiya and Nold suggested that P_j be estimated by $\hat{P}_j = R_j / n_j$. As a consistent estimator for σ^2, they used

$$\hat{\sigma}^2 = (1/J)\{\Sigma_j(U_j - \mathbf{b}\mathbf{x}_j)^2 - \Sigma_j[n_j\hat{P}_j(1 - \hat{P}_j)]^{-1}\} \qquad [17]$$

where $\mathbf{b}$ is the ordinary least squares estimator of $\boldsymbol{\beta}$ obtained by regressing U_j on $\mathbf{x}_j$. Substituting $\hat{P}_j$ and $\hat{\sigma}^2$ into the formula for the variance of U_j and taking its reciprocal, we obtain the weights $w_j = 1/\{\hat{\sigma}^2 + [n_j\hat{P}_j(1 - \hat{P}_j)]^{-1}\}$. The final step is to do a weighted regression of U_j on $\mathbf{x}_j$. An estimate of the covariance matrix[6] of the coefficients is given by $[\Sigma_j w_j \mathbf{x}_j \mathbf{x}_j']^{-1}$. For alternative estimation methods for the same model, based on quasi-likelihood theory, see Williams (1982: 5).

I applied the Amemiya-Nold method to the data in Table 1, with results shown in the third row of Table 2. These results may be compared with those obtained with the conventional WLS estimator (shown in the second row), that assumes that $\sigma^2 = 0$. With the modified method, both the coefficient estimates and their estimated standard errors are increased slightly over those obtained with the conventional estimator.

A PROBIT MODEL

Again we start with equation 15 but now suppose $F = \Phi$ the standard normal c.d.f., and assume that ϵ_j has a normal distribution with mean zero and constant variance σ^2. Amemiya and Nold (1975) suggested that a WLS estimator for the probit case could be developed in much the same way as for the logit case, but they did not pursue that suggestion. An obvious alternative is maximum likelihood estimation. It turns out that the model under consideration is a special case of multivariate probit models described by Ashford and Sowden (1970) and Muthén (1979) who also discussed ML estimation. Unfortunately, the computational difficulty of evaluating multinormal distribution functions makes ML estimation prohibitively expensive at present,

TABLE 2
Estimates of Logit Models for Data in Table 1

	Intercept[a]	Slope[a]
1. Conventional ML	-.7871 (.1748)	.07292 (.01582)
2. Conventional WLS	-.6786 (.1913)	.06124 (.01764)
3. Internal WLS	-.7008 (.2140)	.07496 (.02077)
4. External WLS	-.6543 (.2398)	.07385 (.02418)
5. External ML	-.7460 (.2127)	.08340 (.02109)

a. Standard errors in parentheses.

except in a few special cases, for example, $n_j = 2$ for all j. However, for the general case, Kiefer (1982) has proposed a practical test of the hypothesis that $\sigma^2 = 0$, which may be used to determine whether or not there is dependence among the observations.

EXTERNAL MODELS FOR GROUPED, DEPENDENT DATA

We now consider methods for estimating models that have the general form of equation 16. A well-known special case is the beta-binomial model, which has been applied to data on unemployment (Heckman and Willis, 1977), psychological testing (Wilcox, 1981), marketing (Given, 1980), and toxicology (Haseman and Kupper, 1979).

The beta-binomial model assumes that $F(\boldsymbol{\beta}\mathbf{x}_j) + \epsilon_j$ has a beta distribution (Williams, 1979), with a mean of $\pi_j = F(\boldsymbol{\beta}\mathbf{x}_j)$ and a variance of $\pi_j(1 - \pi_j)\phi_j$, where $0 \leq \phi_j \leq 1$. Note that this indirectly specifies the distribution of ϵ_j. It follows that R_j has a beta-binomial distribution. When $\mathbf{x}_j$ contains quantitative variables, it is customary to impose the restriction that $\phi_j = \phi$ for all j, which is analogous to the assumption of homoscedasticity in linear models (Crowder, 1978; Williams, 1982). Under this specification, estimates of $\boldsymbol{\beta}$ and ϕ may be obtained by ML (Crowder, 1978).

MAXIMUM LIKELIHOOD

For the data in Table 1, I assumed a beta-binomial regression model with the additional specification that F(.) is the logistic function. I then estimated $\boldsymbol{\beta}$ and ϕ by ML using a Newton-Raphson algorithm programmed in the PROC MATRIX language of SAS.[7] Coefficient estimates and their estimated standard errors are shown in line 5 of Table 2. These estimates should be compared with those in line 1 obtained by ML under the assumption that $\phi = 0$, which is just the conventional logit model with no disturbance term. The coefficient for NIH funding is about 14% larger for the beta-binomial model, while the estimated standard error is about 33% larger. The estimate for ϕ is .0508 with an estimated standard error of .0266.

Since the conventional model is a special case of the beta-binomial model, a likelihood-ratio chi-square test of the hypothesis that $\phi = 0$ is obtained by taking twice the positive difference between the log-likelihoods under the two models. The log-likelihoods are –306.30 for conventional ML and –300.92 for modified ML, yielding a chi-square of 10.76 with 1 degree of freedom. This indicates that there *is* dependence, and that the beta-binomial model is preferable.

WEIGHTED LEAST SQUARES

Since ML estimation of the beta-binomial model is computationally demanding and not available in standard software packages, an attractive alternative is WLS estimation. The model just estimated assumed that $\pi_j = F(\boldsymbol{\beta}\mathbf{x}_j)$ with F specified as the logistic function. It also implied that $\text{var}(\epsilon_j) = \pi_j(1 - \pi_j)\phi$. We shall retain those two conditions, but will relax the assumption that $F(\boldsymbol{\beta}\mathbf{x}_j) + \epsilon_j$ has a beta distribution. As with the Amemiya-Nold estimator, the working dependent variable is the empirical logit,[8] $U_j = \ln[R_j/(n_j - R_j)]$. The aim is to regress this variable on the explanatory variables, weighting inversely by the variance of U_j.

Using results in Kleinman (1973) together with the delta method (Bishop, Fienberg, and Holland, 1975: 486ff.), it can be shown that the approximate variance (as n_j gets large) of U_j is

$$\frac{1 + (n_j - 1)\phi}{n_j \pi_j (1 - \pi_j)} \qquad [18]$$

A consistent estimator of expression 18 is needed in order to construct the weights for the regression.

A consistent estimator of π_j is just $\hat{P}_j = R_j / n_j$. In general, one also needs a consistent estimator of ϕ. However, in the special case in which all the n_j's are equal, the numerator in expression 18 is a constant and can therefore be ignored in constructing the weights. In this case, the WLS estimator of $\boldsymbol{\beta}$ is gotten by choosing $\boldsymbol{\beta}$ to minimize

$$\Sigma_j w_j (U_j - \boldsymbol{\beta} \mathbf{x}_j)^2 \qquad [19]$$

where $w_j = \hat{P}_j(1 - \hat{P}_j)$. But this is just the conventional WLS estimator for the logit model when all the n_j's are equal (Berkson, 1944); hence no special procedures are necessary for estimating the coefficient vector. For the conventional WLS estimator, the covariance matrix for the coefficients is ordinarily estimated by $(1/n)(\Sigma_j w_j \mathbf{x}_j \mathbf{x}_j')^{-1}$ where n is the common number of observations in each group (Berkson, 1953). But this is *not* a consistent estimator when $\phi > 0$. A consistent estimator is given below in equation 22.

Estimation is more difficult when groups are unequal in size since a consistent estimator of ϕ is necessary for constructing the weights. Such an estimator can be rather simply obtained from the Pearson chi-square statistic for the goodness-of-fit of the binomial model (Brier, 1980; Williams, 1982). Let X^2 be the chi-square statistic from fitting the conventional logit model by WLS, that is,

$$X^2 = \Sigma_j n_j (\hat{P}_j - \hat{\pi}_j)^2 / [\hat{\pi}_j (1 - \hat{\pi}_j)] \qquad [20]$$

where $\hat{\pi}_j = 1/(1 + \exp[-\hat{\boldsymbol{\beta}} \mathbf{x}_j])$. Let $\tilde{n}$ be the arithmetic mean of the n_j's, and let K be the number of estimated coefficients. Then

$$\tilde{\phi} = \frac{X^2 - (J - K)}{(\tilde{n} - 1)(J - K)}$$

where J is the number of groups. Note that since $\tilde{\phi}$ is an increasing function of X^2, the usual chi-square test can be interpreted as a test for the hypothesis that $\phi = 0$.

The modified WLS estimator is then obtained by choosing $\boldsymbol{\beta}$ to minimize expression 19 with $w_j = n_j\hat{P}_j(1 - \hat{P}_j)/(1 + [n_j - 1]\tilde{\phi})$. The estimated covariance matrix[9] of $\boldsymbol{\beta}$ is $(\Sigma_j w_j \mathbf{x}_j \mathbf{x}_{j'})^{-1}$. However, when $n_j = n$ for all j, this reduces to

$$[X^2/(J - K)]\,(1/n)\,[\Sigma_j \hat{P}_j(1 - \hat{P}_j)\mathbf{x}_j\mathbf{x}_{j'}]^{-1}, \qquad [22]$$

which is the conventional estimator (discussed earlier) inflated by a factor that is the chi-square statistic divided by its degrees of freedom.

To sum up, the WLS procedures can be described by the following steps:

(1) Fit the model by conventional WLS. If the chi-square statistic is not significant, the conventional model is adequate. If it is significant, proceed to the next steps.
(2) If the group sizes are equal (or approximately so), the conventional WLS coefficients are satisfactory. However, the standard errors should be inflated $[X^2/\text{d.f.}]^{1/2}$.
(3) If the group sizes are not equal, estimate ϕ by equation 21 and construct the weights $R_j(1 - R_j/n_j)/[1 + (n_j - 1)\tilde{\phi}]$. Refit the model with the new weights.

Using the GLIM program (Baker and Nelder, 1978), these steps were applied to the data in Table 1. The conventional WLS estimates resulted in a chi-square of 60.66 with 38 d.f. (line 2 of Table 2) yielding a p-value of about .01. This suggests that, in fact, $\phi > 0$. Applying equation 21, we get an estimate for ϕ of .0569 (which is quite close to the ML estimate of .0508). Using this estimate to construct new weights, we get the results in line 4 of Table 2. The coefficient for NIH funding is about 21% higher than that obtained from conventional WLS estimation, but about 11% lower than the ML estimate. The estimated standard error is slightly higher than that for the ML estimate.

QUASI-LIKELIHOOD ESTIMATION

Earlier we saw that when all the n_j's are equal, the WLS procedure for an external disturbance model reduces to the conventional WLS estimator for a grouped logit model. A similar result holds for ML estimation. Wedderburn (1974) introduced a general method of quasi-likelihood estimation that is applicable whenever the variance of the dependent variable is proportional to some known function of the mean. Quasi-likelihood estimators share many properties with ordinary ML estimators, but they are not always efficient. For the beta-binomial model, the condition on the mean and variance is satisfied when the group sizes are constant. In this case, the quasi-likelihood estimator of $\boldsymbol{\beta}$ is identical to the ML estimator for the model that assumes that $\phi = 0$, that is, the conventional ML estimator for a logit regression model. A consistent estimator for the covariance matrix of $\boldsymbol{\beta}$ is obtained by multiplying the conventional covariance matrix by a factor that can be estimated by the Pearson chi-square statistic for the fitted model divided by its degrees of freedom.[10] This is the same adjustment suggested earlier for the WLS estimators in the case of equal group sizes. Williams (1982) has extended this result to the case where the group sizes are not equal.

CONCLUSIONS

There are several different ways to introduce a disturbance into logit and probit regression models, and each model has somewhat different implications. Nevertheless, the models share some properties that have been confused in previous discussions of the problem.

Consider first the case in which all the dichotomous observations are independent. For the external logit and probit models, and for the internal probit model, the form of the likelihood function is unchanged by the introduction of the disturbance term. While the likelihood function is altered for the internal logit model, we saw that the effect of that change is so slight as to be of little practical importance relative to other sources of misspecification. Thus standard estimation procedures should be quite satisfactory for all these models, and there is no reason to expect an increase in the variance of the observed counts. Nevertheless, for the internal models, the disturbance variance leads to a downscaling of the coefficient vector, and this can invalidate cross-population comparisons. Since there is little basis for choosing between the internal and external models, caution is advisable in making such comparisons.

Models appropriate for grouped data with dependent observations within groups yield quite different results. The observed counts are no longer binomially distributed and, in particular, the variance exceeds the binomial variance. Although standard estimation procedures are inefficient in such cases, several efficient methods are available for both internal and external models. The simplest method is probably the WLS estimator for the external logit model; it is easily computed and reduces to the conventional WLS estimator when the groups are equal in size.

NOTES

1. To simplify the notation, the **x** vector is treated as a set of fixed constants rather than as a random vector. Nevertheless, all the results that follow apply equally to the random case. To include an intercept term, the first element of **x** should be 1.

2. In one study, I used numerical integration to generate predicted probabilities based on equation 7 for six different values of σ and a wide range of x's. These were then compared with probabilities based on the approximation given by equations 13 and 9. In all but one of the 90 comparisons, the approximation was within .01 of the true value. In the second study, I generated data based on equation 7 and estimated a standard logit model by maximum likelihood. Even with extremely large samples (e.g., 15,000) the logit model fit extremely well, and the estimated coefficients were very close to those predicted by equation 9. Further details are available from the author.

3. Chamberlain (1980) discusses the case in which some of the explanatory variables vary across individuals and there are equal numbers of individuals in all groups.

4. Even when the observations are independent, the observed variation in R_j will often exceed that expected under the binomial model. This is merely an indication that there is something else wrong with the model, for example, nonlinearities or unspecified interactions.

5. A common recommendation is to adjust the empirical logit to be $\ln[(R_j + ½)/(n_j - R_j + ½)]$. However, Cox (1970) argues that this adjustment is only appropriate for *unweighted* least squares.

6. Standard regression programs that perform WLS with user-specified weights (like PROC REG in SAS) usually estimate the covariance matrix as $s^2[\Sigma_j w_j \mathbf{x}_j \mathbf{x}_j']^{-1}$ where s^2 is the weighted mean residual sum of squares. To get the correct estimates, divide the elements in the reported covariance matrix by s^2. Equivalently, divide the reported standard errors by s (the root mean squared error). In GLIM, s^2 can be forced equal to 1 by specifying $SCALE 1.

7. A listing or BITNET file for this program is available from the author.

8. See note 5.

9. See note 6.

10. In GLIM, this adjustment can be accomplished by specifying $SCALE 0.

REFERENCES

AMEMIYA, T. and F. NOLD (1975) "A modified logit model." Rev. of Economics and Statistics 57: 255-257.

ASHFORD, J. R. and R. R. SOWDEN (1970) "Multivariate probit analysis." Biometrics 26: 535-546.

BAKER, R. J. and J. A. NELDER (1978) The GLIM System. Oxford: Numerical Algorithms Group.

BERKSON, J. (1944) "Application of the logistic function to bio-assay." J. of the Amer. Stat. Assn. 48: 565-599.

BERKSON, J. (1953) "A statistically precise and relatively simple method of estimating the bio-assay with quantal response based on the logistic function." J. of the Amer. Stat. Assn. 48: 565-599.

BISHOP, Y.M.M., S. E. FIENBERG, and P. W. HOLLAND (1975) Discrete Multivariate Analysis. Cambridge, MA: MIT Press.

BRIER, S. S. (1980) "Analysis of contingency tables under cluster sampling." Biometrika 67: 591-596.

COX, D. R. (1970) Analysis of Binary Data. London: Chapman & Hall.

CROWDER, M. J. (1978) "Beta-binomial ANOVA for proportions." Applied Statistics 27: 34-37.

FINNEY, D. J. (1971) Probit Analysis (3rd ed.). Cambridge, England: Cambridge Univ. Press.

GIVEN, M. (1980) "Determination of optimal sample sizes in the beta-binomial brand choice model." J. of Marketing Research 17: 58-62.

HAKKERT, R. (1978) "Some stochastic alternatives to logit analysis." Cornell University. (unpublished)

HANUSHEK, E. A. and J. E. JACKSON (1977) Statistical Methods for Social Scientists. New York: Academic Press.

HASEMAN, J. K. and L. L. KUPPER (1979) "Analysis of dichotomous response data from certain toxicological experiments." Biometrics 35: 281-293.

HECKMAN, J. J. and R. J. WILLIS (1977) "A beta-logistic model for the analysis of sequential labor force participation by married women." J. of Pol. Economy 85: 27-58.

KIEFER, N. (1982) "Testing for dependence in multivariate probit models." Biometrika 69: 161-166.

KLEINMAN, J. C. (1973) "Proportions with extraneous variance: single and independent samples." J. of the Amer. Stat. Assn. 68: 46-54.

LORD, F. M. and M. R. NOVICK (1968) Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

McGINNIS, R., P. D. ALLISON, and J. S. LONG (1982) "Postdoctoral training in bioscience: allocation and outcomes." Social Forces 60: 701-722.

MUTHÉN, B. (1979) "A structural probit model with latent variables." J. of the Amer. Stat. Assn. 74: 907-911.

PARKS, R. W. (1977) "Determinants of scrapping rates for postwar vintage automobiles." Econometrica 45: 1099-1115.

SCHOENBERG, R. (1985) "Latent variables in the analysis of limited dependent variables," pp. 242-276 in N. B. Tuma (ed.) Sociological Methodology 1985. San Francisco: Jossey-Bass.

WEDDERBURN, R.M.W. (1974) "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton Method." Biometrika 61: 439-447.

WILCOX, R. R. (1981) "A review of the beta-binomial model and its extensions." J. of Educ. Statistics 6: 3-32.

WILLIAMS, D. A. (1979) "The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity." Biometrics 31: 949-952.

WILLIAMS, D. A. (1982) "Extra-binomial variation in logistic linear models." Applied Statistics 31: 144-148.

WINSHIP, C. and R. D. MARE (1983) "Structural equations and path analysis for discrete data." Amer. Soc. Rev. 89: 54-110.

WONNACOTT, T. H. and R. J. WONNACOTT (1984) Introductory Statistics for Business and Economics. New York: John Wiley.

Paul D. Allison is Associate Professor of Sociology at the University of Pennsylvania. In addition to categorical data analysis, his research interests include event history analysis and techniques for missing data.