JOURNAL OF APPLIED ECONOMETRICS, VOL. 11, 619–632 (1996)

ECONOMETRIC METHODS FOR FRACTIONAL RESPONSE VARIABLES WITH AN APPLICATION TO 401(K) PLAN PARTICIPATION RATES

LESLIE E. PAPKE AND JEFFREY M. WOOLDRIDGE
Department of Economics, Michigan State University, Marshall Hall, East Lansing, MI 48824-1038, USA

SUMMARY

We develop attractive functional forms and simple quasi-likelihood estimation methods for regression models with a fractional dependent variable. Compared with log-odds type procedures, there is no difficulty in recovering the regression function for the fractional variable, and there is no need to use *ad hoc* transformations to handle data at the extreme values of zero and one. We also offer some new, robust specification tests by nesting the logit or probit function in a more general functional form. We apply these methods to a data set of employee participation rates in 401(k) pension plans.

1. INTRODUCTION

Fractional response variables arise naturally in many economic settings. The fraction of total weekly hours spent working, the proportion of income spent on charitable contributions, and participation rates in voluntary pension plans are just a few examples of economic variables bounded between zero and one. The bounded nature of such variables and the possibility of observing values at the boundaries raise interesting functional form and inference issues. In this paper we specify and analyse a class of functional forms with satisfying econometric properties. We also synthesize and expand on the generalized linear models (GLM) literature from statistics and the quasi-likelihood literature from econometrics to obtain robust methods for estimation and inference with fractional response variables.

We apply the methods to estimate a model of employee participation rates in 401(k) pension plans. The key explanatory variable of interest is the plan's 'match rate,' the rate at which a firm matches a dollar of employee contributions. The empirical work extends that of Papke (1995), who studied this problem using linear spline methods. Spline methods are flexible, but they do not ensure that predicted values lie in the unit interval.

To illustrate the methodological issues that arise with fractional dependent variables, suppose that a variable y, $0 \leq y \leq 1$, is to be explained by a $1 \times K$ vector of explanatory variables $\mathbf{x} \equiv (x_1, x_2, \ldots, x_K)$, with the convention that $x_1 \equiv 1$. The population model

$$E(y \mid \mathbf{x}) = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K = \mathbf{x}\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector, rarely provides the best description of $E(y \mid \mathbf{x})$. The primary reason is that y is bounded between 0 and 1, and so the effect of any particular x_j cannot be constant throughout the range of $\mathbf{x}$ (unless the range of x_j is very limited). To some extent this problem can be overcome by augmenting a linear model with non-linear functions of $\mathbf{x}$, but the predicted

CCC 0883–7252/96/060619–14

Received 25 October 1993
Revised 19 February 1996

values from an OLS regression can never be guaranteed to lie in the unit interval. Thus, the drawbacks of linear models for fractional data are analogous to the drawbacks of the linear probability model for binary data.

The most common alternative to equation (1) has been to model the log-odds ratio as a linear function. If y is *strictly* between zero and one then a linear model for the log-odds ratio is

$$E(\log[y/(1-y)]\,|\,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} \tag{2}$$

Equation (2) is attractive because $\log[y/(1-y)]$ can take on any real value as y varies between 0 and 1, so it is natural to model its population regression as a linear function. Nevertheless, there are two potential problems with equation (2). First, the equation cannot be true if y takes on the values 0 or 1 with positive probability. Consequently, given a set of data, if any observation y_i equals 0 or 1 then an adjustment must be made before computing the log-odds ratio. When the y_i are proportions from a fixed number of groups with known group sizes, adjustments are available in the literature—see, for example, Maddala (1983, p. 30). Estimation of the log-odds model then corresponds to Berkson's minimum chi-square method.

Unfortunately, the minimum chi-square method for a fixed number of categories is not applicable to certain economic problems. First, the fraction y may not be a proportion from a discrete group size—for example, y_i could be the fraction of county land area containing toxic waste dumps, or the proportion of income given in charitable contributions. Second, one may be hesitant to adjust the extreme values in the data if a large percentage is at the extremes. In our application to 401(k) plan participation rates, about 40% of the y_i takes on the value unity. It seems more natural to treat such examples in a regression-type framework.

Even when model (2) is well defined, there is still a problem. Without further assumptions, we cannot recover $E(y\,|\,\mathbf{x})$, which is our primary interest. Under model (2) the expected value of y given $\mathbf{x}$ is

$$E(y\,|\,\mathbf{x}) = \int_{-\infty}^{\infty} \left(\frac{\exp(\mathbf{x}\boldsymbol{\beta} + v)}{1 + \exp(\mathbf{x}\boldsymbol{\beta} + v)} \right) f(v\,|\,\mathbf{x})\, dv \tag{3}$$

where $f(\cdot\,|\,\mathbf{x})$ denotes the conditional density of $u \equiv \log[y/(1-y)] - \mathbf{x}\boldsymbol{\beta}$ given $\mathbf{x}$ and v is a dummy argument of integration. Even if u and $\mathbf{x}$ are assumed to be independent, $E(y\,|\,\mathbf{x}) \neq \exp(\mathbf{x}\boldsymbol{\beta})]/[1 + \exp(\mathbf{x}\boldsymbol{\beta})]$, although $E(y\,|\,\mathbf{x})$ can be estimated using, for example, Duan's (1983) smearing method. If u and $\mathbf{x}$ are not independent, model (3) cannot be estimated without estimating $f(\cdot\,|\,\mathbf{x})$. This is either difficult or non-robust, depending on whether a non-parametric or a parametric approach is adopted. Instead, we prefer to specify models for $E(y\,|\,\mathbf{x})$ directly, without having to estimate the density of u given $\mathbf{x}$.

Naturally, it is always possible to estimate $E(y\,|\,\mathbf{x})$ by assuming a particular distribution for y given $\mathbf{x}$ and estimating the parameters of the conditional distribution by maximum likelihood. One plausible distribution for fractional y is the beta distribution; Mullahy (1990) suggests this as one possible approach. Unfortunately, the estimates of $E(y\,|\,\mathbf{x})$ that one obtains are known not to be robust to distributional failure (this follows from Gourieroux, Monfort, and Trognon(1984); more on this below). Clearly, standard distributional assumptions can fail in certain applications. One important limitation of the beta distribution is that it implies that each value in $[0, 1]$ is taken on with probability zero. Thus, the beta distribution is difficult to justify in applications where at least some portion of the sample is at the extreme values of zero or one.

In the next section we specify a reasonable class of functional forms for $E(y\,|\,\mathbf{x})$ and show how to estimate the parameters using Bernoulli quasi-likelihood methods. These functional forms and estimators circumvent the problems raised above and are easily implemented. Some

new specification tests are offered in Section 3, and Section 4 contains the empirical application relating 401(k) plan participation rates to the plan's matching rate and other plan characteristics.

2. FUNCTIONAL FORMS AND QUASI-LIKELIHOOD METHODS

We assume the availability of an independent (though not necessarily identically distributed) sequence of observations $\{(\mathbf{x}_i, y_i): i = 1, 2, \ldots, N\}$, where $0 \leq y_i \leq 1$ and N is the sample size. The asymptotic analysis is carried out as $N \to \infty$. Our maintained assumption is that, for all i,

$$E(y_i | \mathbf{x}_i) = G(\mathbf{x}_i\beta) \tag{4}$$

where $G(\cdot)$ is a known function satisfying $0 < G(z) < 1$ for all $z \in \mathbb{R}$. This ensures that the predicted values of y lie in the interval $(0, 1)$. Equation (4) is well defined even if y_i can take on 0 or 1 with positive probability. Typically, $G(\cdot)$ is chosen to be a cumulative distribution function (cdf), with the two most popular examples being $G(z) \equiv \Lambda(z) \equiv \exp(z)/[1 + \exp(z)]$—the logistic function—and $G(z) \equiv \Phi(z)$, where $\Phi(\cdot)$ is the standard normal cdf. However, $G(\cdot)$ need not even be a cdf in what follows.

In stating equation (4) we make no assumption about an underlying structure used to obtain y_i. In the special case that y_i is a proportion from a group of known size n_i, the methods in this paper ignore the information on n_i. There are some advantages to ignoring n_i. First, one does not always want to condition on n_i, in which case y_i contains all relevant information. Second, the methods here are computationally simple. Third, under the assumptions we impose, the method suggested here need not be less efficient than methods that use information on group size. (See Papke and Wooldridge (1993) for methods that incorporate information on n_i in a similar framework.)

We have stated the functional form directly in terms of $E(y_i | \mathbf{x}_i)$, where $\mathbf{x}_i$ is observable. Stating the model of interest in terms of $E(y_i | \mathbf{x}_i, \theta_i)$, where θ_i is unobserved heterogeneity independent of $\mathbf{x}_i$, requires one to specify a distribution for θ_i in order to obtain $E(y_i | \mathbf{x}_i)$ (which is ultimately of interest in any case). Generally, although not always, this will lead to a different functional form from equation (4). Allowing for functional forms other than the index structure in equation (4) may be worth-while, but it is not within the scope of this paper. In Section 3 we present a general functional form test that has power against a variety of functional form misspecifications, including those that arise from models of unobserved heterogeneity.

Under equation (4), β can be consistently estimated by non-linear least squares (NLS). The fact that equation (4) is non-linear in β is perhaps the leading reason a linear model for y_i or for the log-odds ratio is used in applied work. Further, heteroscedasticity is likely to be present since $\mathrm{Var}(y_i | \mathbf{x}_i)$ is unlikely to be constant when $0 \leq y_i \leq 1$. Obtaining the NLS estimates and heteroscedasticity-robust standard errors and test statistics requires special programming, and the NLS estimator will not have any efficiency properties when $\mathrm{Var}(y_i | \mathbf{x})$ is not constant. Still, the motivation underlying NLS is sound because it directly estimates $E(y | \mathbf{x})$. See also Mullahy (1990), who suggests NLS for continuously distributed outcomes on a bounded interval.

The estimation procedure we propose is a particular quasi-likelihood method, as in Gourieroux, Monfort, and Trognon (1984) (hereafter GMT) and McCullagh and Nelder (1989) (hereafter MN). The Bernoulli log-likelihood function, given by

$$l_i(\mathbf{b}) \equiv y_i \log[G(\mathbf{x}_i\mathbf{b})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\mathbf{b})] \tag{5}$$

is well defined for $0 < G(\cdot) < 1$ and is attractive for several reasons. First, maximizing the Bernoulli log-likelihood is easy. Second, because equation (5) is a member of the linear exponential family (LEF), the quasi-maximum likelihood estimator (QMLE) of β, obtained

from the maximization problem

$$\max_{\mathbf{b}} \sum_{i=1}^{N} l_i(\mathbf{b})$$

is consistent for $\boldsymbol{\beta}$ provided that equation (4) holds. (This follows from GMT (1984) and is also easily seen by computing the score $\mathbf{s}_i(\mathbf{b}) \equiv \nabla_{\boldsymbol{\beta}} l_i(\mathbf{b})'$ and showing that $E[\mathbf{s}_i(\boldsymbol{\beta}) \mid \mathbf{x}_i] = 0$ under equation (4).) In other words, the Bernoulli, QMLE $\hat{\boldsymbol{\beta}}$ is consistent and $\sqrt{N}$-asymptotically normal *regardless* of the distribution of y_i conditional on $\mathbf{x}_i$; y_i could be a continuous variable, a discrete variable, or have both continuous and discrete characteristics. As we will see below, in some cases for fractional data the Bernoulli QMLE is efficient in a class of estimators containing all QMLEs in the LEF and weighted NLS estimators.

A special case of equation (5)—namely, when $G(\cdot)$ is the logistic function—has been suggested in the GLM framework by MN (1989). The GLM approach has two drawbacks for economic applications. First, for the logit QMLE it assumes that

$$\text{Var}(y_i \mid \mathbf{x}_i) = \sigma^2 G(\mathbf{x}_i\boldsymbol{\beta})[1 - G(\mathbf{x}_i\boldsymbol{\beta})] \text{ for some } \sigma^2 > 0 \tag{6}$$

where $G(\cdot) = \Lambda(\cdot)$. While we prefer equation (6) as a nominal variance assumption to the nominal NLS homoscedasticity assumption $\text{Var}(y_i \mid \mathbf{x}_i) = \sigma^2$, imposing any particular conditional variance when performing inference is too restrictive. Mechanisms for which equation (6) fails are common and are related to the literature on binary choice models with over-dispersion; see, for example, MN (1989, section 4.5). Briefly, if each y_i is computed as the average of n_i independent binary variables, say y_{ij}, such that $P(y_{ij} = 1 \mid \mathbf{x}_i, n_i) = G(\mathbf{x}_i\boldsymbol{\beta})$, then it can be shown that

$$\text{Var}(y_i \mid \mathbf{x}_i) = E(n_i^{-1} \mid \mathbf{x}_i)\text{G}(\mathbf{x}_i\boldsymbol{\beta})[1 - G(\mathbf{x}_i\boldsymbol{\beta})]$$

Unless n_i and $\mathbf{x}_i$ are independent, equation (6) generally fails. In our application, where y_{ij} is a binary indicator for whether worker j at firm i contributes to a 401(k) plan, n_i is the number of workers at firm i, and $\mathbf{x}_i$ contains firm characteristics, n_i and $\mathbf{x}_i$ are unlikely to be independent in the population. In addition, equation (6) can fail if there are unobserved group effects. Notice, however, that neither of these situations invalidates equation (4), which is all that is needed to consistently estimate $\boldsymbol{\beta}$ using the Bernoulli QMLE.

The second drawback to the GLM approach is related to the first: if the variance assumption (6) fails, MN (1989, p. 330) reject the logit quasi-likelihood approach and suggest a more complicated quasi-likelihood. But this begs the issue of whether the conditional mean model (4) is appropriate. Here, we are primarily interested in the conditional mean. Rather than abandoning the Bernoulli QMLE because equation (6) might fail, we propose asymptotically robust inference for the conditional mean parameters.

To find the asymptotic variance of the Bernoulli QMLE, define $g(z) \equiv \text{d}G(z)/\text{d}z$, $\hat{G}_i \equiv G(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \equiv \hat{y}_i$, and $\hat{g}_i \equiv g(\mathbf{x}_i\hat{\boldsymbol{\beta}})$. Then the estimated information matrix is

$$\hat{\mathbf{A}} \equiv \sum_{i=1}^{N} \frac{\hat{g}_i^2 \mathbf{x}_i' \mathbf{x}_i}{[\hat{G}_i(1 - \hat{G}_i)]} \tag{7}$$

The standard error of $\hat{\beta}_j$ reported from standard binary response analysis (regardless of the nature of y_i) would be obtained as the square root of the j^{th} diagonal element of $\hat{\mathbf{A}}^{-1}$. Under equation (4) only, this is not a consistent estimator of the true asymptotic standard error; we also need the outer product of the score. Let $\hat{u}_i \equiv y_i - G(\mathbf{x}_i\hat{\boldsymbol{\beta}})$ be the residuals (deviations

between y_i and its estimated conditional expectation), and define

$$\hat{\mathbf{B}} \equiv \sum_{i=1}^{N} \frac{\hat{u}_i^2 \hat{g}_i^2 \mathbf{x}_i' \mathbf{x}_i}{[\hat{G}_i(1-\hat{G}_i)]^2} \tag{8}$$

Then a valid estimate of the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1} \tag{9}$$

The standard errors are obtained as the square roots of the diagonal elements of equation (9); see GMT (1984) and Wooldridge (1991b) for general treatments.

Interestingly, the robust standard errors from equation (9) in the context of ordinary logit and probit are computed almost routinely by certain statistics and econometrics packages, such as STATA® and SST®. Unfortunately, the packages with which we are familiar automatically transform the dependent variable used in logit or probit into a binary variable before estimation, or do not allow non-binary variables at all (STATA® and SST® fall into the first category). With the minor change of allowing for fractional y in so-called binary response analysis, standard software packages could be used to estimate the parameters in equation (4) and to perform asymptotically valid inference. Alternatively, programming the estimator in a language such as GAUSS®, as we do for our application in Section 4, is fairly straightforward.

If the GLM assumption (6) is maintained in addition to (4) then σ^2 is consistently estimated by

$$\hat{\sigma}^2 = (N-K)^{-1} \sum_{i=1}^{N} \tilde{u}_i^2 \tag{10}$$

where $\tilde{u}_i$ are the *weighted* residuals (sometimes called the *Pearson residuals*):

$$\tilde{u}_i \equiv \hat{u}_i / [\hat{G}_i(1-\hat{G}_i)]^{1/2} \tag{11}$$

(It is standard practice in the GLM literature to use the degrees-of-freedom adjustment in equation (10) in estimating σ^2.) Then the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is estimated as $\hat{\sigma}^2\hat{\mathbf{A}}^{-1}$; see also MN (1989, p. 327). In addition, under equation (6) $\mathrm{Var}(y_i|\mathbf{x}_i)$ is proportional to the variance in the Bernoulli distribution, and so by the results of GMT (1984), the Bernoulli QMLE is efficient in the class of QMLEs in the LEF. This is essentially the same as the class of all weighted NLS estimators, and so it is a non-trivial efficiency result.

To summarize, we have chosen a functional form that ensures estimates of $\mathrm{E}(y|\mathbf{x})$ are between zero and one, and a quasi-likelihood function that leads to a relatively efficient QMLE under a popular auxiliary assumption—namely, equation (6). In addition, we guard against failure of this variance assumption by using equation (9) as the variance estimator. In the next section we suggest specification tests that are valid with and without equation (6).

3. SPECIFICATION TESTING

Specification testing in this framework can be carried out by applying the results of Wooldridge (1991a,b). We discuss two forms of the test. The first is valid under equations (4) and (6); these are non-robust tests because they maintain the GLM variance assumption. The second, robust form of the test requires only equation (4).

We focus primarily on Lagrange multiplier or score tests that nest $E(y|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\beta})$ within a more general model. Let $m(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ be a model for $E(y|\mathbf{x}, \mathbf{z})$, where $\mathbf{z}$ is a $1 \times J$ vector of

additional variables; the elements of $\mathbf{z}$ can be non-linear functions of $\mathbf{x}$ (in which case $E(y|\mathbf{x}) = E(y|\mathbf{x},\mathbf{z})$), or variables not functionally related to $\mathbf{x}$, or both. The vector γ is a $Q \times 1$ vector of additional parameters. The null is assumed to be H_0: $\gamma = \gamma_0$ for a specified vector γ_0 (often $\gamma_0 = \mathbf{0}$). Then, by definition,

$$G(\mathbf{x}\beta) \equiv m(\mathbf{x}, \mathbf{z}, \beta, \gamma_0) \tag{12}$$

Given the estimates under the null, $\hat{\beta}$, define the $1 \times K$ vector $\nabla_\beta \hat{m}_i \equiv \partial m(\mathbf{x}_i, \mathbf{z}_i, \hat{\beta}, \gamma_0)/\partial\beta = \hat{g}_i\mathbf{x}_i$ and the $1 \times Q$ vector $\nabla_\gamma \hat{m}_i \equiv \partial m(\mathbf{x}_i, \mathbf{z}_i, \hat{\beta}, \gamma_0)/\partial\gamma$; these are the gradients of the regression function with respect to β and γ, respectively, evaluated under the null hypothesis. Define the weighted residuals $\tilde{u}_i$ as in equation (11) and the weighted gradients as

$$\nabla_\beta \tilde{m}_i = \nabla_\beta \hat{m}_i/[\hat{G}_i(1-\hat{G}_i)]^{1/2} = \hat{g}_i\mathbf{x}_i/[\hat{G}_i(1-\hat{G}_i)]^{1/2} \tag{13}$$

$$\nabla_\gamma \tilde{m}_i \equiv \nabla_\gamma \hat{m}_i/[\hat{G}_i(1-\hat{G}_i)]^{1/2} \tag{14}$$

As in equation (11), the weights are proportional to the inverse of the estimated nominal standard deviation (see equation (6)). A valid test of H_0: $\gamma = \gamma_0$ depends on what is maintained under the null hypothesis. Under the assumptions

$$E(y_i|\mathbf{x}_i, \mathbf{z}_i) = G(\mathbf{x}_i\beta) \tag{15}$$

and

$$\mathrm{Var}(y_i|\mathbf{x}_i, \mathbf{z}_i) = \sigma^2 G(\mathbf{x}_i\beta)[1 - G(\mathbf{x}_i\beta)] \tag{16}$$

a valid statistic is obtained as NR_u^2 from the OLS regression

$$\tilde{u}_i \text{ on } \nabla_\beta \tilde{m}_i, \nabla_\gamma \tilde{m}_i \qquad i = 1, 2, \ldots, N \tag{17}$$

where R_u^2 is the constant-unadjusted r-squared. Under equations (15) and (16), NR_u^2 is distributed asymptotically as χ_Q^2—see Wooldridge (1991a).

For binary choice models, Engle (1984) and Davidson and MacKinnon (1984) suggest a test based on regression (17) for logit and probit. Gurmu and Trivedi (1993) present results for a class of models that allows testing the logit function against a more general index function. But for fractional dependent variables it is important to use the NR_u^2 form rather than the explained sum of squares form suggested in Davidson and MacKinnon (1984): the latter test requires $\sigma^2 = 1$, which is always the case for binary response variables but is too restrictive for fractional response variables. Alternatively, as in Gurmu and Trivedi (1993), each term in regression (17) can be divided by $\hat{\sigma}$ and then the explained sum of squares can be used. This is essentially the same as the NR_u^2 statistic (although they will differ if $\hat{\sigma}$ is estimated with the degrees-of-freedom adjustment in equation (10)).

It is often useful to have a likelihood-based statistic, especially for testing exclusion restrictions. Under the same two assumptions (15) and (16), a quasi-likelihood ratio (QLR) statistic has a limiting chi-square distribution. Let $\mathcal{L}_N(\hat{\beta}, \gamma_0)$ denote the log-likelihood evaluated under the null, and let $\mathcal{L}_N(\ddot{\beta}, \ddot{\gamma})$ denote the log-likelihood from the unrestricted model (that is, the Bernoulli log-likelihood with $m(\mathbf{x}, \mathbf{z}, \beta, \gamma)$ used in place of $G(\mathbf{x}_i\beta)$). Further, define $\ddot{m}_i \equiv m(\mathbf{x}_i, \mathbf{z}_i, \ddot{\beta}, \ddot{\gamma})$, and let the variance estimator based on the unrestricted estimates be

$$\ddot{\sigma}^2 \equiv (N - K - Q)^{-1} \sum_{i=1}^{N} (y_i - \ddot{m}_i)^2/[\ddot{m}_i(1 - \ddot{m}_i)] \tag{18}$$

(note that the summation is simply the sum of weighted squared residuals from the unrestricted

model). Then the QLR statistic, defined by

$$\text{QLR} \equiv 2[\mathcal{L}_N(\ddot{\boldsymbol{\beta}}, \ddot{\boldsymbol{\gamma}}) - \mathcal{L}_N(\hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}_0)]/\ddot{\sigma}^2 \tag{19}$$

is distributed asymptotically as χ^2_Q under the null hypothesis, provided equation (16) holds in addition to (15). The validity of this statistic follows because the usual information matrix equality holds up to the scalar σ^2 when the conditional mean and conditional variance are correctly specified.

A form of the LM statistic that is valid under equation (15) alone can be computed from an additional regression. First, regress $\nabla_\gamma \tilde{m}_i$ on $\nabla_\beta \tilde{m}_i$ and save the $1 \times Q$ residuals, $\tilde{\mathbf{r}}_i \equiv (\tilde{r}_{i1}, \tilde{r}_{i2}, \ldots, \tilde{r}_{iQ})$, $i = 1, 2, \ldots, N$. (This is the same as regressing each element of $\nabla_\gamma \tilde{m}_i$ on the entire vector $\nabla_\beta \tilde{m}_i$, and collecting the residuals.) Next, obtain the $1 \times Q$ vector $\tilde{u}_i \tilde{\mathbf{r}}_i = (\tilde{u}_i \tilde{r}_{i1}, \tilde{u}_i \tilde{r}_{i2}, \ldots, \tilde{u}_i \tilde{r}_{iQ})$. The robust LM statistic is obtained as $N - \text{SSR}$, where SSR is the usual sum of squared residuals from the auxiliary regression of unity on $\tilde{u}_i \tilde{\mathbf{r}}_i$:

$$1 \text{ on } \tilde{u}_i \tilde{\mathbf{r}}_i \qquad i = 1, \ldots, N \tag{20}$$

Under H_0, which is equation (16) in this case, $N - \text{SSR} \overset{a}{\sim} \chi^2_Q$. The validity of this procedure is discussed further in Wooldridge (1991a,b). Briefly, $N - \text{SSR}$ from equation (20) is a quadratic form in the vector $N^{-1/2} \sum_{i=1}^N \tilde{\mathbf{r}}_i' \tilde{u}_i$, with a weighting matrix that is the inverse of a consistent estimator of its asymptotic variance whether or not equation (16) holds.

In testing for omitted variables, one can use the QLR statistic or the usual LM statistic under equations (15) and (16), or the robust LM statistic under equation (15) only. (Of course, Wald statistics can also be defined for these two cases, but they are computationally more cumbersome than the QLR and LM statistics.) For omitted variables tests, $m(\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = G(\mathbf{x}_i\boldsymbol{\beta} + \mathbf{z}_i\boldsymbol{\gamma})$, $\nabla_\gamma \hat{m}_i = \hat{g}_i \mathbf{z}_i = g(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \cdot \mathbf{z}_i$, and $\nabla_\gamma \tilde{m}_i = \hat{\mathbf{g}}_i \mathbf{z}_i / [\hat{G}_i(1 - \hat{G}_i)]^{1/2}$. One way to test for functional form is to define $\mathbf{z}_i$ as polynomials, interactions, or other functions of $\mathbf{x}_i$.

A general functional form diagnostic is obtained by extending Ramsey's (1969) RESET procedure to index models. For example, let the alternative model be

$$E(y_i \mid \mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\beta} + \gamma_1(\mathbf{x}_i\boldsymbol{\beta})^2 + \gamma_2(\mathbf{x}_i\boldsymbol{\beta})^3) \tag{21}$$

where, again, $G(\cdot)$ is typically the logistic function or the standard normal cdf. This alternative functional form (or including even higher powers of $\mathbf{x}_i\boldsymbol{\beta}$) can be motivated quite generally. Since $G(\cdot)$ is a strictly increasing function in most applications, any index model of the form $E(y_i \mid \mathbf{x}_i) = H(\mathbf{x}_i\boldsymbol{\beta})$ for unknown H can be arbitrarily well approximated by $G(\sum_{h=1}^J \gamma_j(\mathbf{x}_i\boldsymbol{\beta})^j)$ for J large enough (by standard approximation results for polynomials). Since models with unobserved heterogeneity of the form $E(y_i \mid \mathbf{x}_i, \theta_i) = G(\mathbf{x}_i\boldsymbol{\beta} + \theta_i)$, where θ_i is independent of $\mathbf{x}_i$, have an index structure, a test of the null model against equation (21) should have power for alternatives that can be derived explicitly from models of unobserved heterogeneity. In practice, the first few terms in the expansion are the most important, and we use only the quadratic and cubic terms.

In the context of equation (21), the hypothesis that equation (15) holds (with $\mathbf{z}_i = \mathbf{x}_i$) is stated as H_0: $\gamma_1 = 0$, $\gamma_2 = 0$. This is easily tested using the LM procedures outlined above. (By contrast, the QLR statistic is computationally difficult as well as nonrobust.) First, estimate the model under the assumption $\gamma_1 = \gamma_2 = 0$, as is always done. Define $\hat{\boldsymbol{\beta}}$, $\hat{G}_i$, $\hat{g}_i$, $\hat{u}_i$, $\nabla_\beta \tilde{m}_i$, and $\tilde{u}_i$ as before. The gradient with respect to $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2)'$ is $\nabla_\gamma \hat{m}_i = \{\hat{\mathbf{g}}_i \cdot (\mathbf{x}_i\hat{\boldsymbol{\beta}})^2, \hat{\mathbf{g}}_i \cdot (\mathbf{x}_i\hat{\boldsymbol{\beta}})^3\}$, and $\nabla_\gamma \tilde{m}_i$ is defined in equation (14). The statistic obtained from regression (17) is distributed approximately as χ^2_2 under (15) and (16). The robust form is obtained from regression (20).

4. EMPIRICAL APPLICATION: PARTICIPATION IN 401(k) PENSION PLANS

401(k) plans differ from traditional employer-sponsored pension plans in that employees are permitted to make pre-tax contributions and the employer may match part of the contribution. Since participation in these plans is voluntary, the sensitivity of participation to plan characteristics—specifically the employer matching rate—will play a critical role in retirement saving.

Pension plan administrators are required to file Form 5500 annually with the Internal Revenue Service, describing participation and contribution behavior for each plan offered. Papke (1995) uses the plan level data to study, among other things, the relationship between the participation rate and various plan characteristics, including the rate at which a firm matches employee contributions. Papke (1995) also contains a discussion of the theoretical underpinnings relating participation and the size of the match rate. Not surprisingly, under standard assumptions on the utility function, participation is positively related to the match rate.

The participation rate (*PRATE*) is constructed as the number of active accounts divided by the number of employees eligible to participate. An active account is any existing 401(k) account—a contribution need not have been made that plan year. The plan match rate (*MRATE*) is not reported directly on Form 5500, but can be approximated by the ratio of employer to employee contributions for plans that provide some matching. This calculated match rate may exceed the plan's marginal rate because employer contributions include any flat per participant contribution or any helper contribution made to pass anti-discrimination tests. While the calculated match rate exceeds the marginal incentive facing each saver, it may be a better indicator of overall plan generosity. See Papke (1995) for additional discussion.

Papke (1995) uses a spline method to estimate models with the participation rate, *PRATE*, as the dependent variable. She finds a statistically significant positive relationship between *PRATE* and *MRATE*, with some evidence of a diminishing marginal effect. Here, we allow for a diminishing marginal effect of MRATE on PRATE by using a conditional mean of the form (4) with $G(\cdot)$ taken to be the logit function. We compare this directly with linear models where *PRATE* is the dependent variable.

Table I presents summary statistics for the sample of 401(k) plans from the 1987 plan year. Statistics are presented separately for the 80% of the plans with match rates less than or equal to one. Match rates well above one likely indicate end-of-plan year employer contributions made to avoid IRS disqualification; see Papke (1995) for further discussion. Initially, we focus on the subsample with $MRATE \leq 1$.

Participation rates in 401(k) plans are high—averaging about 85% in our sample. Over 40% of the plans (42·73) have a participation proportion of exactly unity—all eligible employees have an active account. This characteristic of the data would make a log-odds approach especially awkward because an adjustment would have to be made to 40% of the observations.

The plan match rate averages about 41 cents on the dollar. Other explanatory variables include total firm employment (*EMP*) which averages 4,622 across the plans. The plans average 12 years in age (*AGE*). *SOLE* is a binary indicator for whether the 401(k) plan is the only pension plan offered by the employer. Sole plans comprise about 37% of the sample.

We begin with the linear model

$$E(PRATE \mid \mathbf{x}) = \beta_1 + \beta_2 MRATE + \beta_3 \log(EMP) + \beta_4 \log(EMP)^2 + \beta_5 AGE + \beta_6 AGE^2 + \beta_7 SOLE \tag{22}$$

which we estimate by ordinary least squares (OLS), initially using the subsample for which $MRATE \leq 1$. The results are given in the first column of Table II. Because of the anticipated

Table I. Summary statistics

Variable	Mean	Standard deviation	Minimum	Maximum
Full sample Number of observations = 4734				
PRATE	0·869	0·167	0·023	1
MRATE	0·746	0·844	0·011	5
EMPLOYMENT	4621·01	16299·64	53	443040
AGE	13·14	9·63	4	76
SOLE	0·415	0·493	0	1
Restricted sample (*MRATE* ≤ 1) Number of observations = 3874				
PRATE	0·848	0·170	0·023	1
MRATE	0·408	0·228	0·011	1
EMPLOYMENT	4621·91	17037·11	53	443040
AGE	12·24	8·91	4	76
SOLE	0·373	0·484	0	1

heteroscedasticity in this equation, the heteroscedasticity-robust standard errors are reported in brackets below the usual OLS standard errors.

All variables are highly statistically significant except for the sole plan indicator. Interestingly, there is very little difference between the usual OLS standard errors and the heteroscedasticity-robust ones. The key variable *MRATE* has a t-statistic well over 10. Its coefficient of 0·156 implies that if the match rate increases by 10 cents on the dollar, the participation rate would increase on average by almost 1·6 percentage points. This is not a small effect considering that the average participation rate is about 85% in the subsample. The linear model implies a constant marginal effect throughout the range of *MRATE* that cannot literally be true.

That the linear model does not fit as well as it should can be seen by computing Ramsey's (1969) RESET (and its heteroscedasticity-robust version). Let $\hat{u}_i$ be the OLS residuals and let $\hat{y}_i$ be the OLS fitted values. Then, the LM version of RESET is obtained as NR^2 from the regression

$$\hat{u}_i \text{ on } \mathbf{x}_i, \hat{y}_i^2, \hat{y}_i^3 \qquad i = 1, 2, \ldots, N$$

Under the null that equation (22) is true, $NR^2 \overset{a}{\sim} \chi_2^2$ (homoscedasticity is also maintained). The heteroscedasticity-robust version is obtained as $N - \text{SSR}$ from regression (20) given the proper definitions: let $\tilde{u}_i \equiv \hat{u}_i$ and let $\tilde{\mathbf{r}}_i$ be the 1×2 residuals from the regression of $(\hat{y}_i^2, \hat{y}_i^3)$ on $\mathbf{x}_i$; see Wooldridge (1991a) for more details. Using either non-robust RESET or its robust form, equation (22) is strongly rejected (the 1% critical value for a χ_2^2 is 9·21). Because RESET is a test of functional form, we conclude that equation (22) misses some potentially important non-linearities. (As usual, there is a potential difference between a statistical rejection of a model and the economic importance of any misspecification.)

We next use the logit QMLE analysed in Section 2 to estimate the non-linear model

$$E(PRATE \mid \mathbf{x}) = G(\beta_1 + \beta_2 MRATE + \beta_3 \log(EMP) + \beta_4 \log(EMP)^2 + \beta_5 AGE + \beta_6 AGE^2 + \beta_7 SOLE) \tag{23}$$

Table II. Results for the restricted sample

Variable	(1) OLS	(2) QMLE	(3) OLS	(4) QMLE
MRATE	0·156	1·390	0·239	1·218
	(0·012)	(0·100)	(0·042)	(0·342)
	[0·011]	[0·108]	[0·046]	[0·378]
$MRATE^2$	—	—	−0·087	0·196
			(0·043)	(0·373)
			[0·044]	[0·425]
log(*EMP*)	−0·112	−1·002	−0·112	−1·002
	(0·014)	(0·111)	(0·014)	(0·111)
	[0·013]	[0·110]	[0·013]	[0·110]
$\log(EMP)^2$	0·0057	0·052	0·0057	0·0522
	(0·0009)	(0·0071)	(0·0009)	(0·0071)
	[0·0009]	[0·0071]	[0·0009]	[0·0071]
AGE	0·0060	0·0501	0·0059	0·0503
	(0·0010)	(0·0087)	(0·0010)	(0·0087)
	[0·0009]	[0·0089]	[0·0009]	[0·0088]
AGE^2	−0·00007	−0·00052	−0·00007	−0·00052
	(0·00002)	(0·00021)	(0·00002)	(0·00021)
	[0·00002]	[0·00021]	[0·00002]	[0·00021]
SOLE	−0·0001	0·0080	0·0008	0·0061
	(0·0058)	(0·0468)	(0·0058)	(0·0470)
	[0·0060]	[0·0502]	[0·0060]	[0·0504]
ONE	1·213	5·058	1·198	5·085
	(0·051)	(0·427)	(0·052)	(0·430)
	[0·048]	[0·4211]	[0·049]	[0·423]
Observations:	3784	3784	3784	3784
SSR:	93·67	92·70	93·56	92·69
SER:	0·157	0·438	0·157	0·438
R-squared:	0·143	0·152	0·144	0·152
RESET:	39·55	0·606	35·06	0·732
	(0·000)	(0·738)	(0·000)	(0·693)
Robust RESET:	45·36	0·782	40·08	0·836
	(0·000)	(0·676)	(0·000)	(0·658)

Notes: The quantities in (·) below estimates are the OLS standard errors or, for QMLE, the GLM standard errors; the quantities in [·] are the standard errors robust to variance misspecification. SSR is the sum of squared residuals and SER is the standard error of the regression; for QMLE, the SER is defined in terms of the weighted residuals. The values in parentheses below the RESET statistics are *p*-values; these are obtained from a chi-square distribution with two degrees-of-freedom.

where $G(\cdot)$ is the logistic function. (The GAUSS® code used for the estimation and testing is available on request from the authors.) The partial effect of *MRATE* on $E(PRATE|\mathbf{x})$ is $\partial E(PRATE|\mathbf{x})/\partial MRATE$, or, for specification (23), $g(\mathbf{x}\boldsymbol{\beta})\beta_2$, where $g(\mathbf{z}) \equiv dG(z)/dz = \exp(z)/[1+\exp(z)]^2$. Because $g(z) \to 0$ as $z \to \infty$, the marginal effect falls to zero as *MRATE* becomes large, holding other variables fixed.

Column (2) of Table II contains the results of estimating equation (23). The variable *MRATE* is highly statistically significant and, with the exception of *SOLE* (which is still not significant), the directions of effects of all other variables are the same as in the linear model. Unlike the linear model, the RESET statistic reveals no misspecification in equation (23); the *p*-value for the robust statistic is 0·676, and it is even larger for the non-robust statistic. Based on this

RESET analog, equation (23) appears to capture the non-linear relationship between *PRATE* and the explanatory variables for $MRATE \leq 1$.

There is other evidence that equation (23) fits better than (22). Table II also contains an *r*-squared for each model, which in either case is defined as 1 – SSR/SST, where SST is the total sun of squares of the y_i. The SSRs, reported in Table II, are based on the *unweighted* residuals, $\hat{u}_i \equiv y_i - \hat{y}_i$ for OLS and QMLE. Thus, the *r*-squareds are comparable across *any* model for $E(PRATE \mid \mathbf{x})$ and for any estimation methods. From Table II we see that the *r*-squared from the logit model is about 6% higher than the *r*-squared for the linear model. Also, while OLS chooses $\hat{\beta}$ to maximize the *r*-squared over all linear functions of **x**, the logit QMLE does *not* maximize *r*-squared given the logit functional form; yet the logit model has a higher *r*-squared than the linear model. Since we are only modelling the conditional expectation, with other features of the conditional distribution left unspecified, the *r*-squared is the most appropriate goodness-of-fit measure.

Before directly comparing estimates of the response functions and the marginal effects, some other comments are worth making about Table II. First, each method comes with an SER (standard error of the regression). These SERs are the estimates of σ for the different models, and thus are not directly comparable. For OLS, $\hat{\sigma}^2$ is based on the unweighted OLS residuals, while for QMLE, $\hat{\sigma}^2$ is based on the weighted residuals; see equation (11). Because $\hat{\sigma} = 0{\cdot}438$ for the QMLE, this implies that the usual logit standard errors obtained from the inverse of the Hessian, $\hat{\mathbf{A}}^{-1}$, are over twice as large as the GLM standard errors that are obtained as the squared roots of the diagonal elements of $\hat{\sigma}^2\hat{\mathbf{A}}^{-1}$. The latter (smaller) standard errors are the appropriate ones under the GLM assumption (6) because they do not assume that $\sigma = 1$. *MRATE* is underdispersed ($\sigma^2 < 1$) relative to the Bernoulli variance ($\sigma^2 = 1$).

We now turn to a direct comparison of the linear and logistic models. To compare the estimated response functions and marginal effects, we need to choose values for *MRATE*, *EMP*, *AGE*, and *SOLE*. Because most 401(k) plans are accompanied by other pension plans, we set $SOLE = 0$. We also set *AGE* at roughly its sample average, $AGE = 13$. To gauge the differences across firms of different sizes we choose three firm sizes: small ($EMP = 200$), average ($EMP = 4620$), and large ($EMP = 100{,}000$) The estimated relationships between $E(PRATE \mid \mathbf{x})$ and *MRATE* for the three different firm sizes are graphed in Figure 1. Interestingly, for a small firm the linear and logistic predictions are most different at high match rates; for the average sized firm, the difference is largest at low match rates; and for a large firm the largest difference is at a match rate between 0·5 and 0·75.

As is seen from Table II, the marginal effect of *MRATE* on $E(PRATE \mid \mathbf{x})$ for the linear model is 0·156 for any value of **x**. For the logistic model, we set $SOLE = 0$, $AGE = 13$, and $EMP = 4{,}620$, and compute the estimated partial effect at three different match rates: $MRATE1 = 0$, $MRATE = 0{\cdot}50$, and $MRATE = 1{\cdot}0$. The estimated derivatives are 0·288, 0·197, and 0·118, respectively, which illustrates the diminishing marginal effect as *MRATE* increases. Perhaps not surprisingly, the marginal effect estimated from the linear model is bracketed by the low and high estimates from the non-linear model. The differences in the estimated marginal effects are not trivial; for example, the non-linear model predicts an increase in participation of approximately 2·9 percentage points in moving from a zero match rate to $MRATE = 0{\cdot}10$, rather than the 1·6 percentage point increase obtained from the linear model. Similarly, at high match rates the marginal effect from increasing the match rate is estimated to be lower in the non-linear model.

One way to try to salvage the linear model is to use a more flexible functional form in the match rate. A popular functional form that allows a diminishing marginal effect is a quadratic. Column (3) contains estimates of the linear model that includes a quadratic in *MRATE*. The

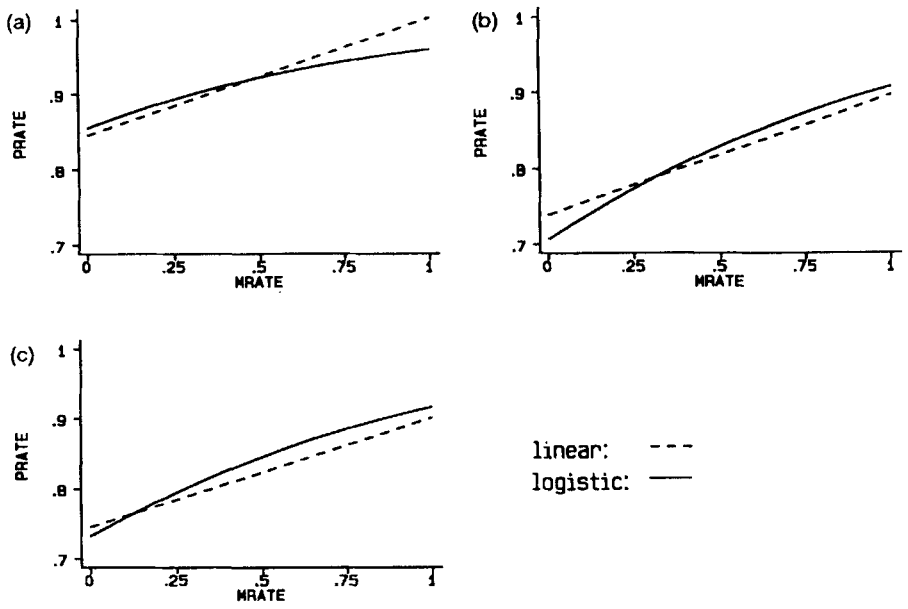

Figure 1. *PRATE* versus *MRATE* for various firm sizes: (a) EMP = 200; (b) EMP = 4620; (c) EMP = 100,000

squared term is marginally significant (robust t-statistic $\approx -1{\cdot}98$), and this does give a diminishing marginal effect. But even with this additional regressor the model in column (3) does not fit as well as the logistic model without the quadratic term (the r-squared for the linear model with the quadratic term is only 0·144). Further, the rejection of the model by RESET is almost as strong as it was without the quadratic. Thus, we conclude that simply adding $MRATE^2$ to equation (22) is not sufficient. (The spline approach used by Papke (1995) is more effective in capturing a diminishing effect in this application, but the coefficients are more difficult to interpret.)

When $MRATE^2$ is added to equation (23) it turns out to be insignificant. Thus, the logistic functional form, with the term linear in *MRATE*, appears to be enough to capture the diminishing effect, at least for $MRATE \leq 1$. This is a useful lesson: a significant quadratic term in a linear model might be indicating that an entirely different, more parsimonious, functional form can provide a better fit. Model (23) is clearly the preferred specification thus far.

As another test of model (23), we interact $\log(EMP)$ with each of *MRATE*, *AGE*, AGE^2, and *SOLE* and test for exclusion of these four interactions using the LM and QLR tests discussed in Section 3. This is similar in spirit to a Chow test where the sample is split based on firm size, but here we do not need to make an arbitrary choice about where to split the sample. The LM statistic is 16·52, the robust LM statistic is 14·41, and the QLR statistic, computed from equation (19), is 15·78 ($\ddot{\mathcal{L}} = -1547{\cdot}33$, $\hat{\mathcal{L}} = -1548{\cdot}84$, and $\ddot{\sigma}^2 = 0{\cdot}1914$). The associated p-value for the robust LM statistic is 0·006, which rejects equation (23) at the 1% significance level. Thus, equation (23) apparently misses some non-linearities, although the significance level is not very small given the large sample size (compare the p-value for RESET in the linear model).

From a practical perspective, the story about the relationship between expected *PRATE* and *MRATE* does not change: the t-statistic on the term $\log(EMP)\cdot MRATE$ is only $-1{\cdot}27$ (the robust t-statistic is $-1{\cdot}13$). In fact, when $\log(EMP)\cdot MRATE$ is dropped from the more general model, the coefficient on *MRATE* becomes 1·396, which is a trivial change from 1·390, the

estimate from equation (23). The most significant interaction term is $\log(EMP)\cdot SOLE$, with a t-statistic of -3.48 (robust t-statistic $= -3{\cdot}47$). We report only equation (23) because of its simplicity and because it captures the economically important relationship between *PRATE* and *MRATE*. The full set of results is available on request from the authors.

The basic story does not change when we estimate the models over the entire sample. One notable difference is that a quadratic term in *MRATE* is now significant in equation (23), reflecting a faster diminishing effect at high match rates. Table III presents the same models as Table II, now estimated over the full sample. First consider the models without $MRATE^2$. The discrepancy in r-squareds between equations (23) and (22) is even greater than before, but RESET now rejects both equations, although the logistic model is rejected less strongly. In columns (3) and (4) we put $MRATE^2$ into each equation. Model (22) is still soundly rejected, whereas (23) with $MRATE^2$ passes the RESET test with a p-value above 0·50. For the full sample, it seems that a quadratic in *MRATE*—or some other way to capture additional non-linearities—is needed to provide a reasonable fit.

Table III. Results for the full sample

Variable	(1) OLS	(2) QMLE	(3) OLS	(4) QMLE
MRATE	0·034	0·542	0·143	1·665
	(0·003)	(0·045)	(0·008)	(0·089)
	[0·003]	[0·079]	[0·008]	[0·104]
$MRATE^2$	—	—	−0·029	−0·332
			(0·002)	(0·021)
			[0·002]	[0·026]
$\log(EMP)$	−0·101	−1·038	−0·099	−1·030
	(0·012)	(0·121)	(0·012)	(0·112)
	[0·012]	[0·110]	[0·012]	[0·110]
$\log(EMP)^2$	0·0051	0·0540	0·0050	0·0536
	(0·0008)	(0·0078)	(0·0008)	(0·0072)
	[0·0008]	[0·0071]	[0·0008]	[0·0071]
AGE	0·0064	0·0621	0·0056	0·0548
	(0·0008)	(0·0089)	(0·0008)	(0·0082)
	[0·0007]	[0·0078]	[0·0007]	[0·0077]
AGE^2	−0·00008	−0·00071	−0·00007	−0·00063
	(0·00002)	(0·00021)	(0·00002)	(0·00019)
	[0·00002]	[0·00018]	[0·00001]	[0·00018]
SOLE	0·0140	0·1190	0·0066	0·0642
	(0·0050)	(0·0510)	(0·0049)	(0·0471)
	[0·0052]	[0·0503]	[0·0051]	[0·0498]
ONE	1·213	5·429	1·170	5·105
	(0·045)	(0·467)	(0·044)	(0·431)
	[0·044]	[0·422]	[0·042]	[0·416]
Observations:	4734	4734	4734	4734
SSR:	120·70	109·51	107·76	105·73
SER:	0·154	0·502	0·151	0·461
R-squared:	0·144	0·168	0·182	0·197
RESET:	85·22	50·56	83·80	1·370
	0·000)	(0·000)	(0·000)	(0·504)
Robust RESET:	69·15	9·666	98·51	1·275
	(0·000)	(0·008)	(0·000)	(0·529)

Note: See Table II.

Putting $MRATE^2$ into equation (23) has the usual drawback for quadratics: it implies an eventual negative marginal effect. In this case, the marginal effect becomes negative at a match rate of about 2·51. This is a high value for *MRATE*, but there are some match rates this large in the full sample.

5. CONCLUSION

The functional forms offered in this paper are viable alternatives to linear models that use either y or the log-odds ratio of y as the dependent variable. No special data adjustments are needed for the extreme values of zero and one, and the conditional expectation of y given the explanatory variables is estimated directly. The quasi-likelihood method we propose is fully robust and relatively efficient under the GLM assumption (6). The empirical application to 401(k) plan participation rates illustrates the usefulness of these methods: while a linear model to explain the fraction of participants is strongly rejected, the logistic conditional mean specification is not.

Methods for fractional dependent variables have many applications in economics. For example, Hausman and Leonard (1994) have recently applied the methods suggested here to estimate a model for Nielsen ratings for telecasts of NBA basketball games.

ACKNOWLEDGEMENTS

We are grateful to John Mullahy and two anonymous referees for helpful comments. The second author would like to thank the Alfred P. Sloan Foundation for financial support.

REFERENCES

Davidson, R. and J. G. MacKinnon(1984), 'Convenient specification tests for logit and probit models', *Journal of Econometrics*, **24**, 241–262.

Duan, N. (1983), 'Smearing estimate: a nonparametric retransformation method', *Journal of the American Statistical Association*, **78**, 605–610.

Engle, R. F. (1984), 'Wald, likelihood ratio, and Lagrange multiplier statistics in econometrics', in Z. Griliches and M. D. Intriligator (eds), *Handbook of Econometrics*, Volume 2, 776–828, North-Holland, Amsterdam.

Gourieroux, C., A. Monfort and A. Trognon (1984), 'Pseudo-maximum likelihood methods: theory', *Econometrica*, **52**, 681–700.

Gurmu, S. and P. K. Trivedi (1993), 'Variable augmentation specification tests in the exponential family', *Econometric Theory*, **9**, 94–113.

Hausman, J. A. and G. K. Leonard (1994), 'Superstars in the NBA: economic value and policy', MIT Department of Economics Working Paper No. 95-2.

Maddala, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.

McCullagh, P. and J. A. Nelder (1989), *Generalized Linear Models*, 2nd edition, Chapman and Hall, New York.

Mullahy, J. (1990), 'Regression models and transformations for beta-distributed outcomes', mimeo, Trinity College Department of Economics.

Papke, L. E. (1995), 'Participation in and contributions to 401(k) pension plans: evidence from plan data,' *Journal of Human Resources*, **30**, 311–325.

Papke, L. E. and J. M. Wooldridge (1993), 'Econometric methods for fractional response variables with an application to 401(k) plan participation rates', National Bureau of Economic Research Technical Working Paper No. 147.

Ramsey, J. B. (1969), 'Tests for specification errors in classical linear least squares regression analysis', *Journal of the Royal Statistical Society*, Series B 31, 350–371.

Wooldridge, J. M. (1991a), 'On the application of robust, regression-based diagnostics to models of conditional means and conditional variances', *Journal of Econometrics*, **47**, 5–46.

Wooldridge, J. M. (1991b), 'Specification testing and quasi-maximum likelihood estimation', *Journal of Econometrics*, **48**, 29–55.