TECHNICAL WORKING PAPER SERIES

MUCH ADO ABOUT TWO: RECONSIDERING RETRANSFORMATION AND THE TWO-PART MODEL IN HEALTH ECONOMETRICS

John Mullahy

Technical Working Paper 228
http://www.nber.org/papers/T0228

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 1998

This research has been supported by Grant AA10393 from the NIH Office of Research on Women's Health and NIAAA to NBER, by NIAAA Grant AA10392 to the University of Minnesota, and by a grant from the David and Lucile Packard Foundation to NBER. The initial stimulus for this paper was provided by some enlightening remarks by Will Manning on heteroskedastic retransformations (see Manning, 1998, for a formal exposition). Thanks are owed to Will Manning, João Santos Silva, Jon Skinner, Frank Windmeijer, Jeff Wooldridge, and Gary Zarkin for insightful comments and exchanges. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

Much Ado About Two: Reconsidering
Retransformation and the Two-Part Model
in Health Econometrics
John Mullahy
NBER Technical Working Paper No. 228
March 1998
JEL Nos. I1, C2

ABSTRACT

In health economics applications involving outcomes (y) and covariates (**x**), it is often the case that the central inferential problems of interest involve E[y|**x**] and its associated partial effects or elasticities. Many such outcomes have two fundamental statistical properties: $y \geq 0$; and the outcome y=0 is observed with sufficient frequency that the zeros cannot be ignored econometrically. Common approaches to estimation in such instances include Tobit, selection, and two-part models. This paper (1) describes circumstances where the standard two-part model with homoskedastic retransformation will fail to provide consistent inferences about important policy parameters; and (2) demonstrates some alternative approaches that are likely to prove helpful in applications.

John Mullahy
Departments of Preventive Medicine and Economics
University of Wisconsin-Madison
Madison, WI 53706
and NBER
jmullahy@facstaff.wisc.edu

I. INTRODUCTION

A. *Prologue*

Many outcomes (y) studied empirically in health economics have two fundamental statistical properties: (a) $y \geq 0$; and (b) the outcome $y=0$ is observed sufficiently frequently that the zeros cannot be ignored econometrically. Such data structures are observed in health applications as diverse as health care utilization/expenditure, the use of unhealthy commodities like tobacco and alcohol, and physicians' time allocations to alternative uses. Given exogenous covariates $\mathbf{x}$, econometric applications in which such data structures are encountered have typically relied on one or more of the following (and, generally, competing) three well-known strategies.

The *two-part model* (2PM) assumes that $\Pr(y>0|\mathbf{x})$ is governed by a parametric binary probability model like logit or probit (part one), and that $E[\ln(y)|y>0,\mathbf{x}]$ is a linear function of $\mathbf{x}$, e.g. $E[\ln(y)|y>0,\mathbf{x}]=\mathbf{x}\beta$ (part two).[1] The *sample selection model* (SSM) assumes that there are two linear equations determining the observed outcome. The first equation is $z=\mathbf{x}\xi_1+\nu$, the second equation is $w=\mathbf{x}\xi_2+\upsilon$, where the error terms (ν,υ) are typically assumed to follow a bivariate normal distribution. In this model, the outcome $\ln(y)=w$ is observed only if $z>0$; regression

---

[1] In some instances this has been referred to as a hurdle model. In addition, other transformations, e.g. $E[\sqrt{y}|y>0,\mathbf{x}]=\mathbf{x}\beta$, have also been suggested, but (a) the log-transformation is by far the most commonly used in practice and (b) essentially the same issues as are discussed below arise whether $\ln(y)$, $\sqrt{y}$, or some other transformation is used. See Manning, 1998, for the particulars as they apply to the 2PM and Carroll and Ruppert, 1988, for a general discussion of transformations.

methods like Heckman's approach (Heckman, 1979) then estimate a Mills-ratio-corrected linear regression of ln(y) on **x** using only the subsample of observations for which z>0.[2] *Tobit* and related models assume that $y^*|\mathbf{x} \sim N(\mathbf{x}\omega, \tau^2)$ and that the observed y is given by $y=\max(0,y^*)$.

B. Inference with the Two-Part Model

While the choice among these or other competing estimation strategies is clearly a first-order analytical issue,[3] this paper tackles a set of somewhat more subtle issues in estimation and inference encountered with applications of two-part models. To wit: While part two of the 2PM has been demonstrated in many empirical settings to be a useful estimator of the parameters β, how these estimates are used is an altogether separate matter. In addressing such concerns, this paper has two main purposes.

The first is to suggest that reliable/consistent estimates of β -- while *necessary* elements of the 2PM framework for conducting inference about important policy parameters -- will generally not be *sufficient* for such purposes. The second is to

---

[2] See Manning et al., 1987a,b for discussion. Note that the modification for the "true zeros" case rather than the "missing observations" case is sometimes referred to as the adjusted Tobit model.

[3] Other strategies are, of course, available (some of these are discussed below). Indeed, depending on the particular objective of the analysis, relatively simple strategies like nonparametric regression (e.g. cell means), linear specification of $E[y|\mathbf{x}]$ with OLS regression, and other approaches may be perfectly acceptable. The intent here is not to assess the relative merits of such alternatives but rather to take interest in a two-part estimation strategy as given and proceed to assess some of the key implications of such an approach.

demonstrate some alternative approaches that are likely to prove useful in applications. In particular, since it will often be the case that many inferential problems of interest involve $E[y|\mathbf{x}]$ and its associated partial effects $\delta(\mathbf{x}) = [\delta_j(\mathbf{x})] = [\partial E[y|\mathbf{x}]/\partial x_j]$ and/or elasticities $\eta(\mathbf{x}) = [\eta_j(\mathbf{x})] = [\partial \ln(E[y|\mathbf{x}])/\partial \ln(x_j)]$, it is fundamentally important to recognize that the parameters β are but one feature of such expectations and related quantities.

That is, how one proceeds from inferences about properties of $E[\ln(y)|y>0,\mathbf{x}]$ (where the elements of β are the key to inference) to inferences about properties of $E[y|\mathbf{x}]$ entails at least two separate considerations. The first is removing the conditioning on $y>0$; the second is transforming back from $\ln(y)$-space to y-space. Both issues have been discussed extensively in the literature and both are involved in the following analysis, although the perspective here departs materially from that typically maintained in the literature.

C. Applications

Some prominent areas of potential applicability of the ideas discussed here are noted at this juncture.

Outcomes Research

Analysts working in the fields of outcomes research, disease management, etc., often utilize large samples of individual-level outcomes on various measures of health care utilization, expenditures, or outcomes ("claims data"). Such datasets typically contain information only on individuals for whom some positive amount of utilization or expenditure is observed over

some specific time period. As such, a common objective in such research is to draw inferences about the determinants of $E[y|y>0,\mathbf{x}]$, perhaps augmenting such inferences with information about $Pr(y>0|\mathbf{x})$ obtained from other data sources. As the main arguments of this paper will illustrate, using common methods like loglinear regression with retransformation must be approached with care if inferences about properties of $E[y|y>0,\mathbf{x}]$ drawn from analysis of claims datasets are to be reliable. The alternative approaches proposed here should be useful in many applications of interest to outcomes researchers.

Models of the Demand for Health Care

It is common practice in empirical health economics to model individuals' demands for health care in a two-part context: whether, over some time period, the individual obtains or uses any care at all; and, if so, how much care (e.g. how many physician visits) is obtained. The two components of this process may differ in their economic determinants as well as their policy relevance (e.g. Pohlmeier and Ulrich, 1995). Consider the example of childhood immunizations (Mullahy, 1997b): An analyst may be concerned both about whether a child has obtained any immunizations by age two, as well as the extent to which the child is on-schedule for immunizations by that age. Another example is screening: Issues may arise regarding both whether an individual has ever been screened for a particular disease and, if so, the frequency with which such screening occurs.

Substance Abuse

In many econometric studies of substance use/abuse (tobacco,

alcohol, illicit drugs), analysts have often examined phenomena that bifurcate naturally into two components: whether or not individuals consume the commodity, and how much of the commodity is consumed by users; or whether or not the use of the commodity influences labor market participation and, if so, whether or not the commodity's use affects hours worked or wages earned.[4] Lost in much of this discourse, however, is a consideration of what these two sets of estimates imply overall for key parameters like $E[y|\mathbf{x}]$ and its associated partial effects. In some cases, it may be the case that parameters other than $E[y|\mathbf{x}]$ are of interest (e.g. Manning et al., 1995), but in other applications the conditional mean is likely to be a prominent consideration (e.g. Mullahy, 1997a).

The following sections describe some fundamental properties of the 2PM model that -- while overlooked often in applications -- turn out to have critical implications for inference, and suggest some reformulations of the 2PM that provide for straightforward inference in the context of some nonlinear regression structures.

D. Plan for the Paper

The plan for the paper is as follows. Section II presents the statistical preliminaries and describes in detail the two-part model. Section III discusses issues involved in inference based on the 2PM about properties of $E[y|\mathbf{x}]$. Section IV suggests alternatives to the 2PM, discusses issues involved in their estimation, and proposes a set of specification tests. Section V

---

[4] See, e.g., French and Zarkin, 1995, and Manning et al., 1995 for applications involving alcohol use.

presents results of a simulation exercise. Section VI reports an empirical study of doctor visits based on the 1992 National Health Interview Survey. Section VII offers conclusions.

II. STATISTICAL PRELIMINARIES: E[y|x] AND THE TWO-PART MODEL

A. *Fundamental Statistical Issues*

It assumed that the analyst observes a random sample of N observations on $(y_i, \mathbf{x}_i)$, where $\mathbf{x}_i=(1,\mathbf{x}_{i1})$ is a k-vector of covariates. There are assumed to be N_+ observations for which $y_i>0$ and N_0 observations for which $y_i=0$, with $N=N_+ + N_0$. The index sets for observations i corresponding to these samples are denoted $S_+=\{i|y_i>0\}$ and $S_0=\{i|y_i=0\}$. Unless necessary for clarity, the "i" subscripts will be suppressed.

With $y\geq 0$, it is meaningful[5] to write a regression model for $y|\mathbf{x}$ using the decomposition[6]

$$E[y|\mathbf{x}] = Pr(y>0|\mathbf{x}) \times E[y|y>0,\mathbf{x}]. \qquad (1)$$

Given $\mathbf{x}$, the realizations of y are generated as

---

[5] An appropriate caveat to this statement is that it is valid so long as such expectations can reasonably be maintained to exist. This could at least in principle be a tenuous matter for highly-skewed distributions of outcomes like cost or utilization. In general, any distribution $f(y|\mathbf{x})$ falling off to zero more slowly than at a rate proportional to y^{-2} will have a mean that does not exist (see Cramer, 1946, for general discussion).

[6] Cragg, 1971, was probably the first prominent econometric study to consider the utility of such decompositions; see Mullahy, 1986, for additional discussion.

$$y = \Psi(\mathbf{x}) \bigstar u \qquad (2)$$

where u is a stochastic error term and where the operator "★" can denote either "×" or "+".[7] *If* $E[u|\mathbf{x}]=1-1\bigstar 0$, then $\Psi(\mathbf{x})$ implicitly defines $E[y|\mathbf{x}]$. However, if $E[u|\mathbf{x}]=h(\mathbf{x})$ then $E[y|\mathbf{x}]=\Psi(\mathbf{x})\bigstar h(\mathbf{x})$. Possible dependence of $E[u|\mathbf{x}]$ on $\mathbf{x}$ is usually "normalized away," but it is ultimately useful for purposes at hand to keep this distinction in mind. As a matter of notation, $\Psi(\mathbf{x})$ will be used to denote $E[y|\mathbf{x}]$ in what follows.

B. The Two-Part Model

Rather than working directly in a regression context like (2) with the level of y *per se*, the 2PM decomposes $E[y|\mathbf{x}]$ by specifying a parametric model for $Pr(y>0|\mathbf{x})=\pi(\mathbf{x};\alpha)$ (part one), and then taking the log-transformation

$$\ln(y) = \ln(\mu(\mathbf{x};\beta)) + \varepsilon, \quad y>0 \qquad (3)$$

to characterize part two. Specifying $\ln(\mu(\mathbf{x};\beta))=\mathbf{x}\beta$ and $E[\varepsilon|y>0,\mathbf{x}]=0$ implies

$$E[\ln(y)|y>0,\mathbf{x}] = \mathbf{x}\beta. \qquad (4)$$

Some basic probability algebra then permits one, in principle, to

---

[7] Whether the error is best modeled as additive or multiplicative in the case of nonnegative dependent variables is, to a considerable degree, immaterial. See Wooldridge, 1992, for additional discussion.

recover $E[y|\mathbf{x}]$ as

$$\begin{aligned} E[y|\mathbf{x}] &\equiv \Psi(\mathbf{x}) \\ &= Pr(y>0|\mathbf{x}) \times E[y|y>0,\mathbf{x}] \\ &= \pi(\mathbf{x}) \times \{ \mu(\mathbf{x}) \times E[\exp(\varepsilon)|y>0,\mathbf{x}] \} \\ &= \pi(\mathbf{x}) \times \{ \mu(\mathbf{x}) \times \rho(\mathbf{x}) \}, \end{aligned} \tag{5}$$

since $y=\mu(\mathbf{x})\exp(\varepsilon)$ for $y>0$ from (3).[8] For reasons to be discussed below, the notation $\rho(\mathbf{x})$ is used in lieu of the more familiar constant ϕ to emphasize that the error retransformation is, in general, a function of $\mathbf{x}$ as opposed to a scalar constant.

Assuming each component of $\Psi(\mathbf{x})$ in (5) has a parametric representation, then

$$\begin{aligned} E[y|\mathbf{x}] &\equiv \Psi(\mathbf{x};\theta) \\ &= \pi(\mathbf{x};\alpha) \times \mu(\mathbf{x};\beta) \times \rho(\mathbf{x};\gamma), \end{aligned} \tag{6}$$

where, for example, $Pr(y>0|\mathbf{x})=\pi(\mathbf{x})=\pi(\mathbf{x};\alpha)$ is typically given by a distribution function with linear index, $F(\mathbf{x}\alpha)$. This paper will concentrate on the logit specification

$$Pr(y>0|\mathbf{x}) = \frac{\exp(\mathbf{x}\alpha)}{1+\exp(\mathbf{x}\alpha)}, \tag{7}$$

---

[8] It should be emphasized that the analysis undertaken here encompasses "one-part" models where $Pr(y>0|\mathbf{x})=\pi(\mathbf{x})=1$ for all $\mathbf{x}$, e.g. an application in which y is a measure of cost in a clinical trial setting where cost is strictly positive for all subjects.

although probit and linear probability models are also obvious candidates that have been employed in empirical applications of the 2PM.[9] It is sometimes useful to refer to (7) as the *hurdle* component of the model and (3) as the *levels* component of the model.

Equation (5) underscores the fact that $E[y|y>0,\mathbf{x}] \neq \mu(\mathbf{x})$. Indeed, from Jensen's inequality on convex functions

$$E[\exp(\ln(y))|y>0,\mathbf{x}] = E[y|y>0,\mathbf{x}] > \exp(E[\ln(y)|y>0,\mathbf{x}]) = \mu(\mathbf{x}). \quad (8)$$

As such, it must be the case that the error retransformation function $\rho(\mathbf{x})$ exceeds one at any $\mathbf{x}$.[10]

C. Fundamental Properties of E[y/x]

Often ignored is that a far more primitive assumption about data on nonnegative outcomes than offered by either the 2PM or the SSM is simply that $E[y|\mathbf{x}]>0$. If all realizations of y are nonnegative, then $E[y|\mathbf{x}]$ must clearly be nonnegative. If $E[y|\mathbf{x}]=0$ then the analytical problem is not interesting. As such, $E[y|\mathbf{x}]>0$ is the only reasonable assumption in such instances. In the microeconometric settings familiar to health economists, focusing on $E[y|\mathbf{x}]$ is natural since inference about the partial effects $\delta(\mathbf{x})$ is typically one of the major

---

[9] Much of the original Rand work used probit models for describing part one of the 2PM. Eichner et al., 1997, is a recent example of the use of linear probability models to characterize part one.

[10] Again using Jensen's inequality, this is seen to be true since $E[\exp(\varepsilon)|y>0,\mathbf{x}]>\exp(E[\varepsilon|y>0,\mathbf{x}])=\exp(0)=1$.

considerations in such applications. For example, obtaining reliable estimates of $E[y|\mathbf{x}]$ is obviously a central concern of analysts wishing to undertake "$\Delta\mathbf{x}$" policy analysis.

Yet in many applications of the 2PM and related methods, consideration of issues attending estimation of the quantity $E[y|\mathbf{x}]$ and its associated partial effects has not been as prominent as (this paper would contend) appropriate. Analysts are sometimes satisfied conducting inference directly on the parameters α and β without regard to how such inferences relate to inferences about quantities like $\delta(\mathbf{x})$. While methods like the homoskedastic version of Duan's smearing estimator (Duan, 1983) have been used widely (and, perhaps, almost too automatically) with the objective of estimating $E[y|\mathbf{x}]$ and its corresponding partial effects, this paper suggests that a more direct consideration of and focus on $E[y|\mathbf{x}]$ may be desirable and useful.

As such, on the basis of primitive assumptions about the structure of $E[y|\mathbf{x}]$ -- i.e. $E[y|\mathbf{x}]>0$ -- one might consider as an alternative to the 2PM and SSM approaches modeling *directly* the regression $E[y|\mathbf{x}]$ without recourse to transformation and retransformation. As suggested below in section IV, this might be accomplished in a single equation context using *all* sample observations ($y=0$ *and* $y>0$), or may be undertaken using two-step methods analogous to those used in most applications of the 2PM.

III. INFERENCE WITH THE TWO-PART MODEL

A. Identification and Estimation

Should the data be up to the task of identifying the parameters α and β, then the 2PM is -- in one sense -- identified. Yet, in the absence of further assumptions (e.g.

lognormality of $f(y|y>0,\mathbf{x})$), it is important to note that the standard specification of the 2PM ((3) and (7)) does *not* generally permit one to recover $E[y|y>0,\mathbf{x}]$ and, therefore, $E[y|\mathbf{x}]$ since identification of $E[\ln(y)|y>0,\mathbf{x}]$ (as given in (4)) is *not* sufficient to identify $E[y|y>0,\mathbf{x}]$. As such, the 2PM thus formulated does not have an interpretation as a natural parametric regression model

$$y = \Psi(\mathbf{x};\theta) \bigstar u, \quad \forall y, \tag{9}$$

where $\Psi(\mathbf{x};\theta)=E[y|\mathbf{x}]$ and $E[u|\mathbf{x}]=1-1\bigstar 0$.

Estimation of the components $\pi(\mathbf{x};\alpha)$ and $\mu(\mathbf{x};\beta)$ is generally not problematic given the assumptions commonly underlying the 2PM and related estimation approaches. In particular, it should be emphasized at the outset that this paper accepts one of the basic econometric underpinnings of the two-part model, i.e. that for functions of interest $\psi(.)$, the parameters characterizing $E[\psi(y)|y>0,\mathbf{x}]$ can be estimated consistently using only the S_+ subsample. This has been a matter of some contention in the literature,[11] but it will be ignored for the remainder of the paper. As such, it is assumed that regression of $\ln(y)$ on $\mathbf{x}$ using observations in S_+ generates consistent estimates $\hat{\beta}$ of β and $\mu(\mathbf{x};\hat{\beta})$ of $\mu(\mathbf{x};\beta)=\exp(\mathbf{x}\beta)$.

In applications of the 2PM, the main impediment to consistent estimation of $E[y|\mathbf{x}]=\Psi(\mathbf{x};\theta)$ -- and, therefore, to

[11] Witness the "cake" debates of the 1980s (Hay et al., 1987; Hay and Olsen, 1984; Manning et al., 1987a; Newhouse et al., 1980; Newhouse et al., 1987; Welch et al., 1987).

consistent estimation of the partial effects $\delta(\mathbf{x})$ that are assumed to be of central concern -- concerns the error retransformation component $\rho(\mathbf{x})=\rho(\mathbf{x};\gamma)$. Many applications ignore this consideration altogether; others have used methods like Duan's smearing estimator (described below), but have typically failed to account for the possible dependence of ρ on $\mathbf{x}$.[12]

Two Examples

First suppose $v\sim U[-.5,.5]$ and $\mathbf{x}=[1,x_1]$ with x_1 a scalar and v and x_1 statistically independent. Let $\varepsilon=g(x_1)v$ for $y>0$ with $g(.)>0$. As such,

$$\begin{aligned} E[\varepsilon|y>0,\mathbf{x}] &= E[g(x_1)\times v|y>0,\mathbf{x}] \\ &= g(x_1) \times E[v|y>0,\mathbf{x}] \\ &= g(x_1) \times E[v|y>0] \\ &= 0. \end{aligned} \tag{10}$$

Moreover,

$$\begin{aligned} E[\exp(\varepsilon)|y>0,\mathbf{x}] &= \int_{-.5}^{.5} \exp(g(x_1)v)dv \\ &= \frac{\exp(.5g(x_1)) - \exp(-.5g(x_1))}{g(x_1)}. \end{aligned} \tag{11}$$

For the second example, suppose $v\sim N(0,1)$ and $\mathbf{x}=[1,x_1]$ with x_1 a scalar and v and x_1 statistically independent. Let

[12] See Manning, 1998, for discussion. Manning notes that the HIE studies typically circumvented such biases by using plan-specific smearing retransformations rather than a single retransformation based on a full-sample smearing estimate.

$\varepsilon=\sqrt{2(\gamma_0 + x_1\gamma_1)}\times v$ with $\gamma_0+x_1\gamma_1>0$ so that $\varepsilon|y>0,\mathbf{x}\sim N(0,2(\gamma_0+x_1\gamma_1))$ implying $E[\varepsilon|y>0,\mathbf{x}]=0$. $\exp(\varepsilon)|y>0,\mathbf{x}$ is, therefore, lognormally distributed with

$$E[\exp(\varepsilon)|y>0,\mathbf{x}] = \exp(.5\sigma^2(\mathbf{x})) = \exp(\gamma_0 + x_1\gamma_1), \quad (12)$$

which is an ECM formulation for $E[\exp(\varepsilon)|y>0,\mathbf{x}]$.

B. Retransformation and Smearing

In typical applications of the 2PM, analysts interested in recovering estimates of $E[y|\mathbf{x}]$ have generally resorted to one of two alternative methods. The first is to assume that $y|y>0,\mathbf{x}$ is lognormally distributed with constant variance parameter σ^2. Given a consistent estimate of β obtained from linear regression of $\ln(y)$ on $\mathbf{x}$, the lognormality assumption then enables consistent estimation of $E[y|y>0,\mathbf{x}]=\exp(\mathbf{x}\beta+.5\sigma^2)$ via the aforementioned estimate of β and an estimate of σ^2 obtained in a method of moments fashion based on estimated residuals from the linear regression.

Duan, 1983, suggested that the robustness of retransformations based on the lognormality assumption could hinge critically on whether the data on $y>0$ were, indeed, conditionally lognormally distributed. As a robust alternative, Duan suggested the *smearing estimator*. The idea is that instead of using the estimated residuals to obtain an estimate of σ^2 as in the lognormal case, the estimated residuals $\hat{\varepsilon}=\ln(y)-\mathbf{x}\hat{\beta}$ can be used to provide a consistent estimate of a homoskedastic

distribution-robust retransformation factor $\hat{\phi}=N_{+}^{-1}\sum_{i \in S_{+}} \exp(\hat{\varepsilon}_i)$. Since $\ln(y)=\mathbf{x}\beta+\varepsilon$ is the maintained model for $y>0$, $E[y|y>0,\mathbf{x}]$ is then estimated as $\hat{\phi}\times\exp(\mathbf{x}\hat{\beta})$ and an estimate of $E[y|\mathbf{x}]$ follows given an estimate of $Pr(y>0|\mathbf{x})$ from, e.g., the logit model.

To appreciate why this version of the smearing retransformation can be problematic, however, it is useful to work through the algebra underlying the retransformation. Consistent estimation of the maintained model $\ln(y)=\mathbf{x}\beta+\varepsilon$, $y>0$, hinges most fundamentally on the orthogonality condition $E[\varepsilon|y>0,\mathbf{x}]=0$. No other orthogonality condition need be specified in order to recover a consistent estimate of β by least squares, and analysts using linear regression models do not typically specify restrictions beyond this. In particular, it is *not* necessary to maintain that ε and $\mathbf{x}$ are statistically independent (conditional on $y>0$) for consistent estimates to obtain.

As such, it is important to emphasize that the (perhaps implicitly) maintained assumption $E[\varepsilon|y>0,\mathbf{x}]=0$ does *not* imply for functions $\psi(.)$ that $E[\psi(\varepsilon)|y>0,\mathbf{x}]$ is a constant not depending on $\mathbf{x}$; e.g. $\varepsilon|y>0,\mathbf{x}$ could have a heteroskedastic distribution with variance $\tau(\mathbf{x})$. Retransforming to recover $E[y|y>0,\mathbf{x}]$ would then use

$$\begin{aligned} E[y|y>0,\mathbf{x}] &= E[\exp(\mathbf{x}\beta)\exp(\varepsilon)|y>0,\mathbf{x}] \\ &= \exp(\mathbf{x}\beta) \times E[\exp(\varepsilon)|y>0,\mathbf{x}] \\ &= \exp(\mathbf{x}\beta) \times \rho(\mathbf{x}), \end{aligned} \tag{13}$$

where $\rho(\mathbf{x})$ is the retransformation factor under the assumption $E[\varepsilon|y>0,\mathbf{x}]=0$.

In general, $\rho(\mathbf{x})$ may depend on $\mathbf{x}$ in a nontrivial manner. As

such estimates of $E[y|y>0,\mathbf{x}]$ and, consequently, of $E[y|\mathbf{x}]$ that fail to recognize the possible dependence of the retransformation factor on $\mathbf{x}$ and that use instead the standard (homoskedastic) smearing retransformation factor are likely (as shown below) to yield misleading (i.e. biased) estimates of key parameters of interest like $\delta(\mathbf{x})$ and $\eta(\mathbf{x})$. This occurs because the effects of $\mathbf{x}$ on $E[y|\mathbf{x}]$ that work through $\rho(\mathbf{x})$ have been ignored.

As a matter of historical context, it should be noted that much of the seminal work on retransformations and the two-part model that grew from the Rand Health Insurance Experiment (HIE) *did* -- appropriately -- emphasize that ϕ (or ρ, in present notation) is in general a function of $\mathbf{x}$ (Duan et al., 1983). One can only speculate as to why most applications have subsequently sidestepped this consideration, but the fact that the Rand analysts found that in some (albeit not all; see below) of their applications it was *empirically* "cumbersome and noisy" (Duan et al., 1983, p. 120) to account for the dependence of ϕ on $\mathbf{x}$ may have led some analysts to surmise thereafter that this issue was not important *conceptually*.

C. The Structure of $\rho(\mathbf{x})$

Parametric Structures

What might be a reasonable parametric model for $\rho(\mathbf{x})$? Since $\exp(\varepsilon)$ is necessarily positive, it might not be unreasonable to expect an *exponential conditional mean* (ECM) specification $\rho(\mathbf{x})=\exp(\mathbf{x}\gamma)$. As such, a multivariate version of the smearing estimate might be obtained by nonlinear regression of $\exp(\hat{\varepsilon})$ on $\exp(\mathbf{x}\gamma)$. At least in principle, determining whether the elements γ_1 of γ corresponding to $\mathbf{x}_1$ (the nonconstant elements of $\mathbf{x}$) are

nonzero would provide an indication of whether the dependence of the retransformation on $\mathbf{x}$ was important, or whether the standard homoskedastic Duan smearing method would be acceptable. Working out the asymptotics of such a problem is beyond this paper's scope, but it should be noted that if $\gamma_1=0$, then this amounts (albeit circuitously) to the standard homoskedastic smearing method if indeed $\rho(\mathbf{x})$ is log-linear in $\mathbf{x}$.

Nonparametric Approaches

If the elements of $\mathbf{x}$ are all discrete, then it may be possible to use a nonparametric approach like that used by the Rand HIE analysts in some of their applications. That is, for each of p distinct values of the vector $\mathbf{x}$ in S_+ (or, more precisely, the $p \leq k-1$ distinct values of the subvector of $\mathbf{x}$ that induce variation in $\rho(\mathbf{x})$) one would compute the smearing estimate in each subset S_{+j}, $j=1,\ldots,p$, as the analog of $\rho(\mathbf{x})$ for that particular subset of observations. So long as there are sufficiently many replicates of $\mathbf{x}$ for each of these subset estimators to have acceptable finite-sample behavior, this approach could be of practical use.

It might also appear that this would be an ideal case for nonparametric regression to provide informative estimates. While this might be so in some circumstances, the particular (and, presumably, most common) context in which the analyst is concerned with conducting *parametric* inference about parameters like $\delta(\mathbf{x})$ would not be helped much by the availability of nonparametric estimates of $\rho(\mathbf{x})$.

D. Biases Arising with the Homoskedastic 2PM

Since it ignores the dependence of $\rho(\mathbf{x})$ on $\mathbf{x}$, reference to the homoskedastic 2PM may result in biased inferences. This discussion considers how both marginal mean predictions -- i.e. predictions about $E_{\mathbf{x}}[\Psi(\mathbf{x})]$ -- as well as predictions about the marginal elasticities and partial effects $E_{\mathbf{x}}[\eta(\mathbf{x})]$ and $E_{\mathbf{x}}[\delta(\mathbf{x})]$ will be influenced by failure to consider the structure of $\rho(\mathbf{x})$.

To begin, it should be noted that $\hat{\beta}$ from the homoskedastic 2PM is a consistent estimator of β and, since exp(.) is continuous in its argument, $\exp(\mathbf{x}\hat{\beta})$ is consistent for $\exp(\mathbf{x}\beta)$ by Slutsky's theorem. Moreover, note that the homoskedastic 2PM smearing estimator $\hat{\phi}$ is a consistent estimator of $E_{\mathbf{x}}[\rho(\mathbf{x})]$ (since $\phi=E[\exp(\varepsilon)|y>0]=E_{\mathbf{x}}E[\exp(\varepsilon)|y>0,\mathbf{x}]=E_{\mathbf{x}}[\rho(\mathbf{x})]$). Finally, it is assumed that $\hat{\pi}(\mathbf{x})=\pi(\mathbf{x};\hat{\alpha})$ is a consistent estimator of $\pi(\mathbf{x};\alpha)$. As such, "bias" issues entail not so much whether the homoskedastic 2PM estimates consistently its structure as they are concerned with whether the structure being estimated is informative.

Biases Arising with Marginal Means

Taking (5) as the true specification of $E[y|\mathbf{x}]$, consider first predictions about the marginal mean $E_{\mathbf{x}}[\Psi(\mathbf{x})]$. Truth is

$$E[y|\mathbf{x}] \equiv \Psi(\mathbf{x}) = \pi(\mathbf{x}) \times \mu(\mathbf{x}) \times \rho(\mathbf{x}), \qquad (14)$$

but the homoskedastic 2PM with smearing estimates the quantity

$$E_{2PM}[y|\mathbf{x}] = \pi(\mathbf{x}) \times \mu(\mathbf{x}) \times E_{\mathbf{x}}[\rho(\mathbf{x})]. \quad (15)$$

For notational shorthand, let $A(\mathbf{x})=\pi(\mathbf{x})\times\mu(\mathbf{x})$. Then using the standard covariance formula $E[W\times V]=E[W]\times E[V]+Cov(W,V)$ for random variables W,V, it follows from (14) that the true marginal mean of y is

$$\begin{aligned} E_{\mathbf{x}}E[y|\mathbf{x}] &= E_{\mathbf{x}}[\Psi(\mathbf{x})] \quad (16) \\ &= E_{\mathbf{x}}[A(\mathbf{x})] \times E_{\mathbf{x}}[\rho(\mathbf{x})] + Cov_{\mathbf{x}}(A(\mathbf{x}),\rho(\mathbf{x})). \end{aligned}$$

The homoskedastic 2PM with smearing thus estimates the quantity

$$\begin{aligned} E_{\mathbf{x}}E_{2PM}[y|\mathbf{x}] &= E_{\mathbf{x}}[A(\mathbf{x}) \times E_{\mathbf{x}}[\rho(\mathbf{x})]] \quad (17) \\ &= E_{\mathbf{x}}[A(\mathbf{x})] \times E_{\mathbf{x}}[\rho(\mathbf{x})] \\ &= E_{\mathbf{x}}[\Psi(\mathbf{x})] - Cov_{\mathbf{x}}(A(\mathbf{x}),\rho(\mathbf{x})). \end{aligned}$$

As such, the *population* "mean prediction error" $E_{\mathbf{x}}E_{2PM}[y|\mathbf{x}]-E_{\mathbf{x}}[\Psi(\mathbf{x})]$ is governed by the quantity $Cov_{\mathbf{x}}(A(\mathbf{x}),\rho(\mathbf{x}))$, i.e. on how the "systematic" part of the model $A(\mathbf{x})=\pi(\mathbf{x})\times\mu(\mathbf{x})$ varies over $\mathbf{x}$ with the retransformation factor $\rho(\mathbf{x})$. Should it turn out that $\rho(\mathbf{x})$ and $A(\mathbf{x})$ covary positively over $\mathbf{x}$ -- an intuitively plausible scenario, but one by no means suggested by economic theory -- then the homoskedastic 2PM will tend to underestimate the true marginal mean $E_{\mathbf{x}}[\Psi(\mathbf{x})]$, i.e. the mean prediction error would be negative.

Since 2PM estimators are used often with the objective of predicting marginal means -- the example of health care expenditure analysis being perhaps most prominent -- nonzero mean

prediction error may be a paramount consideration. It should be noted in this context, however, that one can attain zero MPE for the homoskedastic 2PM with smearing simply by specifying $\mathbf{x}$ to contain only a constant term,[13] and that the issue of nonzero MPE only arises when there is variation in $\mathbf{x}$ over the sample (for only then can there be co-variation between $A(\mathbf{x})$ and $\rho(\mathbf{x})$).

Biases Arising with Elasticities and Partial Effects

Beyond consideration of mean predictions, the main focus in many policy analysis and forecasting settings will be on obtaining estimates of the effect of changes in elements of $\mathbf{x}$ on statistics describing the outcomes y, e.g. conditional means ($E[y|\mathbf{x}]$), conditional quantiles ($Q_\alpha[y|\mathbf{x}]$), etc. A central question that should be of concern to practitioners using the 2PM is: When it comes to elasticities and partial effects, precisely what is it that the 2PM based on the standard homoskedastic smearing estimator is actually estimating?

The analysis for elasticities is straightforward. The j-th conditional elasticity from the homoskedastic 2PM is given by

$$\frac{\partial \ln E_{2PM}[y|\mathbf{x}]}{\partial \ln x_j} = \frac{\partial \ln \pi(\mathbf{x})}{\partial \ln x_j} + \frac{\partial \ln \mu(\mathbf{x})}{\partial \ln x_j} \qquad (18)$$
$$= x_j[(1-\pi(\mathbf{x}))\alpha_j + \beta_j],$$

whereas the true j-th conditional elasticity is given by

[13] This is also true of the alternative estimators proposed in section IV.

$$\frac{\partial \ln \Psi(\mathbf{x})}{\partial \ln x_j} = \frac{\partial \ln \pi(\mathbf{x})}{\partial \ln x_j} + \frac{\partial \ln \mu(\mathbf{x})}{\partial \ln x_j} + \frac{\partial \ln \rho(\mathbf{x})}{\partial \ln x_j} \qquad (19)$$

$$= x_j [(1-\pi(\mathbf{x}))\alpha_j + \beta_j] + \frac{\partial \ln \rho(\mathbf{x})}{\partial \ln x_j}.$$

As such, the difference involves only the term $\frac{\partial \ln \rho(\mathbf{x})}{\partial \ln x_j}$, and its sign is not determined *a priori*. The overall or marginal elasticity will typically be obtained as $E_{\mathbf{x}}\left[\frac{\partial \ln E_{2PM}[y|\mathbf{x}]}{\partial \ln x_j}\right]$, so the difference between what the homoskedastic 2PM estimates and the true elasticity involves only the term $E_{\mathbf{x}}\left[\frac{\partial \ln \rho(\mathbf{x})}{\partial \ln x_j}\right]$, which again is not signed *a priori*.

Things are more complicated with the partial effects. Focusing on $E[y|\mathbf{x}]$, and on the basis of (6), the true partial effect $\delta_j(\mathbf{x})$ is given by

$$\frac{\partial E[y|\mathbf{x}]}{\partial x_j} \qquad (20)$$

$$= \pi(\mathbf{x})\rho(\mathbf{x})\exp(\mathbf{x}\beta)\beta_j + \pi(\mathbf{x})\exp(\mathbf{x}\beta)\frac{\partial \rho(\mathbf{x})}{\partial x_j} + \rho(\mathbf{x})\exp(\mathbf{x}\beta)\frac{\partial \pi(\mathbf{x})}{\partial x_j}$$

$$= \exp(\mathbf{x}\beta) \times \{\pi(\mathbf{x})\rho(\mathbf{x})\beta_j + \pi(\mathbf{x})\frac{\partial \rho(\mathbf{x})}{\partial x_j} + \rho(\mathbf{x})\frac{\partial \pi(\mathbf{x})}{\partial x_j}\},$$

whereas the partial effect that would be estimated from the homoskedastic 2PM is given by

$$\frac{\partial \hat{E}_{2PM}[y|\mathbf{x}]}{\partial x_j} = \hat{\phi} \times \{\hat{\pi}(\mathbf{x})\exp(\mathbf{x}\hat{\beta})\hat{\beta}_j + \exp(\mathbf{x}\hat{\beta})\frac{\partial \hat{\pi}(\mathbf{x})}{\partial x_j}\} \quad (21)$$

$$= \hat{\phi} \times \exp(\mathbf{x}\hat{\beta}) \times \{\hat{\pi}(\mathbf{x})\hat{\beta}_j + \frac{\partial \hat{\pi}(\mathbf{x})}{\partial x_j}\}.$$

What can be said about (20) vs. (21)? The results above imply that the 2PM estimator of $\delta_j(\mathbf{x})$ is consistent for the quantity

$$\mu(\mathbf{x}) \times \{\pi(\mathbf{x})\beta_j + \frac{\partial \pi(\mathbf{x})}{\partial x_j}\} \times E_{\mathbf{x}}[\rho(\mathbf{x})], \quad (22)$$

which is a well-defined quantity, albeit a quantity not equal to the true conditional partial effect (20). The difference between (20) and (22) can be written as

$$D_j(\mathbf{x}) = \quad (23)$$

$$\mu(\mathbf{x}) \times \{\pi(\mathbf{x})\beta_j + \frac{\partial \pi(\mathbf{x})}{\partial x_j}\} \times \{\rho(\mathbf{x}) - E_{\mathbf{x}}[\rho(\mathbf{x})]\} + \pi(\mathbf{x})\mu(\mathbf{x})\frac{\partial \rho(\mathbf{x})}{\partial x_j}.$$

The sign of the bias, $E_{\mathbf{x}}[D_j(\mathbf{x})]$, is ambiguous in general, but some of its features can be assessed as follows. For notational shorthand, let $B_j(\mathbf{x}) = \mu(\mathbf{x}) \times \{\pi(\mathbf{x})\beta_j + \frac{\partial \pi(\mathbf{x})}{\partial x_j}\}$ so that (23) can be rewritten as

$$D_j(\mathbf{x}) = B_j(\mathbf{x}) \times \{\rho(\mathbf{x}) - E_{\mathbf{x}}[\rho(\mathbf{x})]\} + \pi(\mathbf{x})\mu(\mathbf{x})\frac{\partial \rho(\mathbf{x})}{\partial x_j}. \quad (24)$$

Using the same logic as used to obtain (16), and simplifying, it follows that

$$E_{\mathbf{x}}[D_j(\mathbf{x})] \tag{25}$$

$$= \mathrm{Cov}_{\mathbf{x}}(B_j(\mathbf{x}), \rho(\mathbf{x})) + E_{\mathbf{x}}[\pi(\mathbf{x})\mu(\mathbf{x})\frac{\partial \rho(\mathbf{x})}{\partial x_j}].$$

$$= \beta_j \mathrm{Cov}_{\mathbf{x}}(A(\mathbf{x}), \rho(\mathbf{x})) + \mathrm{Cov}_{\mathbf{x}}(\mu(\mathbf{x})\frac{\partial \pi(\mathbf{x})}{\partial x_j}, \rho(\mathbf{x})) + E_{\mathbf{x}}[A(\mathbf{x})\frac{\partial \rho(\mathbf{x})}{\partial x_j}]$$

From the earlier discussion, observing the MPE allows a prediction about the magnitude and sign of $\mathrm{Cov}_{\mathbf{x}}(A(\mathbf{x}), \rho(\mathbf{x}))$, so the first term in the second line of (25) can be estimated by appeal to the homoskedastic 2PM estimates (since $\hat{\beta}_j$ is consistent for β_j here). Moreover, in many applications it will turn out empirically that $\mathrm{sgn}(\beta_j) = \mathrm{sgn}(\partial \pi(\mathbf{x})/\partial x_j)$ -- i.e. factors that positively (negatively) influence the magnitude of the outcome also positively (negatively) influence the probability of a positive outcome, all else equal -- although this is clearly not a restriction always given *a priori* by theory.[14] If so, then it may be reasonable to expect the second term in the second line of (25) to have the same sign as the first term. In conjunction with speculation about (or even estimation of) the sign of

---

[14] In some instances, however -- e.g. price effects for non-Giffen commodities -- it would be predicted that lower commodity prices ($p_y \in \mathbf{x}$) would tend both to reduce the probabilities of corner solutions, i.e. $\left[\frac{\partial \Pr(WTP_y < p_y | y = 0, \mathbf{x})}{\partial p_y}\right] > 0$, *and* increase quantities demanded when positive.

$\partial\rho(\mathbf{x})/\partial x_j$, speculation about the direction of bias might then be possible.[15]

But the bottom line is that there are no unambiguous general biases that can be determined from the conditional differences $D_j(\mathbf{x})$. The important message is that the extent to which $\rho(\mathbf{x})$ varies with $\mathbf{x}$ determines the magnitude of any bias that might be present. If $\rho(\mathbf{x})$ is constant over $\mathbf{x}$ then $D_j(\mathbf{x})=0$ for any $\mathbf{x}$, bias disappears, and the 2PM in conjunction with the homoskedastic smearing estimator is -- as would be expected in such circumstances -- a perfectly satisfactory estimator upon which to construct estimates of the partial effects or elasticities.

IV. RECONSIDERING ESTIMATION: TWO ALTERNATIVES

A. A Reformulation: The Modified Two-Part Model (M2PM)

Main Ideas

The central idea of this paper is the following. A model that captures the basic essence of -- but is in general not identical to -- part two of the homoskedastic 2PM replaces (4) with the *assumption* that

$$\begin{aligned} E[y|y>0,\mathbf{x}] &= \exp(\mathbf{x}\beta_M) \\ &= \mu_M(\mathbf{x}) \end{aligned} \tag{26}$$

so that

---

[15] Yet as the next section demonstrates, there are alternative estimation strategies that obviate the need for such speculation and/or auxiliary estimation and that permit direct unbiased (or, more accurately, consistent) estimation of the means and partial effects of interest.

$$y = \exp(\mathbf{x}\beta_M) \times \exp(\varepsilon_M), \quad y>0, \tag{27}$$

where $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]=1$[16] and where the symbol β_M is used to distinguish this parameter from β in the homoskedastic 2PM yet to emphasize that they play similar roles. The contrast between (26) and (13) is the fundamental distinction between this reformulation -- which will be referred to as M2PM, the "M" prefix standing for "modified" -- and the 2PM model retransformed via the standard homoskedastic smearing estimator.

While dampening the influence of skewness and high-end outliers on parameter estimates as well as other considerations have commonly -- and, it should be stressed, not unreasonably -- been advanced as rationales for specification of the log-linear model (3), the analytical core of the 2PM formulation is the linear conditional mean function $\mathbf{x}\beta$ for $\ln(y)|y>0,\mathbf{x}$ regardless of the skewness or other properties of $f(y|y>0,\mathbf{x})$. Central to this specification is the fact that $y|y>0,\mathbf{x}$ is a positive random variable whose logarithmic transformation permits an easily-estimated linear model to properly characterize the conditional mean without restriction (i.e. $\mathbf{x}\beta$ can be positive or negative just as $\ln(y)$ can be positive or negative).

Analogous reasoning suggests that the critical restriction on $E[y|y>0,\mathbf{x}]$ is $E[y|y>0,\mathbf{x}]>0$. The leading practice for specification of the conditional mean of necessarily nonnegative

---

[16] Assuming that the $\mathbf{x}$ vector contains a constant one can then assume that $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]=1$ without any loss of generality. As a general matter in this case, the key requirement on $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]$ is that it is a constant not depending on $\mathbf{x}$.

and perhaps strictly positive random variables is, as in (26), to specify the ECM model $E[y|y>0,\mathbf{x}]=\exp(\mathbf{x}\beta_M)$.[17] It should also be emphasized that in the leading case considered in the 2PM literature where $f(y|y>0,\mathbf{x})$ is taken to be homoskedastic lognormal, (26) implies and is implied by (4), with the exception that the constant terms in β and β_M may differ in the two specifications owing to the offset by the variance parameter.

By respecifying the 2PM as (7) and (26) it is now possible to recover $E[y|\mathbf{x}]$ from the two parts of M2PM, viz

$$\begin{aligned} E[y|\mathbf{x}] &= \Pr(y>0|\mathbf{x}) \times E[y|y>0,\mathbf{x}] \\ &= \frac{\exp(\mathbf{x}\alpha)\exp(\mathbf{x}\beta_M)}{1+\exp(\mathbf{x}\alpha)} \\ &= \frac{\exp(\mathbf{x}(\alpha+\beta_M))}{1+\exp(\mathbf{x}\alpha)} \\ &= \Psi_M(\mathbf{x};\theta_M), \end{aligned} \tag{28}$$

where $\theta_M=[\alpha,\beta_M]$. Assembling $E[y|\mathbf{x}]$ now entails *two* components, not *three* as in the 2PM (6). As such, one can write the regression model as

$$y = \Psi_M(\mathbf{x};\theta_M) \bigstar u_M, \quad \forall y, \tag{29}$$

where the key orthogonality restriction is $E[u_M|\mathbf{x}]=1\text{-}1\bigstar 0$.

---

[17] This ECM assumption is prominent in, e.g., the count model literature. See Pohlmeier and Ulrich, 1995, for a count model application of two-part models in a health care utilization context.

Basic Properties of $\Psi_M(\mathbf{x};\theta_M)$

The partial relationship $\delta_j(\mathbf{x})$ between the conditional expectation $E[y|\mathbf{x}]=\Psi_M(\mathbf{x};\theta_M)$ and the j-th element of $\mathbf{x}$, x_j is given by

$$\frac{\partial\Psi_M(\mathbf{x};\theta_M)}{\partial x_j} = \Psi_M(\mathbf{x};\theta_M)\{\ \beta_{Mj} + (1-\pi(\mathbf{x}))\alpha_j\ \} \qquad (30)$$

and

$$\frac{\partial^2\Psi_M(\mathbf{x};\theta_M)}{\partial x_j^2} =$$

$$\Psi_M(\mathbf{x};\theta_M)\{\ [\beta_{Mj} + (1-\pi(\mathbf{x}))\alpha_j]^2 - \pi(\mathbf{x})(1-\pi(\mathbf{x}))\alpha_j^2\}. \qquad (31)$$

If $\mathrm{sgn}(\alpha_j)=\mathrm{sgn}(\beta_{Mj})$ then it is clear from (30) that $\mathrm{sgn}(\partial\Psi_M(\mathbf{x};\theta_M)/\partial x_j)=\mathrm{sgn}(\alpha_j)=\mathrm{sgn}(\beta_{Mj})$. If $\mathrm{sgn}(\alpha_j)\neq\mathrm{sgn}(\beta_{Mj})$, then several possibilities emerge. If $\mathrm{abs}(\alpha_j)<\mathrm{abs}(\beta_j)$ then $\mathrm{sgn}(\partial\Psi_M(\mathbf{x};\theta_M)/\partial x_j)=\mathrm{sgn}(\beta_{Mj})$ unambiguously. However, if $\mathrm{abs}(\alpha_j)>\mathrm{abs}(\beta_{Mj})$ then there may be values of x_j at which $\pi(\mathbf{x})$ is sufficiently large or small to result in a change in $\mathrm{sgn}(\partial\Psi_M(\mathbf{x};\theta_M)/\partial x_j)$ as x_j -- and, therefore, $\pi(\mathbf{x})$ -- vary. If x_j varies over $(-\infty,+\infty)$, this will be true in general.[18]

---

[18] The sign of $\partial^2\Psi_M(\mathbf{x};\theta_M)/\partial x_j^2$ -- i.e. the convexity/concavity of $\Psi_M(\mathbf{x};\theta_M)$ -- is more complicated to determine. If $\mathrm{sgn}(\alpha_j)=\mathrm{sgn}(\beta_{Mj})=1$, then $\partial^2\Psi_M(\mathbf{x};\theta_M)/\partial x_j^2>0$ unambiguously for $\pi(\mathbf{x})<.5$. If $\mathrm{abs}(\alpha_j)<\mathrm{abs}(\beta_{Mj})$ and $\mathrm{sgn}(\alpha_j)\times\mathrm{sgn}(\beta_{Mj})=1$ then $\partial^2\Psi_M(\mathbf{x};\theta_M)/\partial x_j^2$ is also unambiguously positive. Note finally that (continued)

The elasticities of $\Psi_M(\mathbf{x};\theta_M)$ in the M2PM model have a particularly simple form:

$$\eta_j(\mathbf{x}) = \frac{\partial \ln(\Psi_M(\mathbf{x};\theta_M))}{\partial \ln(x_j)} = (1-\pi(\mathbf{x}))\alpha_j x_j + \beta_{Mj} x_j, \qquad (32)$$

i.e. the $\eta_j(\mathbf{x})$ are just the logit probability elasticity plus the (conditional) ECM elasticity.

Two-Step Estimation (M2PM-2)

The idea underlying two-step estimation of the M2PM model is fully analogous to that of the usual 2PM model. In step one, a logit (or probit) model is used to estimate α in $\Pr(y>0|\mathbf{x};\alpha)$. In step two, nonlinear least squares (NLLS) or some comparable method with residual function $y-\exp(\mathbf{x}\beta_M)$ is used in the subsample S_+ to estimate β_M.[19]

The key orthogonality condition for identification here is

---

at $\pi(\mathbf{x})=.5$ $\partial^2\Psi_M(\mathbf{x};\theta_M)/\partial x_j^2$ equals simply $\Psi_M(\mathbf{x};\theta_M)\beta_{Mj}(\beta_{Mj}+\alpha_j)$.

[19] The literature on ECM models recognizes that either an additive or multiplicative error can be maintained since the two will, in general, be observationally equivalent. See Wooldridge, 1992, for additional discussion. Inference for NLLS is conducted using the heteroskedasticity-robust covariance estimator (Davidson and MacKinnon, 1993)

$$\hat{V}(\hat{\beta}_M) = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\hat{\Omega}\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1},$$

where $\mathbf{G}$ is the (T×p) matrix of estimated gradients $[\nabla_{\beta_M}\mu(\mathbf{x};\hat{\beta}_M)]$ and $\hat{\Omega}$ is the (T×T) diagonal matrix of squared NLLS residuals. This basic setup is used for all NLLS estimation undertaken here.

that $\rho_M(\mathbf{x})=E[\exp(\varepsilon_M)\,|\,y>0,\mathbf{x}]$ is a constant not depending on $\mathbf{x}$; the normalization $\rho_M(\mathbf{x})=1$ will again be available so long as $\mathbf{x}$ contains a constant term.

One-Step Estimation (M2PM-1)

Most fundamentally, the M2PM model (28)-(29) is a nonlinear regression model for all y. As such, direct estimation of (28) in a single step via NLLS using all N observations might be contemplated. Such a strategy would maintain the orthogonality restriction $E[u_M|\mathbf{x}]=1-1\bigstar 0$ as noted after (29). Cast thusly, estimation via M2PM-1 is a nonlinear regression problem executed on the *full* sample of observations on y and $\mathbf{x}$.

Yet it should be pointed out that an identification issue arises with the M2PM-1 estimator of $\theta=[\alpha,\beta_M]$.[20] Specifically, the T×2k gradient matrix $\mathbf{G}(\theta_M)=[\nabla_{\theta_M}\Psi_M(\mathbf{x};\theta_M)]$ will fail to have full column rank at the null hypothesis $\theta=[\alpha,\beta_M]=0$, although $\mathbf{G}(\theta_M)$ will generally have full column rank at values of θ_M away from zero. If the true θ_M is nonzero, then the model is identified in the sense of $\mathbf{G}(\theta_M)$ having full column rank at the true parameter value.[21]

---

[20] In the empirical example reported below, attempts to estimate θ_M using the starting values $[\alpha,\beta_M]=0$ were, not surprisingly, unsuccessful. Convergence was not problematic, however, when the starting values used were the converged values from the M2PM-2 estimator, for which convergence was quite rapid (the globally concave logit component and the ECM specification for observations y>0).

[21] This identification question would appear to be a general
(continued)

B. An Exponential Conditional Mean Model for E[y/x]

Main Ideas

Probably the most straightforward parametric assumption consistent with the requirement that $E[y|\mathbf{x}]>0$ is to assume that the distribution of $y|\mathbf{x}$ *no longer conditional on* $y>0$ has an exponential conditional mean structure with a linear index function, i.e. $E[y|\mathbf{x}]=\exp(\mathbf{x}\zeta)$. In this instance, the outcomes y are generated by the model

$$\begin{aligned} y &= \Psi_E(\mathbf{x};\zeta) \bigstar u_E, \quad \forall y \\ &= \exp(\mathbf{x}\zeta) \bigstar u_E, \quad \forall y, \end{aligned} \tag{33}$$

where $E[u_E|\mathbf{x}]=1-1\bigstar 0$ is maintained.

Properties

If it can reasonably be maintained that $E[y|\mathbf{x}]=\exp(\mathbf{x}\zeta)$ then it is possible to exploit several potentially very useful properties of ECM specifications. First, with an ECM specification it is possible to use instrumental variable methods to obtain consistent estimates of ζ should some elements of $\mathbf{x}$ be correlated with unobservable determinants of the conditional mean. That is, with unobservables Θ, if

$$E[y|\mathbf{x},\Theta] = \exp(\mathbf{x}\zeta + \Theta) \tag{34}$$

---

property of any regression model whose conditional expectation function is defined on two or more linear index functions in the same covariate vector $\mathbf{x}$. See Ichimura and Lee, 1991, for additional discussion.

with $E[\Theta|\mathbf{x}]$ some nontrivial function of $\mathbf{x}$, then standard methods like nonlinear least squares or quasi-ML will be inconsistent for ζ. However, given classical instruments $\mathbf{z}$, a GMM-type IV estimator can circumvent this difficulty and provide consistent estimates (see Mullahy, 1997a). It should be stressed that this estimation strategy is *not* available with most nonlinear expectation function specifications.

Second, the ECM specification results in linear elasticities:

$$\eta_j(\mathbf{x}) = \frac{\partial \ln(\Psi_E(\mathbf{x};\zeta))}{\partial \ln(x_j)} = x_j\zeta_j . \qquad (35)$$

These are markedly simpler, and perhaps more tenable *a priori*, than the elasticities from the 2PM and even the M2PM which -- from (6) and (28) -- obviously consist of three and two summands, respectively.

Finally, in keeping with the idea that the parametric functional forms considered here are all based on linear index functions, it should be emphasized that even a relatively simple specification of the conditional mean $E[y|\mathbf{x}]$ like ECM is capable of capturing important nonlinearities in parameters like the partial effects if sufficiently rich definitions of the covariate vector $\mathbf{x}$ are employed. As such, a choice between a two-part specification (e.g. M2PM or 2PM) with $\mathbf{x}$ containing main effects only *versus* a one-part model like ECM with $\mathbf{x}$ redefined to include, e.g., low-order polynomials in and interactions between the main effects might ultimately suggest a preference for a one-part estimation strategy (although, as discussed below, this is

largely a testable proposition).[22]

C. Specification Testing

M2PM vs. 2PM

A central implication of the preceding discussion is that if $\rho(\mathbf{x})$ is constant over $\mathbf{x}$ then both the 2PM and the M2PM estimates of the slope parameters β_1 and β_{M1} should converge in the limit to the same (true) value. Conversely, if the 2PM and M2PM estimates of these slope parameters diverge significantly, this would be a strong indication that the M2PM estimate is absorbing some of the effect of $\mathbf{x}$ that works through $\rho(\mathbf{x})$ which, of course, the 2PM estimate fails to do at all.

A simple diagnostic for such departures can be conducted via a split-sample test, as follows. Randomly split the S_+ sample into two subsamples of approximately $N_+/2$ observations each (e.g. assign each observation to one or other subsample as a pseudo-random uniform variate is greater or less than .5). In the first subsample regress ln(y) on $\mathbf{x}$ via linear regression to estimate β; in the second subsample estimate β_M in $\exp(\mathbf{x}\beta_M)$ via nonlinear regression. Denote the respective estimates of the slope parameters $\hat{\beta}_1$ and $\hat{\beta}_{M1}$ and their corresponding (heteroskedasticity-robust) covariance matrix estimates $\hat{V}_1$ and $\hat{V}_{M1}$. Since the subsamples are independent by construction, then the test statistic

---

[22] Of course, richer definitions of the covariate vector $\mathbf{x}$ can be used in any of the linear index function formulations considered here.

$$w = (\hat{\beta}_1 - \hat{\beta}_{M1})'(\hat{V}_1 + \hat{V}_{M1})^{-1} (\hat{\beta}_1 - \hat{\beta}_{M1}) \quad (36)$$

can be treated as a $\chi^2_{(k-1)}$ Wald test statistic under the null hypothesis of parameter equality.

M2PM vs. ECM

Note that the M2PM model nests two important models. If the data suggest $\alpha_1=0$ then covariates are not important features of the hurdle process. Alternatively, if $\beta_{M1}=0$ then only the hurdle part of the model is significantly affected by covariates. Perhaps the key point to note here is that the M2PM model (28) effectively reduces to the ECM model (33) under the restriction $\alpha_1=0$. Letting $\lambda=\beta_0+\ln\left(\frac{\exp(\alpha_0)}{1+\exp(\alpha_0)}\right)$, expression (28) reduces to $\left(\frac{\exp(\alpha_0)\times\exp(\mathbf{x}\beta_M)}{1+\exp(\alpha_0)}\right)=\exp(\mathbf{x}_1\beta_{M1}+\lambda)$ -- an ECM specification -- when $\alpha_1=0$.

Goodness of Fit

Considerations of how well the estimated model fits the data may also be of importance in some applications. Since the estimation strategies considered here are based most fundamentally on specifications of the conditional mean functions $E[y|\mathbf{x};\theta]$ as opposed to specifications of the likelihood functions $\ell(\theta|y;\mathbf{x})$, formal χ^2 or probability-based goodness-of-fit tests (Andrews, 1988) are not available. This does not imply, however, that analysts should not conduct alternative forms of goodness-of-fit testing such as graphical analysis of predicted values vs.

residuals, RESET-type Conditional Moment (CM) tests (Pagan and Vella, 1989), etc.

The goodness-of-fit tests undertaken here are operationalized as CM specification tests. For CM tests, moment functions $m(y,\mathbf{x},\theta)$ are specified such that $E_{\mathcal{N}}[m(y,\mathbf{x},\theta)|\mathbf{x}]=0$ under the null ("$\mathcal{N}$") hypothesis of correct moment specification but $E_{\mathcal{A}}[m(y,\mathbf{x},\theta)|\mathbf{x}]\neq 0$ under the alternative ("$\mathcal{A}$") hypothesis of moment function misspecification. Then $E_{\mathcal{N}}[q(\mathbf{x})'m(y,\mathbf{x},\theta)]=0$ unconditionally while $E_{\mathcal{A}}[q(\mathbf{x})'m(y,\mathbf{x},\theta)]\neq 0$ in general given suitably defined 1×r vectors $q(\mathbf{x})$.[23] Given a consistent estimate $\hat{\theta}$ of θ, the empirical analog of $\Gamma(\theta)=E[q(\mathbf{x})m(y,\mathbf{x},\theta)]$, $\Gamma(\hat{\theta})=N^{-1}\sum_{i=1}^{N} q(\mathbf{x}_i)'m(y_i,\mathbf{x}_i,\hat{\theta})$, is used as the basis of a test. Under the null hypothesis of correct specification, the statistic

$$w(\hat{\theta};\mathbf{R}) = (\mathbf{R}\Gamma(\hat{\theta}))'(\mathbf{R}\hat{\mathbf{V}}(\Gamma(\hat{\theta}))\mathbf{R}')^{-1}(\mathbf{R}\Gamma(\hat{\theta})) \qquad (37)$$

will have a $\chi^2_{\text{rank}(\mathbf{R})}$ distribution, where $\hat{\mathbf{V}}(\Gamma(\hat{\theta}))$ is the estimated covariance matrix of $\Gamma(\hat{\theta})$, and $\mathbf{R}$ is a s×r ($s\leq r$) selection matrix.[24]

---

[23] The choice of $q(\mathbf{x})$ influences the power of the test.

[24] Often $\mathbf{R}=I_r$, but the more general formulation will be useful for the empirical application presented below. Note that if $\mathbf{R}$ picks out only one element of $\Gamma(.)$, then the signed $\sqrt{w(\hat{\theta};\mathbf{R})}$ serves as a "t-statistic" for that particular orthogonality test.

The particular concern here is with proper specification of the conditional mean function $E[y|\mathbf{x}]$. As such, $m(y,\mathbf{x},\theta)$ is specified as

$$m(y,\mathbf{x},\theta) = y - \Psi(\mathbf{x};\theta). \tag{38}$$

$q(\mathbf{x})$ will be specified to contain linear and quadratic terms in, and interations between, elements of $\mathbf{x}$, as will be described below. These goodness-of-fit tests will be conducted for the 2PM, M2PM-2, and ECM models as well as for an OLS specification that serves as a useful benchmark. It should be stressed that these CM tests are nonnested in the sense of not being able to offer an unambiguous recommendation of one functional form relative to any other. However, the relative magnitudes of the estimated χ^2 statistics should still provide a useful assessment of the performance of each individual specification and an indication of any particular shortcomings each may suffer.

Prediction Biases with M2PM and ECM

It should be noted that the mean prediction errors for M2PM-2 and ECM estimators will generally be nonzero. Note first that even the *conditional* MPE for part two of M2PM-2 will not generally be forced to zero, which is evident from the first-order conditions defining $\hat{\beta}_M$:

$$\sum_{i \in S_+} [y_i - \exp(\mathbf{x}_i\beta_M)] \times \exp(\mathbf{x}_i\beta_M) \times \mathbf{x}_i' = 0. \tag{39}$$

Even if $\mathbf{x}_i$ contains a constant term, there is no feature of these

solution equations that sums up the prediction errors $y_i - \exp(\mathbf{x}_i\beta_M)$ over S_+ to be zero. Similarly, the first-order conditions solved by ECM's $\hat{\zeta}$ do not result in a MPE that is forced to zero over all observations,

$$\sum_{i \in \{S_+ \cup S_0\}} [y_i - \exp(\mathbf{x}_i\zeta)] \times \exp(\mathbf{x}_i\zeta) \times \mathbf{x}_i' = 0. \qquad (40)$$

In both cases, however, it is possible to use weighted variants of the estimating equations (39) and (40) to attain zero MPE -- conditional on y>0 in the former case -- as well as consistent estimates of β_M and ζ. For instance, a standard Poisson regression estimation algorithm amounts to using weights $\exp(-\mathbf{x}_i\beta_M)$ and $\exp(-\mathbf{x}_i\zeta)$, respectively, in (39) and (40), yielding the estimating equations

$$\sum_{i \in S_+} [y_i - \exp(\mathbf{x}_i\beta_M)] \times \mathbf{x}_i' = 0 \qquad (41)$$

and

$$\sum_{i \in \{S_+ \cup S_0\}} [y_i - \exp(\mathbf{x}_i\zeta)] \times \mathbf{x}_i' = 0, \qquad (42)$$

respectively. If $\mathbf{x}_i$ contains a constant term, then the respective first-order conditions force a zero mean prediction error (as is the case for logit, least squares, etc.).

For M2PM-2, however, even a zero conditional (on y>0) MPE will not generally yield an exact zero overall MPE. A bit of algebra shows that the M2PM-2 *sample* MPE based on the mean prediction $N^{-1}\sum_{i=1}^{N} \hat{\pi}_i\hat{\mu}_{Mi}$ equals

$$N^{-1}\sum_{i=1}^{N} \hat{\pi}_i\hat{\mu}_{Mi} - \bar{y} =$$

$$\text{Cov}(\hat{\pi}_i, \hat{\mu}_{Mi}) + \left[\bar{\hat{\pi}} \times \left(1 - \bar{\hat{\pi}}\right) \times \left(\overline{\hat{\mu}_{M0}} - \overline{\hat{\mu}_{M+}}\right)\right], \qquad (43)$$

where $\bar{y}$ is the full sample mean of the y_i, $\bar{\hat{\pi}}$ is the sample mean of the $\hat{\pi}_i$, $\overline{\hat{\mu}_{M0}}$ and $\overline{\hat{\mu}_{M+}}$ are the S_0 and S_+ subsample means of the predicted $\hat{\mu}_{Mi}$, and the Cov(.) expression is the sample covariance (normed by N^{-1}, not $(N-1)^{-1}$) between the predicted probabilities and predicted conditional means. As such, even if the M2PM-2 estimate $\bar{\hat{\pi}}$ of $E_{\mathbf{x}}[\pi(\mathbf{x})]$ gives exactly the marginal proportion of positive y values (as would logit) and even if the M2PM-2 estimate $\overline{\hat{\mu}_{M+}}$ of $E_{\mathbf{x}}[\mu_M(\mathbf{x})]$ gives exactly the marginal mean of the y *conditional on* y>0 (as would the weighted strategy suggested above) the overall mean prediction will generally not precisely equal $\bar{y}$.

D. Kindred Results from the Count Data Literature

The conditional mean function $\Psi_M(\mathbf{x};\theta_M)$ given in (28) and the specification test for M2PM vs. ECM discussed above have well-known counterparts in the econometric literature on count data. In particular, one prominent class of models developed in that literature to accommodate conditional probability distributions of counts where an excess[25] of zero outcomes is observed is the so-called "zero-inflated" or "with-zeros" class, with the zero-inflated Poisson ("ZIP") model probably the leading

[25] "Excess" is, of course, defined relative to some null, like the Poisson.

example (Cameron and Trivedi, 1996; Lambert, 1992; Mullahy, 1986). Letting $Pr_{\wp}(y|\mathbf{x})$ denote a null Poisson distribution for $y\in\{0,1,2,...\}$, the ZIP model is given by the conditional probability distribution

$$Pr_{ZIP}(y|\mathbf{x}) = \mathbf{1}(y=0)\varphi(\mathbf{x}) + (1 - \varphi(\mathbf{x}))Pr_{\wp}(y|\mathbf{x}), \quad y\in\{0,1,2,...\}, \tag{44}$$

where it is usually specified that $\varphi(\mathbf{x})\in(0,1)$.[26] The conditional mean function for the ZIP model is given by

$$E_{ZIP}[y|\mathbf{x}] = (1 - \varphi(\mathbf{x}))E_{\wp}[y|\mathbf{x}]. \tag{45}$$

If, as is standard, $E_{\wp}[y|\mathbf{x}]$ is specified as $\exp(\mathbf{x}\beta_{\wp})$ and $\varphi(\mathbf{x})$ is specified as a logit function $\exp(\mathbf{x}\alpha_{ZIP})/(1+\exp(\mathbf{x}\alpha_{ZIP}))$, then it is readily apparent that $E_{ZIP}[y|\mathbf{x}]$ has the same functional form as the M2PM conditional mean function $\Psi_M(\mathbf{x};\theta_M)$ in (28). Moreover, if -- as sometimes maintained -- $\varphi(\mathbf{x})$ is taken to be constant over $\mathbf{x}$ ($\varphi(\mathbf{x})=\varphi$ for any $\mathbf{x}$), then $E_{ZIP}[y|\mathbf{x}]$ reduces simply to an ECM specification like (33), differing from $E_{\wp}[y|\mathbf{x}]=\exp(\mathbf{x}\beta_{\wp})$ only by the constant term parameter in $\beta_{\wp}$.[27]

---

[26] Some range of negative values for $\varphi(\mathbf{x})$ can be admitted to accommodate a deficit of zeros relative to the null, but this is not typically a consideration.

[27] It might be noted, too, that (26) has the same structure as would arise if the y were generated by a Poisson(λ) mixture (on N) of binomial(N,p) variates; see Johnson et al., 1992, chapter 9.5 for discussion.

E. 2PM vs. M2PM and ECM: Summing Up

If contemplating a two-part modeling strategy, the preceding arguments suggest that the analyst should, at a minimum, confront directly the question: Why not just start out in the first instance with the M2PM specification? That is, since the M2PM model *maintains* that $y=\exp(\mathbf{x}\beta_M)\exp(\varepsilon_M)$ for $y>0$, and that $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]=1$, then nonlinear least squares applied to the residual function $y-\exp(\mathbf{x}\beta_M)$ should provide a consistent estimate of β_M *as well as* a consistent estimate of $E[y|y>0,\mathbf{x}]$ $(=\exp(\mathbf{x}\hat{\beta}_M))$ without any further recourse as to how $\mathbf{x}$ affects the shape of the probability distribution of ε.

Two distinct but obviously related issues should be considered at this juncture. The first is whether the estimation strategy being contemplated can result in consistent estimates of the parameters (β or β_M) of the index function in $E[y|y>0,\mathbf{x}]$. For the homoskedastic 2PM, the key requirement for consistent estimation of β is the orthogonality condition $E[\varepsilon|y>0,\mathbf{x}]=0$, as discussed above. For the M2PM model estimated in two steps, the key orthogonality condition is that $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]$ is a constant not depending on $\mathbf{x}$.

The second issue is whether the contemplated estimation strategy can provide consistent estimates of important quantities like the partial effects $\delta(\mathbf{x})$ or elasticities $\eta(\mathbf{x})$ The models under consideration for $E[y|y>0,\mathbf{x}]$ and, therefore, $E[y|\mathbf{x}]$ are nonlinear so these partial effects will involve more than just the parameters β_j or β_{Mj}. In the case of the 2PM, the joint requirements that $E[\varepsilon|y>0,\mathbf{x}]=0$ and that $E[\exp(\varepsilon)|y>0,\mathbf{x}]$ is a constant not depending on $\mathbf{x}$ suggest that *statistical independence*

between ε and $\mathbf{x}$ may be the only reasonable assumption sufficiently general to support such joint requirements. For the M2PM specification, the assumption that $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]$ is a constant not depending on $\mathbf{x}$ is already the basis of consistent parameter estimation. No *additional* assumptions need be made to support consistent estimation of quantities like $\delta(\mathbf{x})$ and/or $\eta(\mathbf{x})$.

V. A SIMULATION EXPERIMENT

One clear implication of the preceding discussion is that while the 2PM may be a consistent estimator of the *parameters* β, its utility as concerns estimation of the *partial effects* $\partial E[y|y>0,\mathbf{x}]/\partial \mathbf{x}$ and, therefore, $\delta(\mathbf{x})$, may be limited. A brief simulation experiment underscores the importance of this distinction. (For purposes of this section, β will denote both β from 2PM as well as β_M from M2PM.)

The design is intentionally one where *neither* the 2PM *nor* the M2PM-2 estimator will be a consistent estimator of the partial effects $\partial E[y|y>0,\mathbf{x}]/\partial x_j$. However, the 2PM estimator will be consistent for β, whereas the M2PM-2 estimator will not (due to violation of the orthogonality condition that $\rho(\mathbf{x})=E[\exp(\varepsilon_M)|y>0,\mathbf{x}]$ is a constant not depending on $\mathbf{x}$). The objective of this exercise is to assess the extent to which even a consistent estimator of β (2PM) may provide misleading inferences about the partial effects $\partial E[y|y>0,\mathbf{x}]/\partial x_j$ relative to an estimator that is known to be inconsistent for β but that at least absorbs to some degree the dependence of $\rho(\mathbf{x})$ on $\mathbf{x}$ (which 2PM based on the homoskedastic smearing estimator does not).

The design is as follows. The model is

$$y = \exp(\beta_0 + \beta_1 x_1 + \varepsilon), \quad y>0 \tag{46}$$

with x_1 a scalar and $\beta_0=\beta_1=.5$. As in Example 1 in section III.A above, let $v\sim U[-.5,.5]$ and $x_1\sim U[0,1]$ be statistically independent pseudo-random uniform variates. Define $\varepsilon=x_1 v$. As such, $E[\varepsilon|y>0,\mathbf{x}]=E[x_1 v|y>0,\mathbf{x}]= x_1 E[v|y>0,\mathbf{x}]=x_1 E[v|y>0]=0$, so the orthogonality condition required for 2PM to be consistent for β is satisfied. However, note that

$$\begin{aligned} E[\exp(\varepsilon)|y>0,\mathbf{x}] &= \int_{-.5}^{.5} \exp(x_1 v)dv \\ &= \frac{\exp(.5x_1) - \exp(-.5x_1)}{x_1}, \quad x_1>0, \end{aligned} \tag{47}$$

which is clearly a nontrivial function of $\mathbf{x}$. As such, the fundamental orthogonality requirement for consistency of part two of the M2PM-2 estimator is violated. Moreover, since $E[\exp(\varepsilon)|y>0,\mathbf{x}]$ does not have a representation as an ECM function here, it thus does not follow that the γ parameters in $\rho(\mathbf{x};\gamma)$ get "rolled in" to the β parameters in $\mu(\mathbf{x};\beta)$.

The simulation uses 1,000 replications of a sample size of N=5,000 observations. The vector x_1 is drawn once and then held constant across the 1,000 draws of the vector v. To be accumulated are the 2PM and M2PM-2 estimates of β_1 at each replication as well as the sample median estimate of the true partial effects at each replication

$$\frac{\partial E[y \mid y>0, \mathbf{x}]}{\partial x_1} = \frac{x_1[\beta_1 \Delta(-) + .5\Delta(+)] - \Delta(-)}{x_1^2}, \qquad (48)$$

where $\Delta(-) = \exp(\beta_0+\beta_1 x_1+.5) - \exp(\beta_0+\beta_1 x_1-.5)$ and $\Delta(+) = \exp(\beta_0+\beta_1 x_1+.5) + \exp(\beta_0+\beta_1 x_1-.5)$. Based on the single draw of x_1, and given the true parameters $\beta_0=\beta_1=.5$, the sample mean and median of these partial effects are 1.186 and 1.160, respectively. The estimates of β_1 and of the partial effects are then summarized by the sample means and medians of these estimates over the 1,000 replications. The partial effects are computed using the estimates of $\exp(\beta_0+\beta_1 x_1)\beta_1$ for M2PM-2 and $\phi \times \exp(\beta_0+\beta_1 x_1)\beta_1$ for 2PM, with ϕ estimated by the standard homoskedastic smearing method.

The results are provocative. Not unexpectedly, the 2PM estimates of β_1 outperform those obtained via M2PM-2: mean/median .500/.500 vs. .544/.545, respectively. However, the superior performance in estimation of β_1 is swamped by the failure of the homoskedastic 2PM smearing estimator of $\rho(\mathbf{x})$ to account for any dependence of $\rho(\mathbf{x})$ on $\mathbf{x}$. For the homoskedastic 2PM, the mean and median of the partial effect estimates are 1.074 and 1.074. For M2PM-2, the mean and median partial effect estimates are 1.169 and 1.169, far closer to the true sample mean and median values. As such, despite the fact that the M2PM-2 estimator is the "wrong" estimator for this design, the fact that it incorporates some degree of dependence of $E[\exp(\varepsilon) \mid y>0, \mathbf{x}]$ on $\mathbf{x}$ results (at least in this example) in its superior performance as an estimator of the partial effects of interest.

VI. AN EMPIRICAL EXAMPLE OF HEALTH CARE UTILIZATION

This section presents some empirical illustrations of the concepts and issues discussed in the previous sections. The various estimators are compared and contrasted in terms of their performance in a single sample, and the results of some specification tests are reported.

A. Sample and Estimators

The estimation sample of N=36,111 observations on individuals ages 25-64 is drawn from the 1992 National Health Interview Survey. The dependent variable is the number of doctor visits in the twelve-month period prior to the survey. For this measure of the dependent variable, y=0 in N_0=8,513 cases (23.6%) and y>0 in N_+=27,598 cases (76.4%). The list of covariates is as described in table 1 and the sample frequency distribution of the visits measure is presented in table 2.[28]

The five models estimated here are as follows:

(1) ECM estimated on the full sample (ζ);

(2) two-step M2PM (M2PM-2) (α, β_M);

(3) M2PM conditional mean function estimated in one step on the full sample[29] (M2PM-1) (α, β_M);

---

[28] In a preliminary specification, Age Squared was included as a covariate but in some cases (ECM) was not statistically significant. To facilitate interpretation of the results, it is not included in the specifications reported below. Whether its exclusion is statistically significant should be determined in part by the results of the CM specification tests.

[29] The M2PM-2 estimates were used as starting values here. Convergence was problematic when alternative arbitrary starting

(continued)

(4) standard logit/loglinear 2PM with standard homoskedastic smearing retransformations (α,β);

(5) standard logit/loglinear 2PM with a nonlinear regression for an ECM smearing retransformation, i.e. $\rho(\mathbf{x})=\exp(\mathbf{x}\gamma)$.

B. Results

The point estimates are reported in table 3. Columns 1 and 2 report the estimates of α from the usual logit estimator and from the M2PM-1 estimator, respectively. In all cases the signs of the point estimates are the same, although in some cases the magnitudes and/or the significance levels differ markedly (e.g. Male, Schooling). As a general matter, but not unexpectedly, the M2PM-1 point estimates are noisier than the logit estimates even though the estimates are computed off the same sample.

Columns 3-5 of table 3 report the estimates of β and β_M obtained from 2PM, M2PM-1, and M2PM-2. For 2PM and M2PM-2 the estimates are based on the S_+ subsample whereas the M2PM-1 estimates use the entire sample. With one exception (Male for M2PM-1) the signs of the point estimates are the same across the three estimators. The magnitudes and significance levels are, again, quite variable across the three sets of results, with the 2PM results tending to have markedly larger asymptotic t-statistics than M2PM-2 and, particularly, M2PM-1.[30]

---

values were used. The ECM and M2PM models were estimated in part using programs written by the author. Stata versions of the ECM program and of the dataset are available on request, preferably via email.

[30] All t-statistics reported here are based on (continued)

Estimates of the ECM formulation of $E[y|\mathbf{x}]$ are reported in column 6 of table 3. The signs of the individual point estimates are the same as the 2PM and the M2PM point estimates, and the magnitudes are quite similar as well. Since over 76 percent of the observations in this particular sample are on $y>0$, this similarity of results is not terribly surprising. In a different sample where the fraction of observations on $y=0$ is larger, it is a fairly safe bet that the ECM results for ζ would tend to diverge more dramatically from the 2PM and M2PM estimates of β and β_M.

The last column of table 3 reports the results of a NLLS estimation where the 2PM exponentiated residuals are the dependent variable, the assumption being that $\rho(\mathbf{x})=\exp(\mathbf{x}\gamma)$. While some caution should probably be exercised in interpreting the estimated standard errors, the signs and magnitudes of the point estimates of γ_j are informative. For instance, for Age, Married, and Excellent the M2PM point estimates of β_{Mj} are markedly larger (in absolute value) than the 2PM estimates. This result may be reconciled in part by the corresponding negative point estimates of γ_j which, when added to the corresponding 2PM estimates of β_j would bring the two sets of estimates more closely in line. Conversely, for Male the 2PM point estimate of β_j is larger in magnitude than the M2PM-2 point estimate. However, the *positive* point estimate of γ_j in this case again serves to partially reconcile the estimates. In all cases, discrepancies between the 2PM and M2PM point estimates and their

---

heteroskedasticity-robust covariance estimators.

(at least partial) reconciliation on the basis of the estimates of γ underscore the importance of accounting for the possible dependence of $\rho(\mathbf{x})$ on $\mathbf{x}$ if sound inferences are to be forthcoming. It should be reiterated (yet again) that the M2PM approach embeds the dependence of $\rho(\mathbf{x})$ on $\mathbf{x}$ as a maintained assumption while the ECM model circumvents the issue altogether.

Table 4 computes for ECM, M2PM-1, M2PM-2, and homoskedastic 2PM the estimated partial effects $\delta_j(\mathbf{x})$ for the continuous covariates Age and Schooling. These partial effects are computed for each observation in the full sample (N=36,111), with the sample quartiles and sample mean reported in table 4. For both Age and Schooling the results are striking. The quartiles and means of the estimated effects from the ECM specification are quite close to those from M2PM-1 and M2PM-2. For both covariates, however, the estimated magnitudes of the partial effects for the 2PM model diverge dramatically from those obtained using the other three methods. Of course, the usual 2PM estimator of $E[y|\mathbf{x}]$ based on a homoskedastic smearing retransformation is the only one of these estimators that does *not* account -- either implicitly or explicitly -- for the possible nontrivial dependence of the distribution of $\varepsilon|y>0,\mathbf{x}$ on $\mathbf{x}$.[31] The approximate comparability of the three other estimators suggests that consideration of such dependence is (at least in this example) of paramount importance, perhaps even moreso than correct specification of the functional form of the conditional

[31] For example, figure 1 plots the estimated exponentiated residuals from the standard 2PM against Schooling, giving a rather strong impression that $\rho(\mathbf{x})$ is a nontrivial function of at least one of its components.

mean function itself.

C. Specification Tests

With reference to the specification tests discussed above in section IV.C, the split-sample test statistic for equality of the point estimates of β_1 from 2PM and β_{M1} from M2PM-2 is 38.3 on 8 d.f. ($p<.00005$), suggesting compellingly rejection of the null and supporting an inference that $\rho(\mathbf{x})$ is a nontrivial function of $\mathbf{x}$ in these data.[32]

Second, for testing ECM against M2PM, the test of the null hypothesis $H_0:\alpha_1=0$ from the M2PM-1 model gives a $\chi^2_{(8)}$ statistic of 16.8 ($p=.032$). As such, there is some, albeit not overwhelming,[33] evidence against the ECM model in favor of the M2PM formulation.

The results of the CM tests for misspecification of $\Psi(\mathbf{x};\theta)$ are summarized in tables 5 and 6. The test statistics are

---

[32] The split-sample test was based on independent subsamples of sizes 13,844 and 13,754. To assess whether this single split-sample result might have somehow been anomalous, a simulation exercise was undertaken. The sample splitting was repeated randomly over 1,000 replications. The .05, .25, .50, .75, and .95 quantiles of the 8 d.f. test statistics were 21.7, 28.6, 34.5, 41.3, and 51.0, respectively. Even the .05 quantile test statistic over these replications would have a p-value less than .01. As such, it is probably safe to conclude that the 2PM and M2PM point estimates of β_1 and β_{M1} diverge significantly.

[33] A p-value of .032 on a sample of N=36,111 does not offer tremendously compelling evidence against the null. It should be noted, however, that the corresponding $\chi^2_{(8)}$ statistic from the logit model (column 1 of table 3) is 2177.2 with a p-value less than .00005.

computed for four estimators: standard 2PM based on the homoskedastic smearing estimator, M2PM-2, ECM, and a baseline OLS specification. In constructing $\Gamma(\theta)$, $q(\mathbf{x})$ is specified to contain all r=35 of the possible linear, quadratic, and interaction terms of the elements of the covariate vector $\mathbf{x}$ (excluding the constant term). The statistics are given by equation (37) with the selection matrix $\mathbf{R}$ defined alternatively as: $\mathbf{R}_1=\mathbf{I}_r$ (all elements of $\Gamma(\theta)$); $\mathbf{R}_2=[\mathbf{I}_{k-1},\mathbf{0}_{k-1,r-k+1}]$ (terms in $\Gamma(\theta)$ corresponding only to the linear elements of $q(\mathbf{x})$); $\mathbf{R}_3=[\mathbf{0}_{r-k+1,k-1},\mathbf{I}_{r-k+1}]$ (terms in $\Gamma(\theta)$ corresponding only to the interaction and quadratic elements of $q(\mathbf{x})$); and $\mathbf{R}_{4j}=[0,0,\ldots 0,1,0\ldots 0]$ (individual elements of $\Gamma(\theta)$ are selected to be tested by t-tests given by the signed $\sqrt{w(\hat{\theta};\mathbf{R})}$.[34] The covariance matrix estimate $\hat{\mathbf{V}}(\Gamma(\hat{\theta}))$ is obtained via a simple bootstrap covariance estimator based on b=1,000 bootstrap replications (Efron and Tibshirani, 1993).

The first noteworthy summary statistics in table 5 are the mean prediction error $MPE=N^{-1}\sum_{i=1}^{N}(\Psi(\mathbf{x}_i;\hat{\theta})-y_i)$ and mean squared error $MSE=N^{-1}\sum_{i=1}^{N}(\Psi(\mathbf{x}_i;\hat{\theta})-y_i)^2$ for the four estimators. (Note that only OLS forces the MPE to zero.) While the MPEs for the M2PM-2 and the ECM estimators are quite small, the MPE for the 2PM (=-0.175) is quite large (approximately 3.5% of the sample mean of the dependent variable), suggesting that -- at least for

---

[34] For the OLS specification, only the test statistics corresponding to $\mathbf{R}_3$ and to the $\mathbf{R}_{4j}$ for j>r are interesting because of its defining orthogonality restriction $\mathbf{x}'(y-\mathbf{x}\theta)=0$.

these data -- the homoskedastic 2PM systematically underpredicts the mean level of the dependent variable.[35] The corresponding MSEs suggest a preference ordering of M2PM-2 $\succ$ ECM $\succ$ OLS $\succ$ 2PM on this criterion.

Despite the fact that the CM tests are formally nonnested across the four model specifications, the preference ordering M2PM-2 $\succ$ ECM $\succ$ 2PM again would appear to hold for any of the selection matrix specifications $\mathbf{R}_1$, $\mathbf{R}_2$, or $\mathbf{R}_3$. In no instance does any χ^2 test statistic fall short of even the .0001 critical value. Of course, rejection of the null is not surprising given the fairly large sample size. If the less conservative and sample-size-sensitive Schwartz criteria (rank($\mathbf{R}$)$\times$ln(N); see Schwartz, 1978) are used instead of the standard χ^2 critical values to determine model acceptability, the M2PM-2 and ECM specifications are comfortably inside the acceptable range whereas the 2PM specification would have to be considered tenuous. When the OLS specification is considered with the $\mathbf{R}_3$ selection matrix, it is seen that its performance is roughly comparable to that of ECM but decidely inferior to that of M2PM-2. Finally, while ECM does not appear to perform quite as well as M2PM-2 in terms of χ^2 scores, it should be noted that it represents a much more parsimonious specification (k parameters instead of 2k parameters), and should in fairness be assessed

---

[35] Recalling the discussion around equation (17), this suggests a positive correlation over $\mathbf{x}$ of $A(\mathbf{x})$ and $\rho(\mathbf{x})$. Indeed, fitting $\rho(\mathbf{x})$ as an ECM model based on the exponentiated least-squares residuals from the regression of ln(y) on $\mathbf{x}$ for y>0 gives a full sample covariance of 0.197 between $\hat{\rho}(\mathbf{x})$ and $\hat{A}(\mathbf{x})$, quite close to the observed |MPE| of 0.175.

accordingly.

The CM tests for the individual elements of $q(\mathbf{x})$ (i.e. those corresponding to the $\mathbf{R}_{4j}$) are presented in table 6. These statistics provide an indication of the extent to which any particular orthogonality restriction built into the estimators may be questionable. The statistical significance might be judged by classical methods ($|t|>1.96$) or by the appropriate Schwartz criterion (given here by $|t|>3.24$).

Most striking is the fact that violations of orthogonality for the homoskedastic 2PM estimator are found in the vast majority of cases. Conversely, M2PM-2 appears to perform reasonably well, with significant t-statistics under the Schwartz criterion found in only two of the 35 cases. ECM again performs less well than M2PM-2. For both M2PM-2 and ECM, particular trouble spots would appear to be with respect to variables and interactions involving Male and Schooling. The OLS results again represent a middle ground; interestingly, the problems with the Male interactions seen for M2PM-2 and ECM do not seem quite so prominent with OLS.

VII. SUMMARY AND DISCUSSION

Both the algebraic and the empirical results presented here suggest that one should approach use of the standard (homoskedastic) 2PM with considerable caution in microeconometric applications where interest centers on $E[y|\mathbf{x}]$ and its associated partial effects. The basic identifying assumption for β in that model, namely $E[\varepsilon|y>0,\mathbf{x}]=0$, is not sufficiently powerful to identify other parameters of interest -- $E[y|\mathbf{x}]$, $\delta(\mathbf{x})$, etc. -- even if $\pi(\mathbf{x})$ is properly specified and identified. One may make

the stronger assumption that ε and $\mathbf{x}$ are statistically independent, but such an assumption is of little use -- and will actually tend to be counterproductive -- when $\rho(\mathbf{x})$ is in fact a nontrivial function of $\mathbf{x}$.

The requirements for identifying β_M as well as $E[y|\mathbf{x}]$ and $\delta(\mathbf{x})$ are far less stringent when the M2PM estimator is used. Again, a single orthogonality condition is the basis of the identification of β_M, namely $E[\exp(\varepsilon_M)|y>0,\mathbf{x}]=1$. In this case, however, and unlike the 2PM, this single restriction is also sufficient to identify $E[y|y>0,\mathbf{x}]$ and its associated partial effects. As such, unless there are some *a priori* bases on which the analyst can be comfortable with the assumption that ε and $\mathbf{x}$ are statistically independent (or at least that $\rho(\mathbf{x})$ is constant over $\mathbf{x}$), there would seem to be a clear preference for using M2PM over 2PM if parameters apart from β or β_M *per se* are of interest and if their consistent estimation is the prime objective. On the basis of the applied work undertaken here, there would seem to emerge a preference for using the two-step estimator M2PM-2 over the one-step version M2PM-1. Estimation of the former is only marginally more time-consuming than 2PM, requiring a logit or probit regression (as does 2PM) and a nonlinear least squares exponential regression, which in practice tends to converge rapidly.

The question of whether "truth" is really a one-part or a two-part model should be confronted squarely in applications. One simple specification test for M2PM against ECM was suggested here and, as suggested earlier, goodness-of-fit tests of various sorts might be contemplated as well. As a general premise, an ECM specification for $E[y|\mathbf{x}]$ captures perhaps the single most

prominent feature of $E[y|\mathbf{x}]$, namely $E[y|\mathbf{x}]>0$. Given a rich specification of the covariate vector $\mathbf{x}$ (e.g. low-order polynomials in and interactions between the main covariates), an ECM specification of $E[y|\mathbf{x}]$ with a linear index function may be sufficient to capture any important nonlinearities in parameters like the partial effects without recourse to a two-part structure.

Finally, contemplation of an estimation strategy should include considerations of some of the key rationales that motivated development of the 2PM in the first place. Perhaps most prominent here are considerations of the robustness of point estimates to thick upper tails and/or high-end outliers. The arguments advanced here suggest that there are likely to be some bias-robustness tradeoffs involved in such considerations, but assessing the nature of such tradeoffs is beyond the scope of this paper.

REFERENCES

Andrews, D.W.K., 1988. Chi-square diagnostic tests for econometric models: Theory. Econometrica 56, 1419-1453.

Cameron, A.C., Trivedi, P.K., 1996. The Analysis of Count Data. Book manuscript.

Carroll, R.J., Ruppert D., 1988. Transformation and Weighting in Regression. Chapman and Hall, London.

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829-844.

Cramer, H., 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ.

Davidson, R., MacKinnon, J.G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.

Duan, N., 1983. Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association 78, 605-610.

Duan, N. et al., 1983. A comparison of alternative models for the demand for medical care. Journal of Business and Economic Statistics 1, 115-126.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Eichner, M. et al., 1997. Health expenditure persistence and the feasibility of medical savings accounts. in J. Poterba, ed. Tax Policy and the Economy. Vol. 11. NBER, MIT Press, Cambridge, MA.

French, M.T., Zarkin, G.A., 1995. Is moderate alcohol use related to wages? Evidence from four worksites. Journal of Health Economics 14, 319-344.

Hay, J.W. et al., 1987. Ordinary least squares and sample-

selection models of health-care demand. Journal of Business and Economic Statistics 5, 499-506.

Hay, J.W., Olsen, R.J., 1984. Let them eat cake: A note on comparing alternative models of the demand for medical care. Journal of Business and Economic Statistics 2, 279-282.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153-161.

Ichimura, H., Lee, L.-F., 1991. Semiparametric least squares estimation of multiple index models: Single equation estimation. in W.A. Barnett et al., eds. Nonparametric and Semiparametric Methods in Econometrics and Finance, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge Univ. Press, Cambridge.

Johnson, N.L., Kotz, S., Kemp, A.W., 1992. Univariate Discrete Distributions, Second Edition. Wiley, New York.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-14.

Manning, W.G., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics, This Issue.

Manning, W.G. et al., 1987a. Monte Carlo evidence on the choice between sample selection and two-part models. Journal of Econometrics 35, 59-82.

Manning, W.G. et al., 1987b. Health insurance and the demand for medical care: Evidence from a randomized experiment. American Economic Review 77, 251-277.

Manning, W.G. et al., 1995. The demand for alcohol: The differential response to price. Journal of Health Economics 14, 123-48.

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Chapman and Hall, London.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics 33, 341-365.

Mullahy, J., 1997a. Instrumental variable estimation of count data models: Applications to models of cigarette smoking. Review of Economics and Statistics 79, 586-593.

Mullahy, J., 1997b. Economic aspects of childhood immunizations. In progress.

Newhouse, J.P. et al., 1980. On having your cake and eating it too: Econometric problems in estimating the demand for health services. Journal of Econometrics 13, 365-390.

Newhouse J.P. et al., 1987. The Findings of the Rand Health Insurance Experiment: A response to Welch et al. Medical Care 25, 157-179.

Pagan, A., Vella, F., 1989. Diagnostic tests for models based on individual data: A survey. Journal of Applied Econometrics 4, S29-S59.

Pohlmeier, W., Ulrich, V., 1995. An econometric model of the two-part decisionmaking process in the demand for health care. Journal of Human Resources 30, 339-361.

Schwartz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461-464.

Welch, B.L. et al., 1987. The Rand Health Insurance Study: A summary critique. Medical Care 25, 148-156.

Wooldridge, J.M., 1992. Some alternatives to the Box-Cox regression model. International Economic Review 33, 935-955.

Table 1

Descriptive Statistics: 1992 NHIS 12-Month Doctor Visits, (N=36,111)

Variable	Mean	Min	Max
Visits	4.91	0	370
Age	41.8	25	64
Schooling	13.0	0	18
Male	.378	0	1
White	.819	0	1
Married	.680	0	1
Excellent	.341	0	1
Very Good	.300	0	1
Good	.241	0	1

Table 2

Sample Frequency Distribution: Visits (N=36,111)

Visits	Sample Freq.	Sample Pct.
0	8513	23.57
1	8260	22.87
2	5155	14.28
3	3044	8.43
4	2312	6.40
5	1389	3.85
6	1369	3.79
7	493	1.37
8	617	1.71
9	197	0.55
10	848	2.35
11	108	0.30
12	1000	2.77
13	109	0.30
14	136	0.38
15	417	1.15
16	116	0.32
17	49	0.14
18	92	0.25
19	27	0.07
20+	1860	5.15

Table 3

Estimation Results: Alternative Estimators
(robust asymptotic t-statistics in parentheses)

	α		β, β_M			ζ	γ
Variable	Logit	M2PM-1	2PM	M2PM-1	M2PM-2	ECM	2PM[36]
Constant	.793	1.161	1.814	2.652	2.749	2.328	.871
	(8.6)	(0.5)	(42.5)	(4.8)	(21.2)	(17.1)	(11.5)
Age	.007	.024	-.003	-.012	-.009	-.007	-.004
	(5.9)	(1.9)	(4.8)	(2.0)	(4.9)	(3.7)	(3.8)
Male	-.913	-1.441	-.188	.109	-.047	-.199	.044
	(35.4)	(2.7)	(15.2)	(0.3)	(1.1)	(4.2)	(1.7)
White	.151	-.109	.154	.183	.132	.175	.014
	(4.4)	(0.3)	(9.7)	(2.2)	(2.5)	(3.2)	(0.4)
Schooling	.103	.038	.024	.049	.037	.055	.002
	(23.1)	(0.9)	(11.3)	(3.5)	(5.5)	(7.7)	(0.5)
Married	.120	.318	-.041	-.191	-.148	-.130	-.065
	(4.3)	(1.2)	(3.1)	(2.0)	(3.6)	(3.0)	(2.4)
Excellent	-1.394	-1.575	-1.304	-1.504	-1.612	-1.828	-.218
	(24.7)	(1.2)	(56.5)	(6.5)	(33.8)	(37.4)	(5.3)
Very Good	-1.056	-1.667	-1.083	-1.129	-1.340	-1.480	-.184
	(18.6)	(1.2)	(47.2)	(4.5)	(29.8)	(32.1)	(4.7)
Good	-.898	-1.171	-.752	-.776	-.856	-.973	-.062
	(15.8)	(0.9)	(32.1)	(5.4)	(19.4)	(21.7)	(1.5)
N. Obs.	36,111	36,111	27,598	36,111	27,598	36,111	27,598

[36] This estimator is the ECM estimator in which the exponentiated estimated residuals from the 2PM are the dependent variable. Asymptotic t-statistics for this estimator are based on standard heteroskedasticity-robust formulae but may still be misleading.

Table 4

Summary of Partial Effects $\partial E[y|\mathbf{x}]/\partial x_j$ for Age and Schooling:
25th, 50th, 75th Sample Percentiles and Sample Mean
(Computed on full sample, N=36,111)

	Estimator			
Variable	ECM	M2PM-2	M2PM-1	2PM
Age				
25th Pctl.	-.040	-.043	-.039	-.015
50th Pctl.	-.027	-.028	-.019	-.0081
75th Pctl.	-.020	-.019	-.0011	-.0025
Sample Mean	-.036	-.038	-.030	-.011
Schooling				
25th Pctl.	.15	.18	.17	.28
50th Pctl.	.20	.22	.23	.32
75th Pctl.	.30	.32	.32	.42
Sample Mean	.27	.28	.29	.38

Table 5

Conditional Moment χ^2 Tests for Misspecification of $E[y|\mathbf{x}]$, Mean Prediction Errors, and Mean Squared Errors

Test Statistic	2PM	M2PM-2	ECM	OLS
CM Test χ^2				
All Covariates (d.f.=35)	334.2	188.0	222.2	--
Linear Terms Only (d.f.=8)	105.5	32.3	62.0	--
Higher-Order Terms Only (d.f.=27)	281.9	132.6	197.4	193.2
Mean Prediction Error	-0.175	-0.009	0.010	0
Mean Squared Error	151.92	151.04	151.12	151.54

χ^2 Critical Values:

d.f.	.05	.01	.0001	Schwartz Criterion
35	49.8	57.3	74.9	367.3
27	40.1	47.0	63.2	283.3
8	15.5	20.1	31.8	84.0

Table 6

CM Tests for Misspecification of E[y|**x**]: t-Statistics

Variables	2PM	M2PM-2	ECM	OLS
Age	5.7	1.7	0.5	--
Schooling	7.3	-0.8	-2.8	--
Male	4.8	-2.6	-3.9	--
White	3.8	1.6	-0.1	--
Married	-0.3	1.1	0.6	--
Excellent	-6.9	2.5	-0.7	--
Very Good	-3.9	1.5	-0.8	--
Good	4.2	-0.7	-1.6	--
Age^2	3.5	2.0	1.1	0.8
$Schooling^2$	6.4	-0.5	-1.8	0.8
Age×Schooling	4.8	0.3	-0.8	-0.3
Age×Male	5.3	-0.8	-2.1	4.4
Age×White	2.0	1.4	0.1	-1.3
Age×Married	-2.1	0.2	-0.2	-4.0
Age×Excellent	-8.8	0.5	-1.3	0.6
Age×Very Good	-4.4	1.7	0.5	2.5
Age×Good	4.0	1.4	0.6	2.4
Schooling×Male	3.7	-3.7	-4.8	-3.0
Schooling×White	4.0	1.5	-0.0	1.4
Schooling×Married	0.1	1.2	0.7	1.8
Schooling×Excellent	-7.3	-0.0	-3.4	-6.5
Schooling×Very Good	-3.3	0.9	-1.3	-2.5
Schooling×Good	4.5	-0.1	-0.8	2.0
Male×White	2.9	-3.0	-4.2	-2.0
Male×Married	2.3	-1.1	-1.9	1.2
Male×Excellent	-3.3	-2.7	-4.7	0.8
Male×Very Good	-3.0	-3.4	-4.5	-1.4
Male×Good	2.6	-0.7	-1.5	-0.5
White×Married	-0.2	1.8	1.2	0.8
White×Excellent	-6.0	2.0	0.1	-2.0
White×Very Good	-3.6	1.0	-0.3	-1.2
White×Good	4.3	1.5	1.0	2.1
Married×Excellent	-4.9	2.6	1.9	3.3
Married×Very Good	-2.3	2.8	2.3	3.6
Married×Good	-0.4	-1.6	-1.8	-2.4

Fig. 1. Standard 2PM: exp(residual) and Schooling