Preliminary

UNOBSERVED HETEROGENEITY AND ESTIMATION OF AVERAGE PARTIAL EFFECTS

Jeffrey M. Wooldridge
Department of Economics
Michigan State University
Marshall Hall
East Lansing, MI 48824-1038
(517) 353-5972
WOOLDRI1@PILOT.MSU.EDU

July 2001

Key Words: Average Partial Effect, Average Treatment Effect, Endogeneity, Ignorability, Selection on Observables, Panel Data, Unobserved Heterogeneity

ABSTRACT

I study the problem of identifying average partial effects (APEs), which are partial effects averaged across the population distribution of unobserved heterogeneity, under various assumptions. One possibility is when the unobserved heterogeneity is conditionally independent of the observed covariates. When the unobserved heterogeneity is independent of the original covariates, or conditional mean independent but heteroskedastic, the derivations of APEs provide a new view of traditional specification problems in widely used models such as probit and Tobit. In addition, the focus on average partial effects resolves scaling issues that arise in estimating the parameters of probit and Tobit models with endogenous explanatory variables. I also show that of APEs in unobserved effects panel data models are identified under weaker assumptions than those typically used to identify the parameters of the models.

1. INTRODUCTION

Econometric models, especially at the individual, family, or firm level, often depend on unobserved heterogeneity in addition to observable covariates. Models with unobserved heterogeneity are often derived from economic theory, or specified based on introspection. The models can be specified for cross section data or panel data.

In nonlinear models, much has been made about the deleterious effects that ignoring heterogeneity can have on the estimation of parameters, even when the heterogeneity is assumed to be independent of the observed covariates. A leading case is the probit model with an omitted variable. Yatchew and Griliches (1985) show that when the omitted variable is independent of the explanatory variables and normally distributed, the probit estimators suffer from (asymptotic) attenuation bias. This result is often cited to illustrate how a misspecification that is innocuous in linear models leads to problems in nonlinear models. (See, for example, Greene, 2000, p. 828.) Discussions of bias caused by omitted heterogeneity in nonlinear models tend to focus on the effects of parameter estimates. The goal in this paper is to focus on partial effects, appropriately averaged across the unobserved heterogeneity.

Rather than assume that the unobserved heterogeneity is independent of the observed covariates in the structural conditional expectation, a weaker assumption is that the heterogeneity and covariates are independent conditional on a set of additional controls. In many cases, the additional controls can be viewed as proxy variables for the unobserved heterogeneity. (Sometimes the additional controls are referred to as "control functions," as in Heckman and Robb (1985), Blundell and Powell (2000).) Conditional independence between the heterogeneity and covariates of interest is often implicit in regression analyses which include many explanatory variables, and has been used in the treatment effect literature as well as the literature on nonlinear unobserved effects panel data models.

In this paper I show that the quantities of primary interest for empirical analysis -- partial effects averaged across the population distribution of any unobserved heterogeneity -- are identified by the conditional expectation of the observed response given all observed conditioning variables. While this result is a rather simple application of the law of iterated expectations, it has important practical implications. One special case is where the unobserved heterogeneity is assumed to be independent of the structural covariates -- a common assumption in random coefficient and mixture models. In such cases I show that we tend to focus too much on identification of parameters and not enough on identification of partial effects. The same is true when additional proxy variables are brought into the analysis.

The basic insights of this paper have been used by others. Chamberlain's (1984) method of estimating average partial effects in the context of his random effects probit model is a special case. Angrist (1991) derives the average treatment effect in a probit model with a single binary, endogenous explanatory variable. Blundell and Powell (2000) study semiparametric estimation of a class of nonlinear models with endogenous explanatory variables. I have used the basic results many times in writing my forthcoming book on cross section and panel data methods (Wooldridge (2001)). Here, I hope to give a systematic treatment that helps to unify the discussion of estimating partial effects. I also provide some new, fairly complicated examples of where average partial effects can be identified.

The rest of the paper is organized as follows. Section 2 gives the basic result and a useful corollary, and Section 3 shows how the results apply to some commonly used parametric models for cross section data, including some extensions. Section 4 discusses models of the kind studied by Smith and Blundell (1986), Rivers and Vuong (1988), and Blundell and Powell (2000), among others. I derive scale adjustments that turn estimates from simple two-step procedures into estimates of the parameters indexing the average partial effects, and I show how heteroskedasticity in the unobserved heterogeneity is easily handled. In Section 5 I show that the results from Section 2 have important implications for estimating the partial effects in unobserved effects panel data models, including nonlinear models without full distributional assumptions, where some explanatory variables may be strictly exogenous conditional on unobserved heterogeneity, while others are not. Section 6 contains some caveats and concluding remarks.

2. THE FRAMEWORK AND BASIC RESULTS

It is useful to have a framework for unifying the discussion of partial effects in models with unobserved heterogeneity. Let y be an observed scalar response, let $\mathbf{x}$ be a k-vector of observed explanatory variables taking values in $\mathcal{X}$, and let $\mathbf{q}$ denote an m-vector of unobserved heterogeneity taking values in $\mathcal{Q}$. The random vector $(y,\mathbf{x},\mathbf{q})$ represents the underlying population. (In many examples, $\mathbf{q}$ is a scalar, but the general setup is no easier in that case.)

The structural expectation of interest is

$$E(y|\mathbf{x},\mathbf{q}) \equiv \mu_1(\mathbf{x},\mathbf{q}), \tag{2.1}$$

where $\mu_1(\mathbf{x},\mathbf{q})$ is the conditional mean function. Because $\mathbf{x}$ is the vector of observed covariates in (2.1), as a shorthand we call $\mathbf{x}$ the "structural covariates." Often, we would model μ_1 parametrically, but that is not required at this level. We are interested in how the x_j affect $E(y|\mathbf{x},\mathbf{q})$. When x_j is continuous the partial effect is

$$\partial E(y|\mathbf{x},\mathbf{q})/\partial x_j, \tag{2.2}$$

which, for small changes in x_j, can be multiplied by Δx_j to obtain the approximate change in $E(y|\mathbf{x},\mathbf{q})$, holding the other elements in $\mathbf{x}$ and $\mathbf{q}$ fixed. (Sometimes we are interested in some function of (2.2), such as an elasticity or semi-elasticity.) If x_j is discrete, we can instead look at the difference in $E(y|\mathbf{x},\mathbf{q})$ for two different x_j values (such as one and zero when x_j is binary.) We focus on (2.2) for concreteness, but all of the following discussion applies to partial effects based on differences in $E(y|\mathbf{x},\mathbf{q})$ for different values of $\mathbf{x}$.

Unless $E(y|\mathbf{x},\mathbf{q})$ is separable in $\mathbf{x}$ and $\mathbf{q}$, the partial effect in (2.2) depends on $\mathbf{q}$ as well as $\mathbf{x}$. While we can plug in interesting values for the x_j -- such as sample averages, medians, quartiles, minimums, maximums, and values representing policy changes -- it is less clear what to do about the dependence of (2.2) on $\mathbf{q}$. In parametric models, the heterogeneity usually can be normalized to have a zero mean without loss of generality. Sometimes a different normalization is more convenient, but for now we assume $E(\mathbf{q}) = \mathbf{0}$. Then, we might estimate the effect of x_j by evaluating (2.2) at $\mathbf{q} = \mathbf{0}$.

If $\mathbf{q}$ has a continuous distribution, setting $\mathbf{q} = 0$ may be representative of only a small fraction of the population, a point made by Chamberlain

(1984, p. 1273) in the context of unobserved effects panel data models. [Technically, $P(\mathbf{q} = 0) = 0$ when $\mathbf{q}$ is continuous.] An alternative is to average the partial effect over the distribution of $\mathbf{q}$; in effect, we "integrate out" $\mathbf{q}$ in (2.2). This leads to the *average partial effect* (*APE*) of x_j. To define the APE, either let $\theta_j(\mathbf{x},\mathbf{q})$ be the partial derivative in (2.2) or, for discrete changes, define $\theta_j(\mathbf{x},\mathbf{q})$ as the difference in $E(y|\mathbf{x},\mathbf{q})$ at two different values of x_j, holding $\mathbf{q}$ and the other elements of $\mathbf{x}$ fixed.) The APE of x_j, evaluated at $\mathbf{x}^\circ$, is

$$\delta_j(\mathbf{x}^\circ) \equiv \int_{\mathcal{Q}} \theta_j(\mathbf{x}^\circ,\mathcal{q}) f(\mathcal{q})\eta(d\mathcal{q}) \equiv E_{\mathbf{q}}[\theta_j(\mathbf{x}^\circ,\mathbf{q})], \qquad (2.3)$$

where $f(\cdot)$ is the density of $\mathbf{q}$, which we take to be absolutely continuous with respect to the σ-finite measure, $\eta(\cdot)$. If $\mathbf{q}$ is continuous, $\eta(\cdot)$ can be taken to be Lebesgue measure; if $\mathbf{q}$ is discrete, $\eta(\cdot)$ is counting measure, in which case the integral is a weighted average. These are the leading cases, but heterogeneity distributions with $\mathbf{q}$ neither continuous nor discrete are allowed. We use script variables as dummy arguments in the integration, and $E_{\mathbf{q}}[\cdot]$ denotes expectation with respect to the distribution of $\mathbf{q}$. Given $\mu_1(\cdot,\cdot)$, $f(\cdot)$, and, in the case of (2.2), the assumption that the derivative and integral can be interchanged, $\delta_j(\mathbf{x}^\circ)$ can be computed as a function of $\mathbf{x}^\circ$.

The definition of an average partial effect is implicit in Chamberlain's (1984, p. 1273) treatment of unobserved effects probit panel data models. Blundell and Powell (2000) called $E_{\mathbf{q}}[\mu_1(\mathbf{x}^\circ,\mathbf{q})]$ the average structural function. By taken differences of the ASF or derivatives (and assuming the derivative and expectation can be interchanged) we arrive exactly at the APEs. For the remainder of the paper, I assume that partial derivatives and expectations can be interchanged without saying so explicitly. The assumptions under which this interchange is allowed are quite weak; see, for example, Bartle (1966, Corollary 5.9).

In many applications, $\mu_1(\mathbf{x},\mathbf{q})$ is specified parametrically. But this does not guarantee identification of the APEs, for two reasons: (1) The relevant parameters in $\mu_1(\mathbf{x},\mathbf{q})$ may not be identified; (2) The density of $\mathbf{q}$ might not be known. Often, but certainly not always, the APEs, are identified if we specify a conditional distribution of $\mathbf{q}$ given some observed covariates, $\mathbf{w}$. The following lemma follows by the law of iterated expectations.

LEMMA 2.1: For an h-vector $\mathbf{w}$ taking values in $\mathcal{W}$, let $g(\cdot|\mathbf{w})$ the conditional density of $\mathbf{q}$ given $\mathbf{w}$ with respect to a σ-finite measure, $\eta(\cdot)$. For any $\mathbf{x}^{\circ} \in \mathcal{X}$ and $\mathbf{w}^{\circ} \in \mathcal{W}$, define

$$\mu_2(\mathbf{x}^{\circ},\mathbf{w}^{\circ}) \equiv \int_{\mathcal{Q}} \mu_1(\mathbf{x}^{\circ},\mathfrak{q})g(\mathfrak{q}|\mathbf{w}^{\circ})\eta(d\mathfrak{q}). \tag{2.4}$$

Then

$$\mathrm{E}_{\mathbf{q}}[\mu_1(\mathbf{x}^{\circ},\mathbf{q})] = \mathrm{E}_{\mathbf{w}}[\mu_2(\mathbf{x}^{\circ},\mathbf{w})], \tag{2.5}$$

and so APEs can be obtained by taking derivates or changes of $\mathrm{E}_{\mathbf{w}}[\mu_2(\mathbf{x},\mathbf{w})]$ with respect to the elements of $\mathbf{x}$ and inserting $\mathbf{x}^{\circ}$ for $\mathbf{x}$. ■

Because (2.4) can be written as $\mu_2(\mathbf{x}^{\circ},\mathbf{w}^{\circ}) = \mathrm{E}[\mu_1(\mathbf{x}^{\circ},\mathbf{q})|\mathbf{w} = \mathbf{w}^{\circ}]$, (2.5) follows directly by iterated expectations. At this point, $\mu_2(\mathbf{x}^{\circ},\mathbf{w}^{\circ})$ is not necessarily $\mathrm{E}(y|\mathbf{x} = \mathbf{x}^{\circ},\mathbf{w} = \mathbf{w}^{\circ})$.

Whether Lemma 2.1 is useful hinges on whether $\mu_2(\cdot,\cdot)$ is identified, and this depends, loosely, on whether enough features of $\mu_1(\cdot,\cdot)$ and $g(\cdot|\cdot)$ are identified. If $\mu_2(\cdot,\cdot)$ is identified, then the right hand side of (2.5) can be consistently estimated by averaging $\hat{\mu}_2(\mathbf{x}^{\circ},\mathbf{w}_i)$ across the random sample, $i = 1,\ldots,N$, where $\hat{\mu}_2$ is a consistent estimator of $\mu_2(\cdot,\cdot)$. In particular, a consistent estimator of $\hat{\delta}_j(\mathbf{x}^{\circ})$ when x_j is continuous is simply

$$\hat{\delta}_j(\mathbf{x}^{\circ}) = N^{-1}\sum_{i=1}^{N}\partial\hat{\mu}_2(\mathbf{x}^{\circ},\mathbf{w}_i)/\partial x_j. \qquad (2.6)$$

When x_j is discrete, or we simply want to estimate the discrete difference at two different values, say $x_j^{(1)}$ and $x_j^{(0)}$, we use

$$\hat{\delta}_j(\mathbf{x}^{\circ}) = N^{-1}\sum_{i=1}^{N}[\hat{\mu}_2(x_j^{(1)},\mathbf{x}_{(j)}^{\circ},\mathbf{w}_i) - \hat{\mu}_2(x_j^{(0)},\mathbf{x}_{(j)}^{\circ},\mathbf{w}_i)], \qquad (2.7)$$

where $\mathbf{x}_{(j)}^{\circ}$ denotes fixed values of the other elements of $\mathbf{x}$.

Nothing prevents us from choosing $\mathbf{w} = \mathbf{x}$ in stating Lemma 2.1. However, not surprisingly, allowing $\mathbf{q}$ and $\mathbf{x}$ to be dependent does not generally allow us to identify the APEs, unless we have more information. (Section 3 covers probit and Tobit models when the heterogeneity is heteroskedastic conditional on $\mathbf{x}$, and shows that the APEs are identified.) In Section 5 we discuss how the APEs are generally identified in panel data models where $\mathbf{x}$ becomes $\mathbf{x}_t$ and $\mathbf{w} = (\mathbf{x}_1,\ldots,\mathbf{x}_T)$, the entire observed history of the covariate process. In some interesting cases, the APEs are identified even though the separate parameters in $\mu_1(\cdot,\cdot)$ and $g(\cdot|\cdot)$ might not be. Wooldridge (2000b) uses Lemma 2.1 to identify APEs in nonlinear, dynamic unobserved effects models once a distribution for the unobserved heterogeneity, given the initial conditional and a set of strictly exogenous covariates, has been specified.

In many cases, we want to apply Lemma 2.1 when $\mathbf{w}$ and $\mathbf{x}$ can vary freely from one another. Provided we make appropriate assumptions about the structural expectation and the conditional distribution of heterogeneity, we can identify the APEs by identifying $\mathrm{E}(y|\mathbf{x},\mathbf{w})$. The first assumption is that $\mathbf{w}$ is appropriately excluded from the structural mean.

ASSUMPTION A.1: The h-vector $\mathbf{w}$ is redundant, or ignorable, in (2.1):

$$E(y|\mathbf{x},\mathbf{q},\mathbf{w}) = E(y|\mathbf{x},\mathbf{q}). \quad \blacksquare \tag{2.8}$$

In pure cross section settings, Assumption A.1 is not especially controversial because any element in $\mathbf{w}$ that we think belongs in the structural expectation should be included in $\mathbf{x}$ as well. Assumption A.1 simply allows for the fact that we have observed covariates that are assumed to be properly omitted from a structural equation that contains unobserved heterogeneity. For example, suppose that y is earnings, $\mathbf{x}$ is schooling, and $\mathbf{q}$ is innate ability. Suppose that $\mathbf{w}$ is the outcome on a cognitive test, say an IQ test. Then (2.8) simply means that *if* we could control for "ability," IQ would not be helpful in explaining average earnings. Essentially, this exclusion restriction holds by definition of "ability."

In Section 5 we cover panel data applications, where we replace $(y,\mathbf{x})$ with $(y_t,\mathbf{x}_t)$, the draw from a single time period, and $\mathbf{q}$ is time-constant heterogeneity. We take the vector of controls to be the time average of the covariates, $(\mathbf{x}_1 + \mathbf{x}_2 + ... + \mathbf{x}_T)/T$. Assumption A.1 then implies strict exogeneity of $\{\mathbf{x}_t\}$ conditional on $\mathbf{q}$. (It is also possible to include other time-constant covariates in $\mathbf{w}$, as we discuss in Section 5.)

The second assumption is

ASSUMPTION A.2: Conditional on $\mathbf{w}$, $\mathbf{q}$ and $\mathbf{x}$ are independent: $D(\mathbf{q}|\mathbf{x},\mathbf{w}) = D(\mathbf{q}|\mathbf{w})$, where $D(\cdot|\cdot)$ denotes conditional distribution. $\blacksquare$

In a cross section environment, Assumption A.2 is typically much more restrictive than Assumption A.1, and Assumption A.2 is what makes $\mathbf{w}$ a vector of proxy or control variables rather than a set of instrumental variables. If $\mathbf{w}$ is closely related to $\mathbf{q}$, Assumption A.2 is often reasonable. As we will see in Section 4, it is often reasonable to make Assumption A.2 when $\mathbf{w}$ is a set of reduced-form disturbances and $\mathbf{x}$ contains endogenous explanatory variables correlated with $\mathbf{q}$. For panel data applications when $\mathbf{w}$ is the time average of the $\mathbf{x}_t$, $\bar{\mathbf{x}}$, Assumption A.2 requires that the distribution of the heterogeneity, given $\mathbf{x}_t$ and $\bar{\mathbf{x}}$, depends only on $\bar{\mathbf{x}}$.

Assumption A.2 has been called "selection on observables" in the econometrics literature because $\mathbf{x}$ is not dependent on $\mathbf{q}$ once we control for $\mathbf{w}$. (See, for example, Heckman and Robb (1985).) In the treatment effect literature, Assumption A.2 is the "igorability of treatment" assumption, where $\mathbf{x}$ is the vector of treatments and $\mathbf{q}$ contains the counterfactual outcomes on y.

For special forms of $\mu_1(\mathbf{x},\mathbf{q})$, Assumption A.2 can be relaxed; sometimes a conditional mean independence assumption suffices. But to handle general nonlinear models we need full conditional independence. The following lemma is simple but fundamental.

LEMMA 2.2: Define $\mu_2(\mathbf{x}^\circ,\mathbf{w}^\circ)$ as in equation (2.4). Then, under Assumptions A.1 and A.2,

$$\mu_2(\mathbf{x}^\circ,\mathbf{w}^\circ) = \mathrm{E}(y|\mathbf{x} = \mathbf{x}^\circ,\mathbf{w} = \mathbf{w}^\circ) \qquad (2.9)$$

for $\mathbf{x}^\circ \in \mathcal{X}$, $\mathbf{w}^\circ \in \mathcal{W}$. Therefore, if $\mathrm{E}(y|\mathbf{x},\mathbf{w})$ is identified, so are the average partial effects.

PROOF: The law of iterated expectations implies that

$$\mathrm{E}(y|\mathbf{x},\mathbf{w}) = \mathrm{E}[\mathrm{E}(y|\mathbf{x},\mathbf{q},\mathbf{w})|\mathbf{x},\mathbf{w}]$$

$$= \mathrm{E}[\mu_1(\mathbf{x},\mathbf{q})|\mathbf{x},\mathbf{w}] \tag{2.10}$$

$$= \int_Q \mu_1(\mathbf{x},q)\, g(q|\mathbf{w})\, \eta(\mathrm{d}q) \tag{2.11}$$

where (2.10) follows from Assumption A.1 [$\mathrm{E}(y|\mathbf{x},\mathbf{q},\mathbf{w})$ does not depend on $\mathbf{w}$] and (2.11) follows from A.2 [$\mathrm{D}(\mathbf{q}|\mathbf{x},\mathbf{w})$ does not depend on $\mathbf{x}$.] But (2.11) with $\mathbf{x} = \mathbf{x}^{\circ}$ and $\mathbf{w} = \mathbf{w}^{\circ}$ is simply (2.4). ■

If $\mathrm{E}(y|\mathbf{x},\mathbf{w})$ is identified, usually it can be consistently estimated given a random sample on $(y,\mathbf{x},\mathbf{w})$ from the population, which means that we have a consistent estimator $\hat{\mu}_2(\mathbf{x},\mathbf{w})$ of $\mu_2(\mathbf{x},\mathbf{w})$ for all $\mathbf{x}$ and $\mathbf{w}$. Then we can estimate $\delta_j(\mathbf{x}^{\circ})$ generally as in (2.6) or (2.7). In some of the examples we cover later, we will have the distribution of $\mathbf{w}$, and so we can compute $\mathrm{E}_{\mathbf{w}}[\mu_2(\mathbf{x}^{\circ},\mathbf{w})]$ as a function of parameters and then take derivatives or changes with respect to the $\mathbf{x}$ elements.

One way to view Lemma 2.1 is that it says, under ignorability assumptions when $\mathrm{E}(y|\mathbf{x},\mathbf{w})$ is identified, we should simply estimate $\mathrm{E}(y|\mathbf{x},\mathbf{w})$, compute the partial derivative or discrete change of interest, plug in values for $\mathbf{x}$, and average across the population distribution of $\mathbf{w}$ or the sample of $\mathbf{w}_i$. We need never explicitly account for unobserved heterogeneity -- all quantities of interest are obtained from $\mathrm{E}(y|\mathbf{x},\mathbf{w})$.

Those outside of econometrics, particularly those in applied statistics, might wonder what all the fuss is about unobserved heterogeneity, especially in cross-sectional contexts. Given Lemma 2.2, the value of introducing unobserved heterogeneity explicitly can be questioned, at least under the ignorability assumptions. For example, the treatment effect literature is based on estimating expectations of the form $\mathrm{E}(y|x,\mathbf{w})$, where x is a binary treatment and $\mathbf{w}$ is a set of controls that might be needed to control for self-selection into treatment. The conclusion of Lemma 2.2 is simply to estimate $E(y|x,\mathbf{w})$ for $x = 1$ and $x = 0$, and then average the difference across $\mathbf{w}_i$ in the sample. The resulting estimator is a well-known estimate of the average treatment effect. (See, for example, Heckman, Ichimura, and Todd (1997) and Hahn (1998).)

The claim that it is sufficient to focus on $E(y|\mathbf{x},\mathbf{w})$, without any reference to $E(y|\mathbf{x},\mathbf{q})$, may be objectionable to economists and other social scientists who often prefer to estimate so-called "structural" parameters in models with unobserved heterogeneity. Indeed, when Lemma 2.2 does not hold, we cannot recover estimates of the average partial effect from an estimate of $E(y|\mathbf{x},\mathbf{w})$. Then, models for $E(y|\mathbf{x},\mathbf{q})$ and $D(\mathbf{q}|\mathbf{w})$ are generally important.

In the next section I show how Lemmas 2.1 and 2.2 can be applied to several common parametric models with unobserved effects. The basic message is that if we insist on specifying models of $E(y|\mathbf{x},\mathbf{q})$, we must nevertheless be careful in interpreting the parameters, and in making claims about "biases" caused by unobserved heterogeneity.

As a corollary to Lemma 2.2, we have the case when $\mathbf{q}$ is independent of $\mathbf{x}$, which is a common assumption in nonlinear models with unobserved heterogeneity. We replace Assumption A.2 with

ASSUMPTION A.2′: $\mathbf{q}$ is independent of $\mathbf{x}$.

COROLLARY 2.1: Under Assumption A.2′,

$$E_{\mathbf{q}}[\mu_1(\mathbf{x}^{\circ},\mathbf{q})] = E(y|\mathbf{x} = \mathbf{x}^{\circ}).$$

Therefore, average partial effects are obtained directly from $E(y|\mathbf{x})$. ∎

This corollary is immediate by taking $\mathbf{w}$ to be empty in Lemma 2.2. Note that Assumption A.1 has no content when $\mathbf{q}$ and $\mathbf{x}$ are independent because A.1 simply defines $\mu_1(\mathbf{x},\mathbf{q}) = \mathrm{E}(y|\mathbf{x},\mathbf{q})$.

Many cross-sectional models, and panel data models, assume that unobserved heterogeneity is independent of the observed covariates. A large class of mixture models for all kinds of responses typically assume that the heterogeneity is independent of the covariates. One way to interpret Corollary 2.1 is that complicated ways of modeling unobserved heterogeneity is largely a waste of time: if we are just going to assume that the heterogeneity is indepedent of the covariates, and we are interested in average partial effects on mean responses, it suffices to focus on flexible ways of modeling $\mathrm{E}(y|\mathbf{x})$ and forget about $\mathbf{q}$.

3. APPLICATION TO SOME POPULAR CROSS-SECTIONAL PARAMETRIC MODELS

I now provide several examples that illustrate the usefulness of the simple results from Section 2. Throughout this section the unobserved heterogeneity, q, is a scalar.

3.1. Random Coefficient Model

Let x be a scalar and let $\mathbf{w}$ be a $1 \times h$ vector. Consider a structural model linear in parameters, but where the structural covariate of interest, x, interacts with unobserved heterogeneity:

$$\mathrm{E}(y|x,q,\mathbf{w}) = \eta + \beta x + q + \gamma x \cdot q. \quad (3.1)$$

Without loss of generality, the coefficient on q is normalized at unity and $E(q) = 0$. The partial effect of x on $E(y|x,q)$ is $\partial E(y|x,q)/\partial x = \beta + \gamma q$. (It is the dependence of the partial effect of x on the unobserved random variable, q, that makes (3.1) a random coefficient model.) Therefore, the APE of x is $\beta + \gamma E(q) = \beta$. (Note that β is the APE of x regardless of any correlation between q and x.) In addition, we assume that

$$E(q|x,\mathbf{w}) = E(q|\mathbf{w}), \tag{3.2}$$

-- which is a conditional mean independence version of Assumption A.2. Under (3.1) and (3.2),

$$\begin{aligned} E(y|x,\mathbf{w}) &= \alpha + \beta x + E(q|\mathbf{w}) + \gamma x \cdot E(q|\mathbf{w}) \\ &= \alpha + \beta x + \psi(\mathbf{w}) + \gamma x \psi(\mathbf{w}), \end{aligned} \tag{3.3}$$

where $\psi(\mathbf{w}) \equiv E(q|\mathbf{w})$. If we knew $\psi(\cdot)$ we could estimate β from a regression that includes x, $\psi(\mathbf{w})$, and the interaction between x and $\psi(\mathbf{w})$. In the linear case, $\psi(\mathbf{w}) = (\mathbf{w} - \boldsymbol{\xi})\boldsymbol{\delta}$ where $\boldsymbol{\xi} = E(\mathbf{w})$. (Demeaning $\mathbf{w}$ ensures that $E[\psi(\mathbf{w})] = 0$, which must hold because $E(q) = 0$.) In practice, we would replace $\boldsymbol{\xi}$ with $\bar{\mathbf{w}}$, the sample mean, and run the OLS regression

$$y_i \text{ on } 1,\ x_i,\ \mathbf{w}_i,\ x_i(\mathbf{w}_i - \bar{\mathbf{w}}),\ i = 1,\ldots,N; \tag{3.4}$$

the coefficient on x_i, say $\hat{\beta}$, is a consistent, asymptotically normal estimator of β.

Equation (3.3) suggests the possibility that β can be estimated using a semiparametric method, that is, without imposing a parametric form on $\psi(\cdot)$. Robinson (1988) considers the case without the interaction term. It would be useful to extend Robinson's method to (3.3).

When x is a binary treatment, and we explicitly consider the counterfactual framework in Rosenbaum and Rubin (1983) and many others since, we are led immediately to a standard estimator of the average treatment effect. Estimate $E(y|x = 1,\mathbf{w})$ and $E(y|x = 0,\mathbf{w})$ by flexible methods. If $\hat{m}_1(\mathbf{w})$ and $\hat{m}_0(\mathbf{w})$ denote such estimates. Then the average treatment effect is estimated as $N^{-1}\sum_{i=1}^{N}[\hat{m}_1(\mathbf{w}_i) - \hat{m}_0(\mathbf{w}_i)]$.

3.2. Probit Model

We now consider the structural probit model

$$P(y = 1|\mathbf{x},q,\mathbf{w}) = \Phi(\alpha + \mathbf{x}\boldsymbol{\beta} + q) = E(y|\mathbf{x},q), \tag{3.5}$$

where Φ is the standard normal cumulative distribution function and the coefficient on q is normalized at unity without loss of generality. Yatchew and Griliches (1985) assume that q and $\mathbf{x}$ are independent where q has a normal distributed with zero mean. We allow for the case that independence between q and $\mathbf{x}$ holds only after conditioning on some additional variables, $\mathbf{w}$:

$$q|\mathbf{x},\mathbf{w} \sim \text{Normal}(\eta + \mathbf{w}\boldsymbol{\delta},\tau^2). \tag{3.6}$$

Because Assumptions A.1 and A.2 are satisfied, we can use Lemma 2.1 to compute the APE of a continuous each x_j. This requires first finding $E(y|\mathbf{x},\mathbf{w}) = P(y = 1|\mathbf{x},\mathbf{w})$. We can show this follows a probit model by writing $y = 1[\lambda + \mathbf{x}\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\delta} + a + u \geq 0]$, where $\lambda = \alpha + \eta$, $a = q - (\eta + \mathbf{w}\boldsymbol{\delta})$, and $u|(\mathbf{x},\mathbf{w},a) \sim \text{Normal}(0,1)$. Because $a|(\mathbf{x},\mathbf{w}) \sim \text{Normal}(0,\tau^2)$, it follows that $a + u$ is independent of $(\mathbf{x},\mathbf{w})$ and distributed as $\text{Normal}(0,\tau^2 + 1)$ ($\text{Var}(a + u) = \text{Var}(q) + \text{Var}(u) = \tau^2 + 1$). Therefore

$$P(y = 1|\mathbf{x},\mathbf{w}) = \Phi[(\lambda + \mathbf{x}\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\delta})/(\tau^2 + 1)^{1/2}]. \tag{3.7}$$

It follows immediately that the APE of x_j is the partial derivative of (3.7):

$$\theta_j\phi(\mathbf{z}\boldsymbol{\theta}), \tag{3.8}$$

where $\phi(\cdot)$ is the standard normal density function, $\mathbf{z} = (1,\mathbf{x},\mathbf{w})$, and $\boldsymbol{\theta} = (\lambda,\boldsymbol{\beta}',\boldsymbol{\delta}')'/(\tau^2 + 1)^{1/2}$.

From (3.7) we can read off the plims from a probit of y on 1, $\mathbf{x}$, $\mathbf{w}$. In particular, if $\hat{\theta}_j$ is the coefficient on x_j, then

$$\text{plim } \hat{\theta}_j = \beta_j/(\tau^2 + 1)^{1/2} = \theta_j. \tag{3.9}$$

Equation (3.9) implies an attenuation bias in estimating the β_j. This is true even when q and $\mathbf{x}$ are assumed to be independent, and $\mathbf{w}$ is not available, a fact that has been cited as a serious problem in omitting unobserved heterogeneity in probit models. But why are the β_j of interest? Rarely do the magnitudes of the parameters in the underlying latent variable model have meaning; only the effects of the x_j on the response probability have quantitative meaning. For obtaining the directions and relative magnitudes of the effects, the $\beta_j/(\tau^2 + 1)^{1/2}$ are just as informative as the β_j.

One reason to be interested in the β_j is that they appear directly in the structural partial effects evaluated at $q = 0$:

$$\partial \mathrm{P}(y = 1|\mathbf{x}, q = 0, \mathbf{w})/\partial x_j = \beta_j \phi(\alpha + \mathbf{x}\boldsymbol{\beta}). \tag{3.10}$$

However, as we discussed in Section 2, (3.10) applies only to a small part of the population, and plugging in any other value of q is arbitrary, as q rarely has known units of measurement. But there is another subtle point about attenuation bias that has gone unnoticed. Even if we claim we are interested in (3.10), it is not obvious that unobserved heterogeneity causes an attenuation bias. This is easiest to see when q and $\mathbf{x}$ are independent, in which case we drop $\mathbf{w}$. The average partial effect, which we consistently estimate from the probit of y on $(1, \mathbf{x})$, is

$$[\beta_j/(\tau^2 + 1)^{1/2}]\phi[(\alpha + \mathbf{x}\boldsymbol{\beta})/(\tau^2 + 1)^{1/2}]. \tag{3.11}$$

Now, while the first term is attenuated toward zero, the "bias" works in the opposite direction in the second term because $\phi[\cdot]$ is symmetric about zero and reaches its maximum at zero, and $|(\alpha + \mathbf{x}\boldsymbol{\beta})/(\tau^2 + 1)^{1/2}| < |\alpha + \mathbf{x}\boldsymbol{\beta}|$.

Without knowing α, $\boldsymbol{\beta}$, τ^2, and a particular value of $\mathbf{x}$, we cannot know which is greater in magnitude, (3.10) or (3.11).

A related point involves comparing estimated coefficients on the covariates of interest, $\mathbf{x}$, across studies that use different proxy variable controls. Let $\mathbf{w}_1$ and $\mathbf{w}_2$ denote different proxy variables such that (3.5) and (3.6) are both assumed when $\mathbf{w}$ is replaced by $\mathbf{w}_1$ or $\mathbf{w}_2$ (with appropriate subscripts in (3.6)). Then, for the coefficient on x_j, the probability limits of the MLEs in the two cases are $\beta_j/(\tau_1^2 + 1)^{1/2}$ and $\beta_j/(\tau_2^2 + 1)^{1/2}$, respectively. While it makes sense to compare signs and relative effects, it makes no sense to argue that controlling for $\mathbf{w}_1$ versus $\mathbf{w}_2$ makes the effect of x_j more or less important, unless one estimates the average partial effects in the two cases.

My view is that the so-called attenuation bias from neglected heterogeneity, or even after we include some proxies, is a *good* thing: we want to estimate $\boldsymbol{\beta}/(\tau^2 + 1)^{1/2}$ because this is exactly what appears in the average partial effects, (3.8). The same is true if we look at discrete changes in x_j.

In cases where β_j is clearly the quantity of interest -- where the probit model arises from data censoring -- it turns out that ignoring the heterogeneity *still* does not have any deleterious effects. A leading case is estimating willingness-to-pay functions. Suppose we assume that

$$y^* = \alpha + \mathbf{x}\boldsymbol{\beta} + q + u$$

$$u|\mathbf{x},q,\mathbf{w} \sim \text{Normal}(0,\sigma^2),$$

where y^* is the unobserved willingness to pay. We are interested in $\text{E}(y^*|\mathbf{x},q)$. We also assume (3.6). Then

$$y^* = \lambda + \mathbf{x}\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\delta} + a + u \equiv \mathbf{z}\boldsymbol{\theta} + v,$$

where $\mathbf{z} = (1,\mathbf{x},\mathbf{w})$ and $v = a + u$. Let i denote a random draw from the population. Each individual i is presented with a cost, c_i, of the project or good; c_i is independent of all other factors, or, at least, of a_i and u_i. Individual i approves of the project if $y_i^* \geq c_i$. Let $y_i = 1[y_i^* \geq c_i]$ denote the observed binary response (accept or reject). Then

$$
\begin{aligned}
P(y_i = 1|\mathbf{x}_i,\mathbf{w}_i,c_i) &= P(\mathbf{z}_i\boldsymbol{\theta} + v_i \geq c_i|\mathbf{z}_i,c_i) \\
&= P(v_i \geq -\mathbf{z}_i\boldsymbol{\theta} + c_i) \\
&= P[v_i/\omega \geq -\mathbf{z}_i(\boldsymbol{\theta}/\omega) + (1/\omega)c_i] \\
&= \Phi[\mathbf{z}_i(\boldsymbol{\theta}/\omega) - (1/\omega)c_i], \qquad (3.12)
\end{aligned}
$$

where $\omega^2 = \tau^2 + \sigma^2 = \mathrm{Var}(v_i)$. Equation (3.12) shows that $\boldsymbol{\theta}$ and ω are identified and can be consistently estimated by maximum likelihood. In particular, the β_j are identified. While τ^2 and σ^2 are not separately identified, this is irrelevant for estimating the partial effects of interest, β_j.

The derivations in this example show that there is really no reason to fret about the effects of unobserved heterogeneity in probit models when the heterogeneity and structural covariates satisfy a conditional independence assumption, and the heterogeneity is conditionally normally distributed. (A corollary is that q and $\mathbf{x}$ are independent, we need not worry about the effects of unobserved heterogeneity, period.) If $q|\mathbf{w}$ is not normally distributed, then $P(y = 1|\mathbf{x},\mathbf{w})$ does not have a simple form. Therefore, probit of y on $(1,\mathbf{x},\mathbf{w})$ would not consistently estimate the APEs, or anything of interest. Or, if we begin with a logit structural model, logit of y on $(1,\mathbf{x},\mathbf{w})$ would not estimate the APE. But this does not change the basic point: Since the APEs are computed from $P(y = 1|\mathbf{x},\mathbf{w})$, why not model this probability directly? It is not very difficult to specify and estimate models that are more flexible than probit and logit without being wedded to a particular form of unobserved heterogeneity.

If q and $\mathbf{x}$ cannot be made conditionally independent, things are much different. Still, under a certain (restrictive) assumptions about the dependence between q and $\mathbf{x}$, the APEs can be identified by applying Lemma 2.1. We assume

$$\mathrm{E}(y|\mathbf{x},q) = \Phi(\alpha + \mathbf{x}\boldsymbol{\beta} + q), \tag{3.13}$$

where q has zero mean and variance τ^2. If q is independent of $\mathbf{x}$ and normally distributed, we just saw that probit of y on 1, $\mathbf{x}$ consistently estimates the average partial effects. If x_j is continuous, the APE of x_j at $\mathbf{x}^o$ is

$$\beta_j \mathrm{E}_q[\phi(\alpha + \mathbf{x}^o\boldsymbol{\beta} + q)], \tag{3.14}$$

which can be estimated only if we have consistent estimators of α and $\boldsymbol{\beta}$, and know the distribution of q (or have a consistent estimate). Nevertheless, (3.14) does show that, if we can estimate the β_j consistently up to a common scale factor, as in Manski (1975, 1985) and Horowitz (1992), then the relative effects of the continuous variables on the APEs can be found. (In fact, this is seen to be true if we replace $\Phi(\cdot)$ in (3.13) with an unknown function.) Estimation of the relative APEs for discrete x_j does not seem possible unless we know more, although the β_j give the directions of the effects.

Under what additional assumptions can we estimate the APEs? Certainly this is not possible without specific assumptions. For example, if we assume that q given $\mathbf{x}$ has a conditional normal distribution, say $q|\mathbf{x} \sim \mathrm{Normal}(\eta + \mathbf{x}\boldsymbol{\delta},\omega^2)$, then only $(\boldsymbol{\beta} + \boldsymbol{\delta})/(1 + \eta^2)^{1/2}$ is identified. This is the usual kind of endogeneity problem encounted in econometric models (see Section 4). But suppose we assume the dependence between q and $\mathbf{x}$ is only in the conditional variance, and not the conditional mean:

$$q|\mathbf{x} \sim \text{Normal}[0,\exp(2\mathbf{x}\boldsymbol{\delta}) - 1], \qquad (3.15)$$

where subtracting one is a convenient way of allowing no unobserved heterogeneity when $\boldsymbol{\delta} = 0$ (and leads to a very popular way of modeling heteroskedasticity in probit models). Assumption (3.15), along with (3.13), are enough to identify the APEs. To see why, we apply Lemma 2.1 with $\mathbf{w} = \mathbf{x}$. For fixed $\mathbf{x}^{\circ}$, we first need to find $\mathrm{E}[\Phi(\alpha + \mathbf{x}^{\circ}\boldsymbol{\beta} + q)|\mathbf{x}]$. A standard calculation, using the fact that $q + u$ given $\mathbf{x}$ has a $\text{Normal}[0,\exp(2\mathbf{x}\boldsymbol{\delta})]$ distribution (where u is the standard normal random variable underlying (3.13)), gives

$$\mathrm{E}[\Phi(\alpha + \mathbf{x}^{\circ}\boldsymbol{\beta} + q)|\mathbf{x}] = \Phi[(\alpha + \mathbf{x}^{\circ}\boldsymbol{\beta})/\exp(\mathbf{x}\boldsymbol{\delta})]. \qquad (3.16)$$

Therfore, provided α, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ are identified (they are), the APEs are consistently estimated by taking derivatives, or changes, with respect to $\mathbf{x}^{\circ}$ in

$$N^{-1}\sum_{i=1}^{N}\Phi[(\hat{\alpha} + \mathbf{x}^{\circ}\hat{\boldsymbol{\beta}})/\exp(\mathbf{x}_i\hat{\boldsymbol{\delta}})] \qquad (3.17).$$

Notice the different roles played by $\mathbf{x}^{\circ}$, which multiplies $\hat{\boldsymbol{\beta}}$, and $\mathbf{x}_i$, which appears in the variance function. The latter are averaged out, whereas it is the former that we compute changes or derivatives with respect to. Taking the derivative with respect to x_j° shows that the APE for a continuous variable is estimated as

$$\hat{\beta}_j\left(N^{-1}\sum_{i=1}^{N}[1/\exp(\mathbf{x}_i\hat{\boldsymbol{\delta}})]\phi[(\hat{\alpha} + \mathbf{x}^{\circ}\hat{\boldsymbol{\beta}})/\exp(\mathbf{x}_i\hat{\boldsymbol{\delta}})]\right), \qquad (3.18)$$

which shows that the scale factor needed to turn $\hat{\beta}_j$ into an APE can be easily estimated. APEs defined as discrete changes are also easy to estimate.

Why are the parameters in (3.16) identified? The reason is simple. Under (3.13) and (3.15),

$$\mathrm{P}(y = 1|\mathbf{x}) = \Phi[(\alpha + \mathbf{x}\boldsymbol{\beta})/\exp(\mathbf{x}\boldsymbol{\delta})], \qquad (3.19)$$

and therefore α, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ can be consistently estimated by conditional maximum likelihood. Equations (3.16) and (3.19) reveal something interesting about this example. While the response probability, conditional only on $\mathbf{x}$, and (3.16) have the same form, the average partial effects are *not* obtained by differentiating (3.19) with respect to $\mathbf{x}$. Yet, the APEs are identified.

This somewhat peculiar feature hinges on the specification of $\mathrm{E}(y|\mathbf{x},q)$ and, just as importantly, on the special way that q and $\mathbf{x}$ are assumed to be related. If, instead, $q|\mathbf{x} \sim \mathrm{Normal}(\eta + \mathbf{x}\boldsymbol{\delta},\omega^2)$ -- so that the conditional mean of q depends on $\mathbf{x}$, rather than just the conditional variance, then the APEs are not even identified, let alone obtained from $\mathrm{P}(y = 1|\mathbf{x})$.

An analogous linear example may add further insight. Suppose that, for a scalar x, $\mathrm{E}(y|x,q) = \alpha + \beta x + q$, so that the APE with respect to x is simply β. If $\mathrm{E}(q|\mathbf{x}) = \eta + \delta x$ then β is not identified. But suppose we assume $\mathrm{E}(q|\mathbf{x}) = \eta + \delta x^2$. Then $\mathrm{E}(y|x) = (\alpha + \eta) + \beta x + \delta x^2$, and so β is identified. But the APE of x is not obtained by taking the derivative of or changes in $\mathrm{E}(y|x)$.

It is difficult to imagine a researcher running a regression of y on a quadratic in x but then reporting as the relevant effect the coefficient on x. One cannot distinguish omitted heterogeneity from a misspecified functional form. The same is true in the probit example. While (3.18) is the appropriate partial effect under (3.13) and (3.15), the response probability in (3.19) is indistinguishable from that obtained from the model

$$\mathrm{P}(y = 1|\mathbf{x},q) = \Phi\{\alpha + \mathbf{x}\boldsymbol{\beta} + [\exp(2\mathbf{x}\boldsymbol{\delta}) - 1]^{1/2}q\} \qquad (3.20)$$

$$q|\mathbf{x} \sim \mathrm{Normal}(0,1). \qquad (3.21)$$

Under (3.20) and (3.21), the APEs are computed by taking derivatives of or changes in (3.19), which is the common way of obtaining partial effects in heteroskedastic probit models. Of course model (3.13) is more appealing than (3.20), but more common methods of interacting $\mathbf{x}$ and q also lead to models where the composite error in the latent variable model is heteroskedastic. For example, a probit version of the random coefficient model is obtained as $P(y|\mathbf{x},q) = \Phi(\alpha + \mathbf{x}\boldsymbol{\beta} + q + q\cdot\mathbf{x}\boldsymbol{\gamma})$ with $q|\mathbf{x} \sim \text{Normal}(0,\tau^2)$. Then $P(y = 1|\mathbf{x}) = \Phi\{(\alpha + \mathbf{x}\boldsymbol{\beta})/[1 + \tau^2(1 + \mathbf{x}\boldsymbol{\gamma})^2]^{1/2}\}$, and the APEs are obtained from taking the partial derivatives of this response probability with respect to $\mathbf{x}$. The APE of x_j could have the opposite sign from β_j.

3.3. Tobit Model

For the basic Tobit model, the literature appears to be silent on the effects of unobserved heterogeneity that satisfies an assumption like (3.6). I think this is partly a case of getting the right answer for the wrong reason. Recall that Tobit models are applied to two conceptually different problems. One application is to data censoring, where an underlying response variable satisfies the classical linear model but is censored because of data collection methods or institutional constraints. As in the willingness-to-pay probit example, interest is unambiguously in the coefficients β_j in $E(y^*|\mathbf{x})$. It is easy to see that unobserved heterogeneity satisfying (3.6) causes no problems in estimating $\boldsymbol{\beta}$. But here my interest is in a more common application of Tobit models: the response we are interested in, y, has a population distribution that piles up at zero, but is roughly continuously distributed over strictly positive values. [In Wooldridge (2000a, 2001) I call such responses *corner solution outcomes*.] Then, we are interested in conditional means involving y, and these are nonlinear functions of $E(y^*|\mathbf{x}) =$

$\mathbf{x}\boldsymbol{\beta}$ and $\mathrm{Var}(y^*|\mathbf{x}) = \sigma^2$. So how does the presence of unobserved heterogeneity affect estimation of the quantities of interest? The results are very similar to the probit case: we cannot identify the partial effects that the mean heterogeneity, but we can identify the average partial effects.

Write the Tobit model with unobserved heterogeneity as

$$y = \max(0, \alpha + \mathbf{x}\boldsymbol{\beta} + q + u), \tag{3.22}$$

$$u|\mathbf{x}, q, \mathbf{w} \sim \mathrm{Normal}(0, \sigma^2) \tag{3.23}$$

where, again, we have in mind cases where y is not a censored variable but rather a corner solution outcome. Typically, we are interested in $\mathrm{E}(y|\mathbf{x}, q)$ or $\mathrm{E}(y|\mathbf{x}, q, y > 0)$, and the formulas for these are well known (for example, Wooldridge (2000a, Section 17.2)). These expectations depend on α, $\boldsymbol{\beta}$, and σ^2. Since $\mathrm{E}(q) = 0$, we might evaluate the expectations at $q = 0$. But even if we assume (3.6) -- as we now do -- σ^2 is not identified, and so neither are $\mathrm{E}(y|\mathbf{x}, q, y > 0)$ or $\mathrm{E}(y|\mathbf{x}, q)$ for any value of q. But the average partial effects are identified, as we now see.

Under (3.22), (3.23), and (3.6), we have

$$\begin{aligned} y &= \max(0, \lambda + \mathbf{x}\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\delta} + a + u) \\ &\equiv \max(0, \lambda + \mathbf{x}\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\delta} + v), \end{aligned} \tag{3.24}$$

where λ, a, and v are defined as in the probit case. Under the assumptions made, $v|(\mathbf{x}, \mathbf{w}) \sim \mathrm{Normal}(0, \omega^2)$ where $\omega^2 = \tau^2 + \sigma^2$. It follows immediately that Tobit of y on $(1, \mathbf{x}, \mathbf{w})$ consistently estimates λ, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and ω^2. For data censoring, this is fine, as $\boldsymbol{\beta}$ is the vector of interest, and we can compute the appropriate asymptotic variance by estimating ω^2. But for corner solution outcomes, we might initially worry about our inability to estimate σ^2, as discussed above. But Lemma 2.2 implies that $\partial\mathrm{E}(y|\mathbf{x}, \mathbf{q})/\partial x_j$, averaged across the distribution of $q|\mathbf{w}$, is exactly what we estimate by Tobit of y on

$(1,\mathbf{x},\mathbf{w})$; having estimated λ, $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and ω^2, we just apply the usual formula for the partial effects in a Tobit model and then average across $\mathbf{w}_i$. Let $m(z,\sigma^2) = \Phi(z/\sigma)z + \sigma\phi(z/\sigma)$ for $z \in \mathbb{R}$, $\sigma^2 > 0$. Then the APEs are estimated from

$$N^{-1}\sum_{i=1}^{N} m(\hat{\lambda} + \mathbf{x}^{\circ}\hat{\boldsymbol{\beta}} + \mathbf{w}_i\hat{\boldsymbol{\delta}}, \hat{\omega}^2), \tag{3.25}$$

where all estimates are the Tobit MLEs. Therefore, our initial reaction that we may safely ignore the heterogeneity is correct, but for the wrong reason: we are not just interested in $\boldsymbol{\beta}$, but we obtain consistent estimators of the APEs by nevertheless ignoring the heterogeneity. The same is true for computing the APEs on $\mathrm{E}(y|\mathbf{x},q,y > 0)$, or any other moments that might be of interest.

If we drop $\mathbf{w}$ but introduce heteroskedasticity into q, as in the probit example, the conclusions are very similar. In particular, the parameters indexing the APEs are identified from the distribution of y given $\mathbf{x}$, and can be consistently estimated by conditional maximum likelihood. However, the APEs are not obtained from the distribution of y given $\mathbf{x}$. In fact, they would be obtained from

$$N^{-1}\sum_{i=1}^{N} m[\hat{\alpha} + \mathbf{x}^{\circ}\hat{\boldsymbol{\beta}}, h(\mathbf{x}_i,\hat{\boldsymbol{\theta}})], \tag{3.26}$$

where $h(\mathbf{x},\boldsymbol{\theta})$ is the model for $\mathrm{Var}(q + u|\mathbf{x})$, for example, $h(\mathbf{x}_i,\boldsymbol{\theta}) = \omega^2\exp(\mathbf{x}\boldsymbol{\delta})$.

4. APPLICATION TO MODELS WITH ENDOGENOUS EXPLANATORY VARIABLES

I now turn to some examples of nonlinear models where instrumental variables assumptions are needed to identify the average partial effects. Blundell and Powell (2000) cover semiparametric and nonparametric approaches and so much of what is here is subsumed by their work. Here, I focus on flexible parametric models and the adjustments needed to turn parameter estimates from common two-step procedures into estimates of the APEs.

4.1. A Probit Response Function

We consider a model where y_1 has a conditional expectation that follows a probit model. Of course, this is applicable to the binary response case, but it can also be applied when y_1 is a fractional response that may or may not take on the extreme values of zero and one. (Examples include when y_1 is the fraction of pension assets invested in the stock market and the fraction of students in a class passing a standardized exam.) The probit conditional mean function keeps all predicted values in the unit interval, and, as in Section 3.2, is very convenient for identifying APEs. For simplicity, we consider a model with a scalar endogenous explanatory variable.

The structural expectation is

$$\mathrm{E}(y_1|\mathbf{z},y_2,q_1) = \mathrm{E}(y_1|\mathbf{z}_1,y_2,q_1) = \Phi[\mathbf{g}_1(\mathbf{z}_1,y_2)\boldsymbol{\beta}_1 + q_1], \tag{4.1}$$

where y_2 is the endogenous explanatory variable, $\mathbf{z}$ is the set of exogenous variables, and q_1 is the unobserved heterogeneity. The function $\mathbf{g}_1(\cdot)$ is assumed known, and allows $\mathbf{z}_1$ and y_2 to appear in a very flexible manner. For example, we might have the levels of $\mathbf{z}_1$ and y_2 along with squares and interactions among several or all variables, including y_2. For notational convenience, we will write $\mathbf{x}_1 \equiv \mathbf{g}_1(\mathbf{z}_1,y_2)$. We would virtually always include a constant in $\mathbf{x}_1$ and so, without loss of generality, we assume $\mathrm{E}(q_1) = 0$ and, in fact, $q_1 \sim \mathrm{Normal}(0,\sigma_1^2)$. While q_1 is assumed to be independent of $\mathbf{z}$, it is allowed to be correlated with y_2. Therefore, as in all instrumental variables contexts, we need at least one element in $\mathbf{z}$ that is not also in $\mathbf{z}_1$.

Under the normality assumption for q_1, the average partial effects in (4.1) with respect to elements of $(\mathbf{z}_1, y_2)$ are indexed by the vector $\boldsymbol{\beta}_1/(1 + \sigma_1^2)^{1/2}$, something that follows directly from Section 3.2. Without further assumptions, the APEs are not identified. We add the assumption that a strictly monotonic transformation of y_2 has a linear reduced form with an additive normal disturbance:

$$h(y_2) = \mathbf{z}\boldsymbol{\delta}_2 + v_2, \; v_2|\mathbf{z} \sim \text{Normal}(0,\sigma_2^2), \tag{4.2}$$

where h is strictly monotone on the support of y_2. (It is convenient here to absorn an intercept into $\mathbf{z}$.) The monotonicity of h ensures that we can write $y_2 = h^{-1}(\mathbf{z}\boldsymbol{\delta}_2 + v_2)$. For what follows, we actually assume that (q_1, v_2) is independent of $\mathbf{z}$ with a bivariate normal distribution. Then y_2 is endogenous in (4.1) if and only if q_1 and v_2 are correlated. We write

$$q_1 = \gamma_1 v_2 + e_1 \tag{4.3}$$

where $\gamma_1 = \sigma_{12}/\sigma_1^2$, so that e_1 is independent of $(\mathbf{z}, v_2)$ with a Normal$(0,\sigma_1^2 - \gamma_1^2\sigma_2^2)$ distribution. A simple two-step estimation approach follows Smith and Blundell (1986) and Rivers and Vuong (1988), with the slight extension that we are generally applying a conditional quasi-MLE unless y_1 is binary. (Wooldridge (1997) uses the quasi-MLE approach for estimating exponential regression models with endogenous explanatory variables.) Using assumptions we have made,

$$\text{E}(y_1|\mathbf{z}, y_2, v_2) = \Phi\{[\mathbf{g}_1(\mathbf{z}_1, y_2)\boldsymbol{\beta}_1 + \gamma_1 v_2]/(1 + \eta_1^2)^{1/2}\}, \tag{4.4}$$

where $\eta_1^2 = \text{Var}(e_1) \leq \sigma_1^2$. The parameters we can immediately estimate consistently are $\boldsymbol{\beta}_{\eta 1} \equiv \boldsymbol{\beta}_1/(1 + \eta_1^2)^{1/2}$ and $\gamma_{\eta 1} \equiv \gamma_1/(1 + \eta_1^2)^{1/2}$. That the following two-step procedure works is fairly clear from general two-step estimation results and the fact that the Bernoulli log-likelihood is in the linear exponential family (see Gourieroux, Monfort, and Trognon (1984); Papke and Wooldridge (1996)).

1. Run the OLS regression of

$$h(y_{i2}) \text{ on } \mathbf{z}_i,\ i = 1,...,N \tag{4.5}$$

and save the residuals, $\hat{v}_{i2}$.

2. Using the Bernoulli quasi-log-likelihood function by doing "probit" of

$$y_{i1} \text{ on } \mathbf{x}_{i1},\ \hat{v}_{i2},\ i = 1,...,N \tag{4.6}$$

to obtain $\hat{\boldsymbol{\beta}}_{\eta 1}$ and $\hat{\gamma}_{\eta 1}$.

If $\gamma_1 \neq 1$, the asymptotic variance matrix from step 2 needs to be adjusted for the first-stage estimation of $\boldsymbol{\delta}_2$; see, for example, Newey and McFadden (1994). If $\gamma_1 = 0$ -- as occurs under the null hypothesis that y_2 is exogenous -- first stage estimation of $\boldsymbol{\delta}_2$ can be ignored, but the variance matrix estmator has to be the Huber-White "sandwich" form unless y_1 is a binary response. See, for example, Papke and Wooldridge (1996) for the specific formulas for the Bernoulli case.

The main question is: how can we turn the estimates from the second-stage program into parameter estimates that index the APEs? This is potentially important because, when $\gamma_1 \neq 0$, $\eta_1^2 < \sigma_1^2$. Focusing on the parameter estimates from the second stage probit can lead us to overestimate the importance of the explanatory variables (although, of course, the relative effects are unaffected). It is somewhat common to compare the parameter estimates from a probit that assumes exogeneity of y_2 with those from a two-step procedure. But the magnitudes are not directly comparable because of the rescaling that occurs when adding $\hat{v}_{i2}$ to the probit. (Whether we underestimate or overestimate the partial effects depends on the values of

$(\mathbf{z}_1, y_2)$ that we plug in.) Plus, in order to compare the nonlinear probit model that accounts for endogeneity with those from a simple linear probability model estimated by two stage least squares, we need to estimate the APEs.

Fortunately, it is very easy to adjust $\hat{\boldsymbol{\beta}}_{\eta 1}$ to obtain estimates of $\boldsymbol{\beta}_{\sigma 1} = \boldsymbol{\beta}/(1 + \sigma_1^2)^{1/2}$. One possibility is to simply apply Lemma 2.2 with $\mathbf{x} \equiv (\mathbf{z}_1, y_2)$, $q \equiv q_1$, and $\mathbf{w} \equiv v_2$. First, Assumption A.1 holds because v_2 is a function of $(\mathbf{z}, y_2)$ and so $\mathrm{E}(y_1|\mathbf{z}, y_2, q_1, v_2) = \mathrm{E}(y_1|\mathbf{z}, y_2, q_1)$. Second, Assumption A.2 holds because y_2 is a function of $(\mathbf{z}, v_2)$, and (q_1, v_2) is independent of $\mathbf{z}$: $\mathrm{E}(q_1|\mathbf{z}, y_2, v_2) = \mathrm{E}(q_1|\mathbf{z}, v_2) = \mathrm{E}(q_1|v_2)$. It now follows from Lemma 2.2 that the APEs with respect to $(\mathbf{z}_1, y_2)$ are obtained from

$$\mathrm{E}_{v_2}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_{\eta 1} + \gamma_{\eta 1} v_2)], \tag{4.7}$$

where $\mathbf{x}_1$ is fixed and the expectation is over the distribution of v_2. A simple calculation using the convolution properties of the normal distribution shows that (4.7) equals

$$\Phi[(\mathbf{x}_1\boldsymbol{\beta}_{\eta 1}/(1 + \gamma_{\eta 1}^2\sigma_2^2)^{1/2}]. \tag{4.8}$$

Therfore, having obtained $\hat{\boldsymbol{\beta}}_{\eta 1}$ and $\hat{\gamma}_{\eta 1}$ from (4.6), and $\hat{\sigma}_2^2$ as the usual error variance estimate from (4.5), we simple divide each element of $\hat{\boldsymbol{\beta}}_{\eta 1}$ by $(1 + \hat{\gamma}_{\eta 1}^2\hat{\sigma}_2^2)^{1/2}$ to obtain the parameter estimates used in obtaining estimated APEs. An alternative approach, which does not exploit the normality of v_2, is to directly use the sample analog of (4.7):

$$N^{-1}\sum_{i=1}^{N}\Phi(\mathbf{x}_1\hat{\boldsymbol{\beta}}_{\eta 1} + \hat{\gamma}_{\eta 1}\hat{v}_{i2}). \tag{4.9}$$

This kind of estimate is used by Blundell and Powell (2000) in semiparametric contexts. If all of the assumptions are correct, (4.8) and (4.9) should yield similar estimates of the APEs.

As in Section 3.2, the derivations here have implications for comparing estimates across studies that use different sets of IVs. Different IVs imply different error variances for v_2, which means that different scale factors are implicit in the IV estimation. Once we focus on APEs, the effects different IV estimates are comparable.

As noticed by Blundell and Powell (2000), the focus on APEs is liberating, in the sense that much more flexible methods are possible. A simple parametric extension would be to allow q_1 given v_2 to have normal distribution with mean that is polynomial in v_2 and possibly heteroskedastic, too. (Then, unconditional normality of q_1 would be unrealistic, but that might be tolerable.) If we specify an exponential function for the heteroskedasticity, we would add a low-order polynomial in $\hat{v}_2$ to the quasi-probit in equation (4.6) and also allow exponential heteroskedasticity in $\hat{v}_2$ (or a polynomial in $\hat{v}_2$). The APEs would be estimated from the analog of (4.9). For example, with a quadratic in $\hat{v}_2$ in the mean and $h(v_2,\boldsymbol{\theta}_2)$ the variance function, the estimated APEs would come from

$$N^{-1}\sum_{i=1}^{N}\Phi[(\mathbf{x}_1\hat{\boldsymbol{\beta}}_1 + \hat{\gamma}_1\hat{v}_{i2} + \hat{\rho}_1\hat{v}_{i2}^2)/h(\hat{v}_{i2},\hat{\boldsymbol{\theta}}_2)],$$

where the estimates denote the quasi-MLEs from the two-step heteroskedastic probit. (The scaling is implicit here.) While analytical standard errors would be cumbersome to compute, simulation methods should be tractable. This also suggests a simple test of the assumption that q_1 is independent of $\mathbf{x}_1$, conditional on v_2, even if we only have one instrument for y_2. For example, we could specify a variance function $\exp(\theta_{21}v_2 + \mathbf{x}_1\boldsymbol{\theta}_{22})$ in the probit model and test $\boldsymbol{\theta}_{22} = \mathbf{0}$.

4.2. Tobit Model

Obtaining APEs after two-step estimation of a Tobit model with an endogenous explanatory variable is also simple. Along with (4.2) write

$$y_1 = \max[0, \mathbf{g}_1(\mathbf{z}_1, y_2)\boldsymbol{\beta}_1 + q_1], \qquad (4.10)$$

where (q_1, v_2) is independent of $\mathbf{z}$ with a zero-mean normal distribution. Define e_1 as in (4.3), where $\sigma_1^2 = \mathrm{Var}(q_1)$, $\sigma_2^2 = \mathrm{Var}(v_2)$, and $\gamma_1 = \mathrm{Cov}(q_1, v_2)/\mathrm{Var}(v_2)$.

For estimating APEs, we need the function $m(\mathbf{x}_1^{\circ}\boldsymbol{\beta}_1, \sigma_1^2)$, where the function $m(\cdot,\cdot)$ is defined in Section 3.3. The Smith-Blundell (1986) is to write

$$y_1 = \max(0, \mathbf{x}_1\boldsymbol{\beta}_1 + \gamma_1 v_2 + e_1), \qquad (4.11)$$

where $e_1|\mathbf{z}, v_2 \sim \mathrm{Normal}(0, \eta_1^2)$ and $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, y_2)$. The Smith-Blundell procedure for estimating $\boldsymbol{\beta}_1$, γ_1, and η_1^2 is the same as for the probit, except in step 2 we do Tobit of y_{i1} on $\mathbf{x}_{i1}$, $\hat{v}_{i2}$, $i = 1,\ldots,N$.

Since $\sigma_1^2 = \eta_1^2 + \gamma_1^2\sigma_2^2$, and $\hat{\eta}_1^2$, $\hat{\gamma}_1$, and $\hat{\sigma}_2^2$ are immediately available (with the latter coming from equation (4.5)), the APEs are easy to obtain. The same sorts of embellishments possible for probit apply to Tobit, too.

5. APPLICATION TO PANEL DATA MODELS

The basic insights about estimating partial effects averaged across unobserved heterogeneity apply to panel data models. Because of replicability across time, we can account for time-constant unobserved heterogeneity in ways not possible with cross section data. We let $\{(\mathbf{x}_t, y_t): t = 1,\ldots, T\}$ denote a generic cross section unit observed over T time periods. The vector of unobserved heterogeneity, $\mathbf{q}$, is initially time-constant.

5.1. Models Under Strict Exogeneity

We assume that interest lies in the sequence of structural expectations

$$\mu_{t1}(\mathbf{x}_t,\mathbf{q}) \equiv \mathrm{E}(y_t|\mathbf{x}_t,\mathbf{q}),\ t = 1,\ldots,T. \tag{5.1}$$

The average partial effects of the elements of $\mathbf{x}_t$ are defined by taking the expected value of $\mu_{t1}(\mathbf{x}_t^{\circ},\mathbf{q})$ with respect to the distribution of $\mathbf{q}$. A common assumption in models where $\mathbf{x}_t$ does not contain lagged y_t is strict exogeneity, conditional on $\mathbf{q}$:

$$\mathrm{E}(y_t|\mathbf{x}_1,\ldots,\mathbf{x}_T,\mathbf{q}) = \mathrm{E}(y_t|\mathbf{x}_t,\mathbf{q}),\ t = 1,\ldots,T. \tag{5.2}$$

Under (5.2) and certain functional form restrictions, we can identify the APEs by specifying a distribution for $\mathbf{q}$ given $\mathbf{w} \equiv (\mathbf{x}_1,\ldots,\mathbf{x}_T)$. Mundlak (1978) takes this approach for the linear unobserved effects model and Chamberlain (1980, 1984) analyzes the unobserved effects probit model. The key to estimating APEs is to apply Lemma 2.1, observing that the parameters of

$$\mu_{t2}(\mathbf{x}_t^{\circ},\mathbf{w}^{\circ}) \equiv \int_{Q} \mu_{t1}(\mathbf{x}_t^{\circ},q)\,g(q|\mathbf{w}^{\circ})\,\eta(dq) \tag{5.3}$$

are often identified from the expectations

$$\mathrm{E}(y_t|\mathbf{w}),\ t = 1,\ldots,T. \tag{5.4}$$

As an example, consider the case of a probit response function, so that the structural model is

$$\mathrm{E}(y_t|\mathbf{x}_t,q) = \Phi(\alpha_t + \mathbf{x}_t\boldsymbol{\beta} + q), \tag{5.5}$$

where the α_t denote different time intercepts inside $\Phi(\cdot)$. Now suppose, as in Chamberlain, we add

$$q|(\mathbf{x}_1,\ldots,\mathbf{x}_T) \sim \mathrm{Normal}(\mathbf{w}\boldsymbol{\lambda},\eta^2), \tag{5.6}$$

where an intercept is omitted because a full set of year intercepts are already in (5.5). The same calculation used in Section 3.2 shows that, under (5.2), (5.5), and (5.6),

$$\mathrm{E}(y_t|\mathbf{w}) = \Phi[(\alpha_t + \mathbf{x}_t\boldsymbol{\beta} + \mathbf{w}\boldsymbol{\lambda})/(1 + \eta^2)^{1/2}],\ t = 1,\ldots,T. \tag{5.7}$$

Therefore, we can estimate the scaled versions of α_t, $\boldsymbol{\beta}$, and $\boldsymbol{\lambda}$ -- where each is divided by $(1 + \eta^2)^{1/2}$ -- using pooled quasi-MLE, where the quasi-likelihood function is Bernoulli with a probit response probability depending on year dummies, $\mathbf{x}_{it}$, and $\mathbf{w}_i$, $t = 1,\ldots,T$, $i = 1,\ldots,N$. Using this approach, we cannot separately estimate $\boldsymbol{\beta}$ or η^2, but this is not important for computing average partial effects. (This is one of those cases where the parameters in the structural expectation and the heterogeneity distribution are not separately identified, but the APEs are.) In fact, the same argument that leads to (5.7) shows that

$$\mu_{t2}(\mathbf{x}_t^o,\mathbf{w}) = \Phi(\alpha_{\eta t} + \mathbf{x}_t^o\boldsymbol{\beta}_\eta + \mathbf{w}\boldsymbol{\lambda}_\eta),$$

which means that the APEs are obtained from

$$N^{-1}\sum_{i=1}^{N}\Phi(\hat{\alpha}_{\eta t} + \mathbf{x}_t^o\hat{\boldsymbol{\beta}}_\eta + \mathbf{w}_i\hat{\boldsymbol{\lambda}}_\eta). \tag{5.8}$$

Chamberlain (1984) derived this formula for the pure probit case.

An alternative is to use a more efficient generalized method of moments estimation method. Equation (5.7) implies many moment conditions. Because there is almost certain serial dependence in $\{y_t\}$, conditional on $(\mathbf{x}_1,\ldots,\mathbf{x}_T)$, more efficient estimation is possible.

Often, Chamberlain's unobserved effects probit model is estimated under stronger assumptions, particularly independence of the y_t conditional on $(\mathbf{x}_1,\ldots,\mathbf{x}_T)$ and q. Then, the unscaled coefficients, along with η^2, can be estimated using standard random effects probit, but where $\mathbf{w}_i$ is added as an additional set of controls in each time period; see, for example, Hsiao (1986). The previous analysis shows that the APEs are identified without further assumptions. In particular, the APEs are identified in the fractional response case without distributional assumptions on y_t or conditional independence assumptions.

Sometimes it is important to conserve on parameters in modeling the distribution of q given $(\mathbf{x}_1,\ldots,\mathbf{x}_T)$. A common assumption is that this distribution depends only on the time average:

$$D(q|\mathbf{x}_1,\ldots,\mathbf{x}_T) = D(q|\bar{\mathbf{x}}). \tag{5.9}$$

In the probit response example, we would then choose $\mathbf{w}_i = \bar{\mathbf{x}}_i$ for all i. Then, $\boldsymbol{\lambda}$ would be k x 1, where k is the dimension of $\mathbf{x}_t$, rather than Tk x 1. All calcuations are unchanged.

5.2. Ignorability Assumptions

An assumption such as (5.9) can be restrictive (along with specifying a parametric model for $D(q|\mathbf{x})$), but (5.9) can also have some benefits. In particular, if we assume (5.2) and

$$D(\mathbf{q}|\mathbf{x}_t,\bar{\mathbf{x}}) = D(\mathbf{q}|\bar{\mathbf{x}}), \tag{5.10}$$

then we can apply Lemma 2.2 with $\mathbf{x} = \mathbf{x}_t$ and $\mathbf{w} = \bar{\mathbf{x}}$. In particular, let

$$\mu_{t2}(\mathbf{x}_t,\bar{\mathbf{x}}) = E(y_t|\mathbf{x}_t,\bar{\mathbf{x}}). \tag{5.11}$$

Provided the $\mathbf{x}_t$ vary over time, (5.11) is typically identified for each t. Then, APEs for time period t are estimated from

$$N^{-1}\sum_{i=1}^{N}\hat{\mu}_{t2}(\mathbf{x}_t^{\circ},\bar{\mathbf{x}}_i). \tag{5.12}$$

In principle, the models for $E(y_t|\mathbf{x}_t,\bar{\mathbf{x}})$ can be very flexibly. Under stationarity over time, the estimated APEs could then be averaged across t.

Or, as is more common, the models for $\mu_{t2}(\mathbf{x}_t,\bar{\mathbf{x}})$ could be restricted in how they change over time, and then pooled estimation used. A standard index formulation is

$$E(y_t|\mathbf{x}_t,\bar{\mathbf{x}}) = m(\alpha_t + \mathbf{x}_t\boldsymbol{\beta} + \bar{\mathbf{x}}\boldsymbol{\delta}), \tag{5.13}$$

where $m(\cdot)$ is a known function that accounts for the features of y_t. (There is nothing necessarily "structural" about the parameters.) Then, one could estimate the parameters using a pooled least squares or quasi-likelihood procedure. Or, we could allow the coefficients on $\bar{\mathbf{x}}$ to change over time to allow for the possibility that the effect of the unobserved heterogeneity changes over time. There would almost certainly be serial correlation that could be exploited in a GMM procedure.

If we are going to focus on $E(y_t|\mathbf{x}_t,\bar{\mathbf{x}})$ to estimate APEs, then there is no reason to use traditional formulations derived from specific unobserved effects models. For example, we could interact the elements of $\mathbf{x}_t$ and $\bar{\mathbf{x}}$ inside a standard probit or logit function for a binary or fractional response, or even include these inside a Tobit model for a corner solution response. Or, we can use probit models or Tobit models with heteroskedasticity that depends on $\bar{\mathbf{x}}$, and then use such models to test the key assumption (5.10). A specific model when $0 \leq y_t \leq 1$ is

$$E(y_t|\mathbf{x}_t,\bar{\mathbf{x}}) = \Phi[(\alpha_t + \mathbf{x}_t\boldsymbol{\beta} + \bar{\mathbf{x}}\boldsymbol{\delta})/\exp(\bar{\mathbf{x}}\boldsymbol{\gamma})], \tag{5.14}$$

which can be obtained from an unobserved effects probit model where the unobserved heterogeneity has a normal distribution with mean linear in $\bar{\mathbf{x}}$ and exponential heteroskedasticity in $\bar{\mathbf{x}}$. The parameters in (5.14) are easily estimated by pooled QMLE. The APEs are obtained from

$$N^{-1}\sum_{i=1}^{N}\Phi[(\hat{\alpha}_t + \mathbf{x}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_i\hat{\boldsymbol{\delta}})/\exp(\bar{\mathbf{x}}_i\hat{\boldsymbol{\gamma}})]. \tag{5.15}$$

Assumption (5.10) can be tested in the underlying unobserved effects model by adding $\mathbf{x}_{it}$ to the variance function and testing joint significance in a robust way.

It is straightforward to add time-constant observables to any model for $E(y_t|\mathbf{x}_t,\bar{\mathbf{x}})$. Often, this would be desirable, as (5.10) might be more reasonable if we control for observed heterogeneity that is time-constant. In the structural model (5.1) we would either assume that the time-constant observables do not matter, conditional on $(\mathbf{x}_t,\mathbf{q})$, or admit that we cannot estimate the effects of the time-constant variables. Either way, they can be included as controls in estimating the APEs of the time-varying elements, $\mathbf{x}_t$.

There are two points to take away from these examples. First, if we want to specify a fairly standard parametric unobserved effects models, it is easy to estimate the APEs under (5.2), (5.10), and parametric models for $D(\mathbf{q}|\mathbf{x})$. We do not need extra assumptions, particularly on time dependence, that is usually added. The second point is more controvesial: if we want to estimate APEs in an unobserved effects model, and we are willing to assume (5.2) and (5.10), then we can just specify very flexible models for $E(y_t|\mathbf{x}_t,\bar{\mathbf{x}})$.

The strict exogeneity assumption (5.2) and the ignorability assumption (5.10) may be too strong for some applications. Plus, modeling $D(\mathbf{q}|\bar{\mathbf{x}})$ generally requires having a balanced panel data set. It is possible to relax the strict exogeneity requirement along with the balanced panel assumption by using different ignorability assumptions. Let $\mathbf{x}_0$ be the observed responses on $\mathbf{x}_t$ in an initial time period. Sometimes it may be reasonable to assume that, conditional on the intial values, the path of $\mathbf{x}_t$ is independent of $\mathbf{q}$:

$$D(\mathbf{q}|\mathbf{x}_t,\mathbf{x}_0) = D(\mathbf{q}|\mathbf{x}_0),\ t = 1,\ldots,T. \tag{5.16}$$

(We might also include time-constant observables in both conditioning sets in

(5.16).) For example, suppose y_t is a student performance measure at the school level, and $\mathbf{x}_t$ contains per student spending. It might be reasonable to assume that controlling for initial spending makes future spending independent of the time-constant features of a school that affect performance. This would be the case if, between time zero and time one, a reform took place that made spending in $t \geq 1$ an exogenous function of time zero spending.

To apply (5.16) we would add the assumption

$$\mathrm{E}(y_t|\mathbf{x}_t,\mathbf{q},\mathbf{x}_0) = \mathrm{E}(y_t|\mathbf{x}_t,\mathbf{q}), \tag{5.17}$$

which simply says that $\mathbf{x}_0$ does not matter once $\mathbf{x}_t$ and $\mathbf{q}$ have been controlled for. (Of course, $\mathbf{x}_t$ could contain lagged explanatory variables, so the practical restriction is having enough time periods and a finite distributed lag effect.) Under (5.16) and (5.17), all APEs for $\mathbf{x}_t$ are obtained from $\mathrm{E}(y_t|\mathbf{x}_t,\mathbf{x}_0)$. Given standard models for $\mathrm{E}(y_t|\mathbf{x}_t,\mathbf{q})$ and $\mathrm{D}(\mathbf{q}|\mathbf{x}_0)$, $\mathrm{E}(y_t|\mathbf{x}_t,\mathbf{x}_0)$ can be easy to derive. For example, with a probit response function, we can replace $\bar{\mathbf{x}}$ with $\mathbf{x}_0$ in equations (5.14) and (5.15). Note that we would not have to have a balanced panel to estimate $\mathrm{E}(y_t|\mathbf{x}_t,\mathbf{x}_0)$ for each t (provided the reason the data are missing is appropriately random). Further, the $\mathbf{x}_t$ need not be strictly exogenous: (5.16) and (5.17) do not rule out feedback from idiosyncratic changes in y_t to future changes in $\mathbf{x}_t$. In the school spending example, a school could increase spending next year based on unexpectedly poor performance this year.

5.3. Models with Endogenous Explanatory Variables

Average partial effects can also be estimated for panel data models that combine unobserved effects and explanatory variables that are endogenous due to correlation with time-varying unobservables. Here I provide one example to illustrate the approach and the kinds of assumptions that suffice. Consider again a probit response function, so that $0 \leq y_{t1} \leq 1$, $t = 1,...,T$, and a single continuous, endogenous explanatory variable, y_{t2}. The set of exogenous variables at time t is $\mathbf{z}_t$. Then the structural model is

$$\begin{aligned} \mathrm{E}(y_{t1}|\mathbf{z}_1,...,\mathbf{z}_T,y_{t2},c_1,c_2,u_{t1},u_{t2}) &= \mathrm{E}(y_{t1}|\mathbf{z}_{t1},y_{t2},c_1,u_{t1}) \\ &= \Phi(\alpha_t + \mathbf{x}_{t1}\boldsymbol{\beta}_1 + c_1 + u_{t1}), \end{aligned} \tag{5.18}$$

where α_t is a full set of time intercepts, $\mathbf{x}_{t1}$ is a general, known function of $(\mathbf{z}_{t1},y_{t2})$ (the leading case is $\mathbf{x}_{t1} = (\mathbf{z}_{t1},y_{t2})$), (c_1,c_2) are time-constant unobserved heterogeneity, and (u_{t1},u_{t2}) are time-varying unobserved heterogeneity. A key assumption is that the $\{\mathbf{z}_t\}$ are strictly exogenous conditional on everything else. Plus, (5.18) imposes an exclusion restriction: there must be at least one element in $\mathbf{z}_t$ that is not in $\mathbf{z}_{t1}$. Note that y_{t2} is not assumed to be strictly exogenous. In fact, as we will see, this model is appropriate for cases where y_{t2} is contemporaneously exogenous but might not be strictly exogenous, provided we have an instrument for y_{t2} in $\mathbf{z}_{t2}$.

An example of where this model might apply is estimating the effects of per-student spending on student pass rates, at the school level. Then y_{it1} is the pass rate for school i in year t and y_{it2} would be per-student spending. Allowing for a school effect, c_{i1}, might not be sufficient for accounting for endogeneity of spending. For example, schools might have knowledge of the crop of students coming up to take the test -- information in u_{t1} that the econometrician would not generally observe -- and adjust resources accordingly. In some states an instrument for spending is a grant determined by an earlier level of spending. Allowing for c_{i1} in the equation makes the grant exogenous, and so this can be used as an IV for actual school spending. Papke (2001) analyzes this setup for Michigan using linear models for the pass rate.

As a second example, y_{it1} could be a binary indicator of whether firm i receives a patent in year t, and y_{it2} is a measure of current and past R&D spending. We might be worried that R&D spending, while contemporaneous exogenous (uncorrelated with u_{t1}) is not strictly exogenous because receiving a patent this year might affect next year's R&D spending. Then, $y_{t+1,2}$ could be correlated with u_{t1}. Even though y_{t2} and u_{t1} are uncorrelated, we cannot treat y_{t2} as an element of $\mathbf{z}_t$. If we have a strictly exogenous instrument for y_{t2} (which may still be correlated with c_1), then this model applies.

The reduced form for y_{t2} is

$$y_{t2} = \eta_t + \mathbf{z}_t\boldsymbol{\delta}_2 + c_2 + u_{t2}. \tag{5.19}$$

Therefore, (5.18) assumes, essentially tautologically, that c_2 and u_{t2} do not appear in the structural equation once everything else has been conditioned on. (Replacing y_{t2} in (5.19) with an invertible transformation only requires a slight notational change.)

Evidently, we cannot get very far in estimating the APEs with respect to $(\mathbf{z}_{t1}, y_{t2})$ without making some fairly strong assumptions. While some of the following can be relaxed, we assume that we can write

$$c_1 = \bar{\mathbf{z}}\boldsymbol{\xi}_1 + a_1,\ c_2 = \bar{\mathbf{z}}\boldsymbol{\xi}_2 + a_2, \tag{5.20}$$

where, for all t,

$$(a_1, a_2, u_{t1}, u_{t2})|(\mathbf{z}_1, \ldots, \mathbf{z}_T) \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{5.21}$$

Assumption (5.20) is a Chamberlain-Mundlak device which allows all explanatory variables to be correlated with the unobserved heterogeneity.

Assumption (5.21) is an admittedly strong normality assumption. If we write $v_{t1} = a_1 + u_{t1}$ and $v_{t2} = a_2 + u_{t2}$, it is easily seen that (v_{t1}, v_{t2}) is independent of $\mathbf{z}$ and bivariate normal. Therefore, for all t we can write

$$v_{t1} = \gamma_1 v_{t2} + e_{t1}, \ e_{t1}|v_{t2},\mathbf{z} \sim \text{Normal}(0,\omega_1^2). \tag{5.22}$$

After plugging the equation for c_1 in (5.20) and the equation for v_{t1} into (5.18) and taking the expectation with respect to e_{t1}, we get something familiar from previous sections:

$$\begin{aligned} &\mathrm{E}(y_{t1}|\mathbf{z}_{t1},y_{t2},\bar{\mathbf{z}},v_{t2}) \\ &\qquad = \Phi[(\lambda_t + \mathbf{x}_{t1}\boldsymbol{\beta}_1 + \bar{\mathbf{z}}\boldsymbol{\xi}_1 + \gamma_1 v_{t2})/(1 + \omega_1^2)^{1/2}]. \end{aligned} \tag{5.23}$$

For each t, (5.18) and (5.23) show that Assumptions A.1 and A.2 of Lemma 2.2 hold with $\mathbf{x}_t = \mathbf{x}_{t1}$ and $\mathbf{w}_t = (\bar{\mathbf{z}}, v_{t2})$. Therefore, we have the following simple procedure.

1. Across $t = 1,...,T$, $i = 1,...,N$, estimate the Mundlak (1978)-type equation

$$y_{it2} = \eta_t + \mathbf{z}_{it}\boldsymbol{\delta}_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2} \tag{5.24}$$

by pooled OLS, or some other consistent procedure. (Under the additional assumption that $\{u_{it2}\}$ is homoskedastic and serially independent conditional on $(\mathbf{z}_i, c_{i2})$, random effects applied to (5.24) is asymptotically efficient. Even if this additional assumption fails, one can use random effects to account to some extent for the serial dependence in v_{it2}.) Keep the residuals $\hat{v}_{it2} = y_{it2} - \hat{\eta}_t - \mathbf{z}_{it}\hat{\boldsymbol{\delta}}_2 - \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_2$, all i and t, and the estimated error variance, $\hat{\tau}_2^2$.

2. Perform pooled quasi-MLE of y_{it1} on $\Phi(\lambda_{\omega t} + \mathbf{x}_{it1}\boldsymbol{\beta}_{\omega 1} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_{\omega 1} + \gamma_{\omega 1}\hat{v}_{it2})$, using the Bernoulli log-likelihood function. (Or, use a GMM procedure.) The indexing by ω denotes that the scaled coefficients have been estimated. A test of endogeneity of y_{it2}, that is, correlation of y_{t2} with

u_{t1}, is easily tested as H_0: $\gamma_{\omega 1} = 0$ using a standard robust t statistic.

3. For time period t, estimate average partial effects for the elements of $\mathbf{x}_{t1}$ from

$$N^{-1}\sum_{i=1}^{N}\Phi(\hat{\lambda}_{\omega t} + \mathbf{x}_{t1}^{o}\hat{\boldsymbol{\beta}}_{\omega 1} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{\omega 1} + \hat{\gamma}_{\omega 1}\hat{v}_{it2}) \tag{5.25}$$

or

$$N^{-1}\sum_{i=1}^{N}\Phi[(\hat{\lambda}_{\omega t} + \mathbf{x}_{t1}^{o}\hat{\boldsymbol{\beta}}_{\omega 1} + \bar{\mathbf{z}}_i\hat{\boldsymbol{\xi}}_{\omega 1})/(1 + \hat{\gamma}_{\omega 1}^{2}\hat{\tau}_{2}^{2})^{1/2}] \tag{5.26}$$

Formula (5.26) uses the normality of v_{t2} while (5.25) does not. Plus, equation (5.25) allows for $\mathrm{Var}(v_{it2})$ to change over time. This is easily accomodated in (5.26) by estimating a different error variance from (5.24) for each t, say $\hat{\tau}_{t2}^{2}$, and using this in place of the pooled estimate. Also, we can easily allow nonstationarity in the covariances in (5.21) by replacing $\gamma_{\omega 1}$ with $\gamma_{t\omega 1}$ for each t. In fact, the key features of the second step estimation are that including $\bar{\mathbf{z}}_i$ appropriately controls for the time-constant heterogeneity and that including $\hat{v}_{it2}$ accounts for the endogeneity of y_{it2}. We could replace (5.23) with more general models of $E(y_{t1}|\mathbf{z}_{t1},y_{t2},\bar{\mathbf{z}},v_{t2})$, for example, allow a heteroskedastic probit model with variance depending on $(\bar{\mathbf{z}},v_{t2})$ as in Section 4.1.

A strength of the previous approach is that it allows for abitrary serial dependence in (u_{it1},u_{it2}). This is often important in models where lagged dependent variables are not included among the explanatory variables.

Reporting estimates of APEs from nonlinear panel data models, ideally along with standard errors, is important in order to compare magnitudes of effects with more commonly applied linear models. Even when y_{t1} is discrete, or has clear upper or lower bounds, it is still common to estimate linear models with additive, unobserved effects by instrumental variables, where the partial effects are assumed to be constant. To know what the nonlinear model is buying us, we need to compute APEs at various values of the explanatory variables. Typically, when we use the mean values, we get similar estimates of partial effects from linear and nonlinear models. Of interest is what happens to the APEs in the nonlinear models as the explanatory variables are varied.

6. APPLICATION: EFFECTS OF SPENDING ON STUDENT PERFORMANCE

(To be added.)

7. CONCLUSIONS AND CAVEATS

(To be added.)

REFERENCES

Angrist, J.D. (1991), "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," National Bureau of Economic Research Technical Working Paper Number 115.

Bartle, R.G. (1966), *The Elements of Integration*. New York: John Wiley & Sons.

Blundell, R. and J.L. Powell (2000), "Endogeneity in Nonparametric and Semiparametric Regression Models," mimeo, Department of Economics, University College London.

Chamberlain, G. (1980), "Analysis of Covariance with Qualitative Data," *Review of Economic Studies* 47, 225-238.

Chamberlain, G. (1984), "Panel Data," in: Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Volume II. Amsterdam: North Holland, 1247-1318.

Gourieroux, C., A. Monfort, and A. Trognon (1984), "Pseudo-Maximum Likelihood Methods: Theory," *Econometrica* 52, 681-700.

Greene, W.H. (2000), *Econometric Analysis*, fourth edition. New York: Macmillan.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametic Estimation of Average Treatment Effects," *Econometrica* 66, 315-331.

Heckman, J.J., H. Ichimura, and P. Todd (1997), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261-294.

Heckman J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, J. Heckman and B. Singer (eds.), 156-245. New York: Wiley.

Horowitz, J.L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* 60, 505-531.

Hsiao, C. (1986), *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

Manski, C.F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3, 205-208.

Manski, C.F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics* 27, 313-333.

Mundlak, Y. (1978), "On the Pooling of Time Series and Cross Section Data," *Econometrica* 46, 69-85.

Newey, W.K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Volume 4, R.F. Engle and D. McFadden (eds.). Amsterdam: North Holland, 2111-2245.

Papke, L.E. (2001), "The Effects of Spending on Test Pass Rates: Evidence from Michigan," mimeo, Michigan State University Department of Economics.

Papke, L.E. and J.M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates," *Journal of Applied Econometrics* 11, 619-632.

Rivers, D. and Q.H. Vuong (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics* 39, 347-366.

Robinson, P.M. (1988), "Root-*N*-Consistent Semiparametric Regression," *Econometrica* 56, 931-954.

Rosenbaum, P.R. and D.B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, 41-55.

Smith, R. and R. Blundell (1986), "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica* 54, 679-685.

Wooldridge, J.M. (1997), "Quasi-Likelihood Methods for Count Data," in *Handbook of Applied Econometrics*, Volume 2, M.H. Pesaran and P. Schmidt (eds.). Oxford: Blackwell, 352-406.

Wooldridge, J.M. (2000a), *Introductory Econometrics: A Modern Approach*. Cincinnati, OH: South-Western.

Wooldridge, J.M. (2000b), "The Initial Conditions Problem for Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity," Working Paper No. 00/496, University of Bristol Department of Economics.

Wooldridge, J.M. (2001), *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Yatchew, A. and Z. Griliches (1985), "Specification Error in Probit Models" *Review of Economics and Statistics* 67, 134-139.