Identification and estimation of 'irregular' correlated random coefficient models[1]

Bryan S. Graham[†] and James L. Powell[◇]

Initial Draft: December 2007

This Draft: May 26, 2008

Preliminary and Incomplete

Comments Welcome

Abstract

In this paper we study identification and estimation of the causal effect of a small change in an endogenous regressor on a continuously-valued outcome of interest using panel data. We focus on average effects, over either the full population distribution of unobserved heterogeneity (the average partial effect, APE), or over subpopulations defined by their regressor values (local average responses, LARs). In our primary model the outcome of interest varies linearly with a (scalar) regressor, but with an intercept and slope coefficient that may vary across units in a way which depends on the regressor. Our model is a special case of Chamberlain's (1992a) correlated random coefficients (CRC) model, but not does not satisfy the regularity conditions he imposes. We show how two measures of the outcome and regressor for each unit are sufficient for identification of the APE and LAR, as well as aggregate trends that are mean-independent of the regressors. We identify aggregate trends using units with a zero first difference in the regressor, in the language of Chamberlain (1980b, 1982) 'stayers', and the average partial effect using units with non-zero first differences or 'movers'. We discuss extensions of our approach to models with multiple regressors and more than two time periods, periods. We use our methods to estimate the average elasticity of calorie consumption with respect to total outlay for a sample of poor Nicaraguan households.

JEL Classification: C14, C23, C33

Key Words: Panel Data, Correlated Random Coefficients, Control Function, Average Partial Effects, Local Average Response

---

[1] We would like to thank seminar participants at UC - Berkeley, UCLA, USC and Harvard, members of the Berkeley Econometrics Reading Group, Gary Chamberlain and Michael Jansson for helpful discussions. All the usual disclaimers apply.

[†] Department of Economics, University of California - Berkeley, 508-1 Evans Hall #3880, Berkeley, CA 94720 and National Bureau of Economic Research. E-mail: `bgraham@econ.berkeley.edu`. Web: `http://www.econ.berkeley.edu/~bgraham/`.

[◇] Department of Economics, University of California - Berkeley, 508-1 Evans Hall #3880, Berkeley, CA 94720. E-mail: `powell@econ.berkeley.edu`. Web: `http://www.econ.berkeley.edu/~powell/`.

1 Introduction

That the availability of multiple observations of the same sampling unit (e.g., individual, firm, etc.) over time can help to control for the presence of unobserved heterogeneity is both intuitive and plausible. The inclusion of unit-specific intercepts in linear regression models is among the most widespread methods of 'controlling for' omitted variables in empirical work (e.g., Griliches, 1979; Currie and Thomas, 1995; Card, 1996; Altonji and Dunn, 1996). The appropriateness of this modelling strategy, however, hinges on any time-invariant correlated heterogeneity entering the outcome equation additively. Unfortunately, additivity, while statistically convenient, is difficult to motivate economically (cf., Imbens, 2007).[2] Browning and Carro (2007) present a number of empirical panel data examples where non-additive forms of unobserved heterogeneity appear to be empirically relevant.

In this paper we study the use of panel data for identifying and estimating what is arguably the simplest statistical model admitting nonseparable heterogeneity: the *correlated random coefficients* (CRC) model. Let $Z_t = (Y_t, X_t')'$ be a random variable measured in each of $t = 1, \ldots, T$ periods for N randomly sampled units. In the most basic model we analyze the structural outcome equation is given by

$$Y_t = a\left(A, U_t\right) + b\left(A, U_t\right) X_t \tag{1}$$

where Y_t is a scalar continuously-valued outcome of interest, X_t a scalar choice variable, A time-invariant unobserved unit-level heterogeneity and U_t a time-varying disturbance. Both A and U_t may be vector-valued. Equation (1) is structural in the sense that the *unit-specific* function

$$Y_t\left(x_t\right) = a\left(A, U_t\right) + b\left(A, U_t\right) x_t \tag{2}$$

traces out a unit's period t potential outcome under different hypothetical values of x_t.[3] Equation (2) differs from the the textbook linear panel data model (with unit-specific intercepts, but otherwise constant regressor coefficients) in that the effect of a small change in x_t generally varies across units.

We study estimands which characterize the effect on an exogenous change in X_t on the probability distribution of Y_t. For concreteness we focus on identification and estimation of the *average partial effect* (APE) (cf., Chamberlain, 1984; Blundell and Powell, 2003; Wooldridge, 2005a) and the *local average response* (LAR) (cf., Altonji and Matzkin, 2005; Bester and Hansen, 2007). In the binary regressor case these two objects correspond to the average treatment effect (ATE) and the average treatment effect on the treated (ATT) (cf., Florens, Heckman, Meghir and Vytlacil, 2008).[4]

The average partial effect is given by

$$\beta_t \equiv \mathbb{E}\left[\frac{\partial Y_t\left(x_t\right)}{\partial x_t}\right] = \mathbb{E}\left[b\left(A, U_t\right)\right]. \tag{3}$$

[2] Chamberlain (1984) presents several well-formulated economic models that *do* imply linear specifications with unit-specific intercepts.

[3] Throughout we use capital letters to denote random variables and lower case letters specific realizations of them.

[4] In a companion paper we study quantile partial effects (QPEs) (Graham, Hahn and Powell, 2008).

Because of linearity of (2), β_t does not depend on x_t.[5]

The local average response gives the average partial effect within a subpopulation defined by its choice of $X_t = x_t$; it is given by

$$\gamma_t(x_t) \equiv \mathbb{E}\left[\left.\frac{\partial Y_t(x_t)}{\partial x_t}\right| X_t = x_t\right] = \mathbb{E}\left[\left. b(A, U_t)\right| X_t = x_t\right]. \tag{4}$$

Identification and estimation of (3) and (4) is nontrivial because X_t may vary systematically with A and/or U_t. For example, the derivative of the regression function of Y_t given X_t does not identify $\gamma_t(x_t)$. Differentiating through the integral we have

$$\frac{\partial \mathbb{E}\left[\left. Y_t\right| X_t = x_t\right]}{\partial x_t} = \gamma_t(x_t) + \int\int \left\{a(a, u_t) + b(a, u_t)x_t\right\} \frac{\partial f\left(a, u_t\right| x_t)}{\partial x_t} \mathrm{d}m(a)\,\mathrm{d}m(u_t).$$

The second term is what Chamberlain (1982) calls heterogeneity bias.

To contextualize our contribution within the wider panel data literature it is useful to consider the more general outcome response function:

$$Y_t(x_t) = m(x_t, A, U_t).$$

Identification of the APE and LAR in the above model may be achieved by one of two main classes of restrictions. The *correlated random effects* approach invokes smoothness priors on the joint distribution of $\left.(U, A)\right| X$; with $U = (U_1, \ldots, U_T)'$ and $X = (X_1, \ldots, X_T)'$. Mundlak (1978) and Chamberlain (1980a, 1984) develop this approach for the case where $m(X_t, A, U_t)$ and $F\left(U, A\right| X)$ are parametrically specified. Newey (1994a) considers a semiparametric specification for $F\left(U, A\right| X)$ (cf., Arellano and Carrasco 2003). Recently, Altonji and Matzkin (2005) have extended this idea to the case where $m(X_t, A, U_t)$ is either semi- or non-parametric along with $F\left(U, A\right| X)$ (cf., Bester and Hansen 2007).

The *fixed effects* approach imposes restrictions on $m(X_t, A, U_t)$ and $F\left(U\right| X, A)$, while leaving $F\left(A\right| X)$, the distribution of the time-invariant heterogeneity, the so-called 'fixed effects', unrestricted. Chamberlain (1980a, 1984, 1992), Manski (1987), Honoré (1992) and Abrevaya (2000) are examples of this approach. Depending on the form of $m(X_t, A, U_t)$, the fixed effect approach may not allow for a complete characterization of the effect of exogenous changes in X_t on the probability distribution of Y_t. Instead only certain features of this relationship may be identified (e.g., ratios of the average partial effect of two regressors) (cf., Chamberlain 1984, 1992b, Arellano and Honoré 2001, Arellano 2003).

Our methods are of the 'fixed effect' variety. In addition to assuming the CRC structure for $Y_t(x_t)$ we impose a marginal stationarity restriction on $F\left(U_t\right| X, A)$, a restriction also used by Manski (1987), Honoré (1992) and Abrevaya (2000), however we leave $F\left(A\right| X)$ nonparametric. In our setup both the APE and LAR are identified when X_t is continuously-valued. In fact, we are

[5] We also study the case where X_t is itself a function of a lower-dimensional choice variable R_t. In that case, the APE, defined in terms of r_t, may vary with r_t. Extending our results to this case is straightforward.

able to provide a characterization of when these estimands are semiparametrically just-identified. In that sense, our maintained assumptions are minimally sufficient (although not necessary).[6]

Motivated by heterogeneity in the labor market returns to schooling, Card (1995, 2001) and Heckman and Vytlacil (1998) have studied identification and estimation of the CRC model using cross section data and 'instrumental variables' (cf., Garen 1984, Wooldridge 1997, 2001, 2005a). This work belongs to larger body of research on nonparametric triangular systems (e.g., Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Heckman and Vytlacil, 2001; Blundell and Powell, 2003; Imbens and Newey, 2007; Florens, Heckman, Meghir and Vytlacil, 2008).[7]

The value of panel data for identification and estimation in the CRC model is comparatively less well understood. Mundlak (1961), while primarily focusing on a constant coefficients linear panel data model with unit-specific intercepts, briefly, and verbally, refers to the CRC model (p. 45).[8] The first formal analysis of the CRC model in the context of panel data appears in Chamberlain (1980b, 1982). In later work Chamberlain (1992a, pp. 579 - 585) proposed an ingenious method-of-moments estimator for the APE, however, the regularity conditions required for his estimator, as we discuss further below, rule out substantively important economic models. Wooldridge (2005a) also analyzes a CRC panel data model. His focus is on providing conditions under which the usual linear fixed effects (FE) estimator is consistent despite the presence of correlated random coefficients (cf., Chamberlain, 1982, p. 11). Fernández-Val (2005) develops bias correction methods for the CRC model in a large-N, large-T setting.

Altonji and Matzkin (2005) and Bester and Hansen (2007) have developed new methods for using panel data to control for nonseparable unobserved heterogeneity. As their approaches are of the random effects variety, while our's are of the fixed effect variety, we view our methods as complementary to theirs.

Chernozhukov, Fernández-Val, Hahn and Newey (2007) study identification of the APE in a probit model with unit-specific intercepts (in the index function) and a discrete or continuous scalar regressor. They show that the maximum likelihood estimator which estimates the unit-specific intercepts along with the coefficient on X_t can be used to construct bounds on the ATE despite the incidental parameters problem (cf., Hahn, 2001). Porter (1996) and Das (2003) study nonparametric estimation of panel data model with additive unobserved heterogeneity. Honoré (1992) and Abrevaya (2000) consider models with nonseparable heterogeneity but, like Manski (1987), only identify index coefficients, not the APE or LAR.

The next section reports identification results for the APE and LAR in a two period version of our core model. When X_t is discretely-valued our assumptions generally only bound the APE and LAR (appropriately defined to account for the discreteness of X_t). Our analysis of this case also suggests useful interpretations of the probability limits of the linear fixed effects (FE) estimator and the 'difference-in-differences' (DID) estimator of the program evaluation literature (Card, 1990;

[6] Chesher (2007) provides an extended discussion of the value of 'just identifying' semiparametric restrictions.

[7] Much of this research is surveyed by Imbens (2007).

[8] The exact reference is "The key to the estimation of the [average] slope of the infrafirm function is have at least two points of data on each f_i. In this case it is possible to get the slope of each of the lines f_i, average them and get the final estimate. That requires a combination of time series and cross section data."

Meyer, 1995; Angrist and Krueger, 1999; Athey and Imbens, 2006). When X_t is continuously valued the APE and LAR are point identified. We also contrast our approach to identification with the semiparametric random effects methods employed by Altonji and Matzkin (2005).

Section 3 details our estimation approach. We begin with a discussion of the two period case. We then introduce a general multiple period CRC model and discuss its estimation. In that section we also relate our results to those of Chamberlain (1992a). Under Chamberlain's (1992a) conditions, which are not satisfied in our leading example, the APE is estimable at parametric rates. In contrast, our estimator has asymptotic properties similar to a standard one-dimensional kernel regression problem.

In Section 5, we use our methods to estimate the average elasticity of calorie demand with respect to total household resources in a sample of poor rural communities in Nicaragua. Our sample is drawn from a population that participated in a pilot of the conditional cash transfer program Red de Protección Social (RPS). Hunger, conventionally measured, is widespread in the communities from which our sample is drawn; we estimate that almost forty percent of households have less then the required number of calories needed for all their members to engage in 'light activity' on a daily basis.[9]

Worldwide, the Food and Agricultural Organization (FAO) estimates that 854 million people suffered from protein-energy malnutrition in 2001-03 (FAO, 2006). Halving this number by 2015, in proportion to the world's total population, is the first United Nations Millennium Development Goal. Chronic malnutrition, particularly in early childhood, may adversely affect cognitive ability and economic productivity in the long-run (e.g., Dasgupta, 1993; Grantham-McGregor and Baker-Henningham, 2005; Case and Paxson, 2006; Hoddinott et al., 2008). A stated goal of the RPS program is to reduce childhood malnutrition, and consequently increase human capital, by directly augmenting household income in exchange for regular school attendance and participation in preventive health care check-ups.

The efficacy of this approach to reducing childhood malnutrition largely depends on the size of the average elasticity of calories demanded with respect to income across poor households.[10] While most estimates of the elasticity of calorie demand are significantly positive, several recent estimates are small in value, casting doubt on the value of income-oriented anti-hunger programs (Behrman and Deolalikar, 1987; Strauss and Thomas, 1995; Subramanian and Deaton, 1996; Hoddinott, Skoufias and Washburn, 2000). Wolfe and Behrman (1983), using data from pre-revolutionary Nicaragua, estimate a calorie elasticity of just 0.01. Their estimate, if accurate, suggests that the income supplements provided by the RPS program should have little effect on caloric intake.

Disagreement about the size of the elasticity of calorie demand has prompted a vigorous method-

[9] We use Food and Agricultural Organization (FAO, 2001) gender- and age-specific energy requirements for 'light activity', as reported in Appendix 8 of Smith and Subandoro (2007), and our estimates of total calories available at the household-level to calculate the fraction of households suffering from 'food insecurity'. This approach to measuring food insecurity is not without its critics (e.g., Edmundson and Sukhatme 1990). Ferro-Luzzi (2005) provides a historical and conceptual overview of FAO/WHO food energy recommendations.

[10] Another motivation for studying this elasticity has to do with its role in theoretical models of nutrition-based poverty traps (e.g., Mirlees, 1975; Stiglitz, 1976; Bliss and Stern, 1978; Dasgupta and Ray, 1986, 1987).

ological debate in development economics. Much of this debate has centered, appropriately so, on issues of measurement and measurement error (e.g., Behrman and Deolalikar, 1987; Bouis and Haddad, 1992; Bouis, 1994; Subramanian and Deaton 1996). The implications of household-level correlated heterogeneity in the underlying elasticity for estimating its average, in contrast, have not been examined. If, for example, a households' food preferences, or preferences towards child welfare, co-vary with those governing labor supply, then its elasticity will be correlated with total household resources. An estimation approach which presumes the absence of such heterogeneity will generally be inconsistent for the parameter of interest. Our statistical model and corresponding estimator provides an opportunity, albeit in a specific setting, for assessing the relevance these types of heterogeneities.

Section 5.3 summarizes and suggests areas for further research.

2 Identification: the two period case with a scalar regressor

We illustrate each of our main identification results for the case where X_t is scalar and $T = 2$. We generalize to panels are arbitrary length and multiple regressors in Section 3 below. Our first assumption is that the data generating process takes a correlated random coefficients form.

Assumption 2.1 (Correlated Random Coefficients)

$$Y_t = a\left(A, U_t\right) + b\left(A, U_t\right) X_t.$$

Our second key identifying assumption is marginal stationarity of the time-varying unobserved heterogeneity, U_t.

Assumption 2.2 (Marginal Stationarity) *(i)*

$$U_t | \, X, A \overset{D}{=} U_s | \, X, A, \qquad t \neq s,$$

(ii) the distribution of U_t given X and A is non-degenerate for all $(X, A) \in \mathcal{X} \times \mathcal{A}$.

Assumption 2.2 does not restrict the conditional distribution of A given X. In this sense A is a 'fixed effect'. Nevertheless Assumption 2.2, while allowing for serial dependence in U_t and certain forms of heteroscedasticity, is restrictive. For example it rules out heteroscedasticity over time (cf., Arellano 2003).

To formally close the model we make the following sampling assumption:

Assumption 2.3 (Random Sampling) *$\{(X_{1i}, X_{2i}, Y_{1i}, Y_{2i}, A_i)\}_{i=1}^{\infty}$ is an independently and identically distributed random sequence drawn from the distribution F_0.*

Let $X = (X_1, X_2)'$ and

$$\beta_t\left(x\right) \equiv \mathbb{E}\left[b\left(A, U_t\right) | \, X = x\right]$$

denote the average effect of a small change in x_t within the subpopulation of units where $X = x = (x_1, x_2)'$. Observe that $\beta_t(x)$, while closely related, is the distinct from the LAR. It gives the average effect within a subpopulation defined by a common complete *history* of choices for X_t. Our first result shows that $\beta_t(x)$ is just-identified when $x_1 \neq x_2$.

Proposition 2.1 *Under Assumptions 2.1, 2.2 and 2.3* $\beta_1(x) = \beta_2(x) = \beta(x)$ *is just-identified by the ratio*

$$\beta(x) = \frac{\mathbb{E}[Y_2|X = x] - \mathbb{E}[Y_1|X = x]}{x_2 - x_1} \tag{5}$$

for all $x \in \{x : x \in \mathcal{X},\ x_1 \neq x_2\}$.

Proof. Under Assumptions 2.1 and 2.3 we have

$$\begin{aligned}
\mathbb{E}[Y_1|X] &= \alpha_1(X) + \beta_1(X)X_1 \\
\mathbb{E}[Y_2|X] &= \alpha_2(X) + \beta_2(X)X_2,
\end{aligned}$$

for $\alpha_t(X) = \mathbb{E}[a(A, U_t)|X]$ and $\beta_t(X) = \mathbb{E}[b(A, U_t)|X]$. Iterated expectations (which is allowable by part (ii) of Assumption 2.2), marginal stationarity (part (i) of Assumption 2.2) and time-invariance of A give

$$\beta_t(X) = \mathbb{E}[b(A, U_t)|X] = \mathbb{E}[\mathbb{E}[b(A, U_t)|X, A]|X] = \mathbb{E}\left[\left.\widetilde{b}(X, A)\right|X\right] = \beta(X),$$

for $\widetilde{b}(X, A) = \mathbb{E}[b(A, U_t)|X, A]$. This gives $\beta_1(X) = \beta_2(X) = \beta(X)$; a similar calculation gives $\alpha_t(X) = \mathbb{E}[a(A, U_t)|X] = \alpha(X)$. Taking differences across time periods and solving for $\beta(X)$ then gives (5). That $\beta(x)$ is just-identified follows directly from its definition as a conditional expectation function, linearity of Y_t in $a(A, U_t)$ and $b(A, U_t)$, and just-identification of $\mathbb{E}[Y_1|X]$ and $\mathbb{E}[Y_2|X]$. ■

To recover the APE we average $\beta(X)$ over the marginal distribution of X:

$$\beta = \mathbb{E}[\beta(X)].$$

Since $\beta(x)$ is only identified on those points of the support of X for which $X_1 \neq X_2$ (i.e., for 'movers' or units which alter their choice of X_t across periods) we cannot, in general, calculate $\mathbb{E}[\beta(X)]$ without further assumptions (Chamberlain 1982, p. 13). Consequently, unless all units change their value of X_t across periods, the APE is not identified. When X_t is discrete it is natural to construct bounds for β or to compute the average of $\beta(X)$ among 'movers'. The latter approach is particularly simple and foreshadows our approach to estimation in the continuous case. When X_t is continuous we impose smoothness restrictions on $\beta(x)$ which are sufficient to point identify β. We consider each case in turn.

Discrete regressor If $X_t \in \{0, \ldots, M\}$, then $\beta(x)$ is only identified for the $M(M+1)$ possible sequences of $x = (x_1, x_2)$ where $x_1 \neq x_2$. Although the APE is not identified, we can compute the average partial effect in the subpopulation of units who *change* their values of X_t across the two periods (Chamberlain 1980b, 1982). Define, invoking marginal stationarity, the 'movers' average partial effect (MAPE) as

$$\beta^M \equiv \mathbb{E}\left[b\left(A,U_t\right)\middle|\,\Delta X \neq 0\right] = \frac{\mathbb{E}\left[\mathbf{1}\left(\Delta X \neq 0\right)\beta\left(X\right)\right]}{\mathbb{E}\left[\mathbf{1}\left(\Delta X \neq 0\right)\right]}. \tag{6}$$

Expression (6) is implicit in Chamberlain (1982, p. 13) who also noted that we have no information on $\beta^S = \mathbb{E}\left[b\left(A,U_t\right)\middle|\,\Delta X = 0\right]$, or the 'stayers' average partial effect (SAPE). The data are consistent with β^S taking on any feasible value. When Y is continuously-valued along the real line, then any value for $\beta = \mathbb{E}\left[\beta\left(X\right)\right]$ is consistent with any given value for β^M. However, if Y has bounded support then β^M can be used to construct sharp bounds on β using the general approach of Manski (2003).

Consider, to illustrate the main ideas, the case where X_t is binary, for example an indicator for union membership, and $Y_t \in \left[\underline{y}, \overline{y}\right]$.[11] In that case the APE is given by

$$\beta = \sum_{i,j=0,1} \pi_{ij}\beta\left(i,j\right),$$

where $\pi_{ij} = \Pr\left(X_1 = i, X_2 = j\right)$. Recall, however, that $\beta\left(0,0\right)$ and $\beta\left(1,1\right)$, the partial effects associated with the two types of stayers are only partially identified with respective identification regions of[12]

$$\mathbb{H}\left\{\beta\left(0,0\right)\right\} = \left[\underline{y} - \mathbb{E}\left[Y_t\middle|\,X_1 = 0, X_2 = 0\right], \overline{y} - \mathbb{E}\left[Y_t\middle|\,X_1 = 0, X_2 = 0\right]\right]$$

and

$$\mathbb{H}\left\{\beta\left(1,1\right)\right\} = \left[\mathbb{E}\left[Y_t\middle|\,X_1 = 1, X_2 = 1\right] - \overline{y}, \mathbb{E}\left[Y_t\middle|\,X_1 = 1, X_2 = 1\right] - \underline{y}\right].$$

The identified set for the APE, β, is thus

$$\begin{aligned}
&\mathbb{H}\left\{\beta\right\} \\
&\in \left[\pi_{00}\left(\underline{y} - \mathbb{E}\left[Y_t\middle|\,X_1 = 0, X_2 = 0\right]\right) + \pi_{01}\beta\left(0,1\right) + \pi_{10}\beta\left(1,0\right) + \pi_{11}\left(\mathbb{E}\left[Y_t\middle|\,X_1 = 1, X_2 = 1\right] - \overline{y}\right),\right. \\
&\quad \left.\pi_{00}\left(\overline{y} - \mathbb{E}\left[Y_t\middle|\,X_1 = 0, X_2 = 0\right]\right) + \pi_{01}\beta\left(0,1\right) + \pi_{10}\beta\left(1,0\right) + \pi_{11}\left(\mathbb{E}\left[Y_t\middle|\,X_1 = 1, X_2 = 1\right] - \underline{y}\right)\right].
\end{aligned}$$

[11] Note that for X_t binary the CRC structure is unrestrictive (although our marginal stationarity assumption is restrictive). In independent work Chernozhukov, Fernández-Val, Hahn and Newey (2007) develop bounds for β for the case where Y_t is binary. They show that if, in our notation, $\mathbb{E}\left[Y_t\left(x_t\right)\middle|\,A = a\right] = \Phi\left(x_t'\gamma + a\right)$, tighter bounds are available.

[12] Observe that, from equation (2),

$$\beta\left(0,0\right) = \mathbb{E}\left[Y_t\left(1\right)\middle|\,X_1 = 0, X_2 = 0\right] - \mathbb{E}\left[Y_t\left(0\right)\middle|\,X_1 = 0, X_2 = 0\right]$$

The data reveal

$$\mathbb{E}\left[Y_t\left(0\right)\middle|\,X_1 = 0, X_2 = 0\right] = \mathbb{E}\left[Y_t\middle|\,X_1 = 0, X_2 = 0\right]$$

but do not reveal $\mathbb{E}\left[Y_t\left(1\right)\middle|\,X_1 = 0, X_2 = 0\right]$, which may lie anywhere in the interval $\left[\underline{y}, \overline{y}\right]$.

The width of $\mathbb{H}\{\beta\}$ is $(\pi_{00}+\pi_{11})\left(\overline{y}-\underline{y}\right)$: the more 'stayers' the less informative the data are for β. Unfortunately a preponderance of stayers is common in many microeconometric applications. In Card's (1996) analysis of the union wage premium, less than 10 percent of workers switch between collective bargaining coverage and non-coverage across periods (Table V, p. 971). In such cases β^M is an average over a very particular population, while bounds on β will be quite wide. When X_t is discrete, however, this is the very best we can do without invoking additional assumptions.[13]

Before turning to the continuous case we briefly discuss identification of the LAR when X_t is discrete. As with the APE, the LAR is generally not identified. Instead we can identify the 'movers' local average response (MLAR). This is given by

$$\gamma_t^M(x_t) \equiv \mathbb{E}\left[b(A,U_t)|\, X_t = x_t, \Delta X \neq 0\right] = \frac{\mathbb{E}\left[\mathbf{1}(X_t = x_t)\,\mathbf{1}(\Delta X \neq 0)\,\beta(X)\right]}{\mathbb{E}\left[\mathbf{1}(X_t = x_t)\,\mathbf{1}(\Delta X \neq 0)\right]}. \tag{7}$$

It is straightforward to construct bounds on $\gamma_t(x_t)$ along the lines of those given for β above.

Continuous regressor When X is continuous the set $\{x : x \in \mathcal{X}, \quad x_1 = x_2\}$ will generally be of measure zero. This suggests that, under mild smoothness conditions, $\beta(x)$ should be identifiable for all $x \in \mathcal{X}$. In particular, at those points where $x_1 = x_2$, we can then identify $\beta(x)$ by the limit

$$\beta(x_1, x_1) = \lim_{h \downarrow 0} \frac{\mathbb{E}\left[Y_2|\, X = (x_1, x_1 + h)\right] - \mathbb{E}\left[Y_1|\, X = (x_1, x_1)\right]}{h}. \tag{8}$$

A sufficient condition for the above limit to exist is:

Assumption 2.4 (Smoothness) *$\beta(x)$ is continuous and differentiable in $\mathcal{X}$.*

Under this smoothness restriction we have the following Theorem.

Theorem 2.1 (Identification) *If X_t is continuously-valued and Assumptions 2.1, 2.2, 2.3 and 2.4 hold, then $\beta_t(x_t) = \beta$ and $\gamma_t(x_t)$ are identified by*

$$\beta = \mathbb{E}\left[\beta(X)\right], \qquad \gamma_t(x_t) = \mathbb{E}\left[\beta(X)|\, X_t = x_t\right]$$

with $\beta(x)$ given by (5) or (8) as appropriate.

Proof. Straightforward and therefore omitted. ■

Observe that $\beta(x)$ is an average over the conditional distribution of (A, U_t) given X. Thus smoothness of $\beta(x)$ suggests that the distribution function of A given $X = x$ is smooth in x. Such

[13] As an example of an informative and potential plausible additional restriction assume that selection into collective bargaining coverage implies that

$$\begin{aligned}\beta(0,0) &\leq \beta(0,1) \leq \beta(1,1) \\ \beta(0,0) &\leq \beta(1,0) \leq \beta(1,1),\end{aligned}$$

or that the return to selecting $X_t = 1$ is highest, on average, for the subpopulation of units that make this choice in both periods and lowest in the subpopulation that never choose $X_t = 1$.

smoothness conditions are often implied by correlated random effect specifications for A. A fixed effects purist could thus call our model (when X_t is continuous) a correlated random effects one. We maintain the fixed effects characterization because we view Assumption 2.4 as rather weak. In anycase estimation would be impossible without it.

2.1 Aggregate time effects

Although the APE is only partially identified when X_t is discrete and 'just-identified' when X_t is continuous, our CRC model nevertheless has testable implications. In particular the CRC outcome response and marginal stationarity imply that:

$$\mathbb{E}\left[\Delta Y \mid X = x\right] = \mathbb{E}\left[\Delta Y \mid X = x'\right] = 0,$$

where x and x' denote two different types of 'stayers':

$$\left\{x, x' : x, x' \in \mathcal{X}, \quad x_1 = x_2, \quad x_1' = x_2', \quad x_1 \neq x_1'\right\}.$$

Linearity of $Y_t(x)$ in x, and constancy of the conditional mean of the random slopes over time (given all leads and lags of X_t) means that outcome changes for stayers are driven soley by changes in $a(A, U_t)$ over time. Since marginal stationarity also implies constancy of the conditional mean of $a(A, U_t)$, however, our model implies that, on average, outcomes do not change across periods for stayers. Since, when X_t is continuous, there may be many types of stayers, corresponding to different values of x_2 (with $x_1 = x_2$), our set-up generates many testable restrictions.

We can use these extra model restrictions to incorporate aggregate time effects into our model in a fairly flexible way. Consider the model

$$Y_t = a_t(A, U_t) + b(A, U_t) X_t, \tag{9}$$

where $a_t(A, U_t)$ the mapping from A and U_t into the intercept is time-specific but restricted by the following assumption:

Assumption 2.5 (Conditional Common Average Trends)

$$\mathbb{E}\left[a_2(A, U_t) - a_1(A, U_t) \mid X\right] = \delta(X_2).$$

Assumption 2.5 allows for heterogeneity in the period two aggregate time shock across units. In particular, the average shock may differ across subpopulations defined in terms of their period two choice. For example, if X_t denotes union membership, then Assumption 2.5 allows for the period two shock to affect mean earnings in the union and non-union sectors differently.

Let $x = (x_1, x_2)'$ with $x_1 \neq x_2$ and $x' = (x_1', x_2')'$ with $x_1' = x_2' = x_2$. Observe that

$$\mathbb{E}\left[Y_2 \mid X = x'\right] - \mathbb{E}\left[Y_1 \mid X = x'\right] = \delta(x_2),$$

while

$$\mathbb{E}\left[Y_2\right| X=x] - \mathbb{E}\left[Y_1\right| X=x] = \delta\left(x_2\right) + \beta\left(x\right)\left(x_2 - x_1\right).$$

We therefore have

$$\beta\left(x\right) = \frac{\mathbb{E}\left[Y_2\right| X=x] - \mathbb{E}\left[Y_1\right| X=x] - \left\{\mathbb{E}\left[Y_2\right| X=x'] - \mathbb{E}\left[Y_1\right| X=x']\right\}}{x_2 - x_1} \tag{10}$$

We may adapt expression (8) to get $\beta(x)$ for stayers. With $\beta(x)$ identified, identification of the APE and LAR follows directly.

Assumption 2.5 is a substantial generalization of the deterministic 'common trends' assumption routinely made in program evaluation studies (Heckman and Robb, 1985; Meyer, 1995; Angrist and Krueger, 1999). In that literature Assumption 2.5 is invoked with the additional the requirement that $\delta(X_2) \equiv \delta$ is constant in X_2; which is an 'unconditional' common average trends assumption.

For estimation purposes it is convenient to assume a parametric form for $\delta(X_2)$. A natural specification, given the CRC form for the outcome response, is $\delta(X_2) = \delta_a + \delta_b X_2$. Note that this specification for $\delta(X_2)$ is isomorphic to the pair of restrictions

$$\mathbb{E}\left[a_2\left(A, U_t\right) - a_1\left(A, U_t\right)\right| X] = \delta_a, \quad \mathbb{E}\left[b_2\left(A, U_t\right) - b_1\left(A, U_t\right)\right| X] = \delta_b,$$

in the model where $b_t(A, U_t)$ the mapping from A and U_t into the slope is also time-specific (i.e., $Y_t = a_t(A, U_t) + b_t(A, U_t) X_t$). Put differently, linearity of $\delta(X_2)$ is equivalent to a CRC model which allows for aggregate common intercept and slope drift across periods. Such a model allows for a fairly flexible pattern of heterogeneous macroeconomic shocks over time, while at the same time remaining easy to interpret and, importantly, easy to estimate. At the same time it provides a set of testable restrictions which may be used to judge model adequacy. Namely that for any two stayers with $X = x'$ and $X = x''$ we have

$$\frac{\mathbb{E}\left[\Delta Y\right| X=x''] - \mathbb{E}\left[\Delta Y\right| X=x']}{x_1'' - x_1'} = \delta_b.$$

We work with this specification for $\delta(X_2)$ for the next subsection and with the even simpler constant in X_2 aggregate effect $\delta(X_2) = \delta$ in our initial discussion of estimation. We return to more general models for aggregate time effects in Section 4 below.

2.2 Relationship to linear 'fixed effects' (FE) estimator

Our model can be used to provide a representation of the probability limit of the textbook FE estimator under CRC misspecification. Assume that the researcher posits a model of

$$Y_t = \delta_t + \beta X_t + A + U_t, \qquad \mathbb{E}\left[U_t\right| A, X] = 0, \quad t = 1, 2 \tag{11}$$

when in fact the true model is as described by Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 with $\delta(X_2) = \delta_a + \delta_b X_2$.

In the $T = 2$ case the linear FE estimator has a probability limit equal to the coefficient, b^{FE}, on ΔX in the (mean squared error minimizing) linear predictor of ΔY given ΔX. It is straightforward to show that

$$b^{FE} = \beta + \delta_b \left\{ 1 - \frac{\mathbb{C}(X_1, X_2)}{\mathbb{V}(\Delta X)} \right\} + \mathbb{E}\left[\omega\left(\Delta X\right)\left(\beta\left(X\right) - \beta\right)\right], \qquad \omega\left(\Delta X\right) = \frac{\Delta X\left(\Delta X - \mathbb{E}\left[\Delta X\right]\right)}{\mathbb{V}(\Delta X)}. \tag{12}$$

The first term in (12) reflects the failure of the textbook model to account for aggregate slope drift, while the second is due to its failure to account for slope heterogeneity. This second term is similar to the local average treatment effect (LATE) representation of the Wald-IV estimator's probability limit (Imbens and Angrist, 1994; Angrist, Imbens and Rubin, 1996; Imbens, 2007). If slope drift is not a concern (i.e., $\delta_b = 0$), we can view b^{FE} as a movers weighted average partial effect since $\mathbb{E}\left[\omega\left(\Delta X\right)\right] = 1$ and $\omega\left(0\right) = 0$. An important difference between (12) and the LATE is that 'movers', unlike 'compliers', can be directly identified from the data. Consequently the weights in (12) are estimable.

To get a sense of whether b^{FE} is likely to be interpretable it is helpful to consider some stylized examples. For simplicity assume we assume, for the remainder of this subsection, the absence of slope drift (i.e., that $\delta_b = 0$). If X_1 and X_2 are independent and identically distributed normal random variables, then $\omega\left(\Delta X\right)$ will be a χ_1^2 random variable and b^{FE} will be 'dominated' by those few units with very large values of ΔX. This suggests that b^{FE} will be more representative of the partial effect of those units who change their choice of X_t dramatically across periods.

The binary X_t case is also informative. Let π_{ij} denote the probability that $X_1 = i$ and $X_2 = j$ (with $i, j \in \{0, 1\}$), we can show that

$$b^{FE} = \omega\left(-1\right)\beta\left(0,1\right) + \left(1 - \omega\left(-1\right)\right)\beta\left(1,0\right), \qquad \omega\left(-1\right) = \frac{\pi_{01}\left(1 - \pi_{01} + \pi_{10}\right)}{\pi_{01}\left(1 - \pi_{01}\right) + \pi_{10}\left(1 - \pi_{10}\right) + 2\pi_{01}\pi_{10}}$$

which is a weighted average of the average partial effect of those units who 'move' from $X_1 = 0$ to $X_2 = 1$ and those who move from $X_1 = 1$ to $X_2 = 0$. If $\pi_{10} = \pi_{01}$ such that $\mathbb{E}\left[\Delta X\right] = 0$, then $b^{FE} = \beta^M$, however, in general the two estimands will differ (this equality also holds when (11) does not include time-specific intercepts).[14]

The linear FE estimator is especially interpretable in the 'classical' difference-in-differences (DID) set-up. In that setting there are two sets of regions. In both sets of regions the program is unavailable in period one. In treatment regions it becomes available in period two, while in control regions it remains unavailable. In that case $\pi_{10} = 0$ and it is easy to see that

$$b^{FE} = \beta\left(0,1\right),$$

which also equals the average treatment effect on the treated (ATT).

Wooldridge (2005b), who maintains the CRC structure as we do, imposes the additional re-

[14] In independent work, Chernozhukov, Fernández-Val, Hahn and Newey (2007, Section 3.3) obtain a related result in the context of a fixed effects binary choice model (without time effects).

striction (in our notation) that $\beta(x) = \mathbb{E}[b(A, U_t)]$ for $x_1 \neq x_2$ (cf., Equation (14) on p. 387).[15] In that case equivalency of the FE probability limit and the APE follows directly by the property that $\mathbb{E}[\omega(\Delta X)] = 1$. Chamberlain (1982, p. 11) makes a similar point. He notes, again in our notation, that if $\mathbb{C}(b(A, U_1), X_1) = \mathbb{C}(b(A, U_2), X_2)$, then $\mathbb{E}^*[\beta(X)|\Delta X] = \mathbb{E}[\beta(X)]$ so that $\mathbb{E}[\beta(X)|\Delta X] = \mathbb{E}[\beta(X)]$. Iterated expectations applied to (12) then gives the equality $b^{FE} = \mathbb{E}[\beta(X)]$.

While, covariance stationarity of the random slopes may be plausible in some settings, it will strain credibility in others. Consider a government which allocates a certain program across regions. Assume that initially, in period 1, the program is regressively targeted in the sense that it is placed in those regions with where returns, $b(A, U_1)$, are low, while in period 2 targeting takes an opposite, progressive form. In that case $\mathbb{C}(b(A, U_1), X_1) < 0 < \mathbb{C}(b(A, U_2), X_2)$ and $b^{FE} \neq \mathbb{E}[\beta(X)]$. This example may be of more than intellectual interest: policy 'experiments' are often associated with changes of government or legislation that involves alterations of the implicit targeting rule (e.g., Duflo, 2001). However, in other cases, covariance stationarity may be reasonable. For example, the pattern of selection into unions is plausibly stable across two adjacent years with similar macroeconomic conditions (as in Card, 1996). In any case, our approach does not require these types of restrictions.

2.3 Relationship to semiparametric correlated random effects methods

Altonji and Matzkin (2005) also study semiparametric panel data models. They work with the general model given by

$$Y_t = m_t(X_t, A, U_t) \tag{13}$$

and the following exchangeability assumption:

Assumption 2.6 (Exchangeability) *(i)*

$$A, U_t | X_1, \ldots X_T \overset{D}{=} A, U_t | X_{p(1)}, \ldots X_{p(T)},$$

for $p(t) \in \{1, \ldots, T\}$, $p(t) \neq p(t')$, *(ii) the distribution of* (A, U_t) *given* X *is non-degenerate for all* $X \in \mathcal{X}$.

Observe that Assumption 2.6, unlike Assumption 2.2 above, *does* restrict the conditional distribution of A given X. Under Assumption 2.6 Altonji and Matzkin (2005, pp. 1062 - 3) show that the Fundamental Theorem of Symmetric Functions and the Weierstrass Approximation Theorem imply the distributional equality

$$A| X_1, \ldots X_T \overset{D}{=} A|\zeta_1(X), \ldots, \zeta_T(X),$$

[15] Wooldridge (2005a) also assumes that the correlated random coefficients are time invariant.

where $\zeta_t(X)$ is the t^{th} elementary symmetric polynomial on X.[16] Because Assumption 2.6 is not sufficient to identify $\beta_t(x)$ Altonji and Matzkin (2005, pp. 1063 - 4) suggest either further restricting the conditional distribution of (A, U_t) given X or the form of the structural outcome equation.[17]

Following their second suggestion, the imposition of our CRC structure on (13) and Assumption 2.6 implies that

$$\begin{aligned}\mathbb{E}[Y_t|X] &= \alpha_t(X) + \beta_t(X) X_t \\ &= \alpha_t(\zeta_1(X), \zeta_2(X)) + \beta_t(\zeta_1(X), \zeta_2(X)) X_t,\end{aligned}$$

for $t = 1, 2$.

Now consider x and x' such that $x_1 = x_2'$ and $x_2 = x_1'$ with $x_1 \neq x_2$ (i.e., x' is a permutation of x). It is easy to show that $\beta_t(x)$ is identified by

$$\beta_t(x) = \frac{\mathbb{E}[Y_t|X = x] - \mathbb{E}[Y_t|X' = x']}{x_t - x_t'}.$$

Exchangeability and the CRC structure are sufficient to identify $\beta_t(x)$ even if the outcome variable is only observed for a single period as along as X_t is observed in each period. Altonji and Matzkin (2005, p. 1065 - 66) argue that this feature of their approach is particularly attractive in the context of sibling studies where the outcome (e.g., wages) may only be observed for a single older sibling, while the endogenous regressor (e.g., school quality) might be measured for younger as well as older siblings. In contrast, our approach requires that we observe Y_t in both periods.

Neither Assumption 2.2 or 2.6 nest the other. For example, while Assumption 2.2 does not restrict the conditional distribution of A given X it does exclude time-varying heteroscedasticty allowed by Assumption 2.6.

A natural combination of the two assumptions is:

Assumption 2.7 (Stationarity and Exchangeability) *(i)*

$$U_t|X, A \overset{D}{=} U_s|X, A, \qquad t \neq s,$$

(ii) the distribution of U_t given X and A is non-degenerate for all $(X, A) \in \mathcal{X} \times \mathcal{A}$, (iii)

$$A|X_1, \ldots X_T \overset{D}{=} A|X_{p(1)}, \ldots X_{p(T)},$$

for $p(t) \in \{1, \ldots, T\}$, $p(t) \neq p(t')$.

[16] These polynomials take the form $\zeta_1(X) = \sum_{1 \leq i \leq T} X_i$, $\zeta_2(X) = \sum_{1 \leq i < j \leq T} X_i X_j$, $\zeta_3(X) = \sum_{1 \leq i < j < k \leq T} X_i X_j X_k$, $\zeta_4(X) = \sum_{1 \leq i < j < k < l \leq T} X_i X_j X_k X_l$ and so on up to $\zeta_T(X) = \prod_{i=1}^{T} X_i$.

[17] One suggestion made by Altonji and Matzkin (2005) is to impose a correlated random coefficients structure on $m_t(X_t, A, U_t)$, as we do here (Equation immediately prior to Equation (2.6) on p. 1064).

Under Assumption 2.7 $\beta(x)$ is overidentified since

$$\beta(x) = \frac{\mathbb{E}[Y_2 | X = x] - \mathbb{E}[Y_1 | X = x]}{x_2 - x_1} = \frac{\mathbb{E}[Y_2 | X' = x'] - \mathbb{E}[Y_1 | X' = x']}{x_2' - x_1'},$$

when x' is a permutation of x.

3 Estimation

In this section we discuss estimation of the movers average partial effect, β^M, and movers local average response, $\gamma_t^M(x_t)$, when the regressors are discretely-valued, and the average partial effect, β, and local average response, $\gamma_t(x_t)$, with continuously-valued regressors. To keep the exposition simple we work with the constant in X_2 aggregate time effects specification $\delta(X_2) = \delta$.

3.1 Discrete regressor case

We begin with the discrete regressor case, as it straightforward, and foreshadows our estimation approach for continuous regressors. Under our assumptions we can identify the common trend by the average change in Y_t across the two periods among the subpopulation of 'stayers'. That is

$$\delta \equiv \mathbb{E}[\Delta Y | \Delta X = 0] = \frac{\mathbb{E}[\mathbf{1}(\Delta X = 0)\Delta Y]}{\mathbb{E}[\mathbf{1}(\Delta X = 0)]}.$$

We, of course, require that $\Pr(\Delta X = 0)$ is greater than zero: it is the presence of stayers which identifies δ_0. Now consider the subpopulation of movers, we have

$$\mathbb{E}[\Delta Y | X = x] = \delta_0 + \beta(x)\Delta x,$$

and hence, with δ identified, we may write

$$\beta^M \equiv \frac{\mathbb{E}\left[\mathbf{1}(\Delta X \neq 0)\frac{\mathbb{E}[\Delta Y|X]-\delta}{\Delta X}\right]}{\mathbb{E}[\mathbf{1}(\Delta X \neq 0)]} = \frac{\mathbb{E}\left[\mathbf{1}(\Delta X \neq 0)\frac{\Delta Y-\delta}{\Delta X}\right]}{\mathbb{E}[\mathbf{1}(\Delta X \neq 0)]}.$$

Let $\theta = \left(\delta, \beta^M\right)'$, the above expressions generate the following 2×1 vector of moment restrictions $\mathbb{E}[\psi(Z, \theta_0)] = 0$, with

$$\psi(Z, \theta) = \begin{pmatrix} \mathbf{1}(\Delta X = 0)(\Delta Y - \delta) \\ \frac{\mathbf{1}(\Delta X \neq 0)}{\Delta X}\left(\Delta Y - \delta - \beta^M \Delta X\right) \end{pmatrix}.$$

The GMM estimate $\widehat{\beta}^M$ is very easy to compute, being the coefficient on ΔX in the linear instrumental variables fit of ΔY on a constant and ΔX with $\mathbf{1}(\Delta X = 0)$ and $\frac{\mathbf{1}(\Delta X \neq 0)}{\Delta X}$ serving as excluded instruments (this follows since $\mathbf{1}(\Delta X = 0)\left(\Delta Y - \delta - \beta^M \Delta X\right) = \mathbf{1}(\Delta X = 0)(\Delta Y - \delta)$). Conventional 'robust' standard errors reported by most software packages will be asymptotically valid.

Since it foreshadows portions of our results for the continuous X_t case we present a closed-form expression for the asymptotic sampling variance of $\widehat{\beta}^M$. Let $\Gamma_0 = \mathbb{E}\left[\partial\psi\left(Z,\theta_0\right)/\partial\theta'\right]$ and $\Omega_0 = \mathbb{E}\left[\psi\left(Z,\theta_0\right)\psi\left(Z,\theta_0\right)'\right]$ and further define

$$\pi_0 = \Pr\left(\Delta X = 0\right), \qquad \sigma_0^2 = \mathbb{V}\left(Y \middle| \Delta X = 0\right)$$
$$\xi = \mathbb{E}\left[\frac{1}{\Delta X}\middle| \Delta X \neq 0\right], \qquad \kappa = \mathbb{E}\left[\mathbb{V}\left[\frac{\Delta Y}{\Delta X}\middle| X\right]\middle| \Delta X \neq 0\right] + \mathbb{V}\left(\beta\left(X\right)\middle| \Delta X \neq 0\right).$$

We have

$$\Gamma_0 = -\begin{pmatrix} \pi_0 & 0 \\ \left(1-\pi_0\right)\xi & \left(1-\phi_0\right) \end{pmatrix}, \qquad \Omega_0 = \begin{pmatrix} \pi_0\sigma_0^2 & 0 \\ 0 & \left(1-\pi_0\right)\kappa \end{pmatrix},$$

and hence, by standard results for GMM (e.g., Newey and McFadden, 1994), an asymptotic sampling distribution for $\widehat{\theta}$ of

$$\sqrt{N}\begin{pmatrix} \widehat{\delta}-\delta \\ \widehat{\beta}^M - \beta^M \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\underline{0}, \begin{pmatrix} \frac{\sigma_0^2}{\pi_0} & -\frac{\sigma_0^2}{\pi_0}\xi \\ -\frac{\sigma_0^2}{\pi_0}\xi & \kappa + \frac{\sigma_0^2}{\pi_0}\xi^2 \end{pmatrix}\right).$$

Estimation of the MLAR is also straightforward. An argument analogous to that given for β^M yields the representation

$$\gamma_t^M\left(x_t\right) = \frac{\mathbb{E}\left[\mathbf{1}\left(X_t = x_t\right)\mathbf{1}\left(\Delta X \neq 0\right)\frac{\Delta Y - \delta}{\Delta X}\right]}{\mathbb{E}\left[\mathbf{1}\left(X_t = x_t\right)\mathbf{1}\left(\Delta X \neq 0\right)\right]}.$$

We may therefore estimate $\theta = \left(\delta, \gamma_t^M\left(x_t\right)\right)'$ by the analog estimator based on the population moment restriction $\mathbb{E}\left[\psi\left(Z,\theta_0\right)\right] = 0$ with

$$\psi\left(Z,\theta\right) = \begin{pmatrix} \mathbf{1}\left(\Delta X = 0\right)\left(\Delta Y - \delta\right) \\ \frac{\mathbf{1}\left(X_t = x_t\right)\mathbf{1}\left(\Delta X \neq 0\right)}{\Delta X}\left(\Delta Y - \delta - \gamma_t^M\left(x_t\right)\Delta X\right) \end{pmatrix}.$$

3.2 Continuous regressor case

3.2.1 No time effect

When X_t is continuously distributed – or, more precisely, when ΔX is continuously distributed in a neighborhood of zero – and no aggregate time effects are present ($\delta\left(X_2\right) = 0$), then Theorem 2.1 implies that the average partial effect β is identified by

$$\beta = \mathbb{E}\left[\frac{\mathbb{E}\left[\Delta Y \middle| X\right]}{\Delta X}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[\Delta Y \middle| X\right]}{\Delta X}\left|\Delta X \neq 0\right.\right].$$

Given the second equality a natural estimator of the APE, β, would be that proposed for the discrete case above, that is,

$$\widetilde{\beta} = \frac{\sum_{i=1}^N \mathbf{1}(\Delta X_i \neq 0)\left(\frac{\Delta Y_i}{\Delta X_i}\right)}{\sum_{i=1}^N \mathbf{1}(\Delta X_i \neq 0)} = \frac{1}{N}\sum_{i=1}^N \frac{\Delta Y_i}{\Delta X_i}..$$

This estimator was informally suggested by Mundlak (1961, p.45); as Chamberlain (1980b, 1982) notes, it will be strongly consistent if $\mathbb{E}[|\Delta Y/\Delta X|] < \infty$ by the strong law of large numbers. However, if ΔX has a positive, continuous density at zero – and if $\mathbb{E}[|\Delta Y| \mid \Delta X = d]$ does not vanish at $d = 0$ – then $\widetilde{\beta}$ will be inconsistent in general, since $\Delta Y/\Delta X$ will not have finite expectation (unlike $\beta(X) = \mathbb{E}[\Delta Y | X] / \Delta X$ whose expectation exists by assumption). For example, if (Y_t, X_t) is independently and identically distributed according to the bivariate normal distribution then $\Delta Y/\Delta X$ will be distributed according to the Cauchy distribution.

To ensure quadratic-mean convergence, we consider instead a 'trimmed' estimator of the form

$$\widehat{\beta}(h_N) \equiv \frac{\sum_{i=1}^{N} \mathbf{1}(|\Delta X_i| > h_N) \left(\frac{\Delta Y_i}{\Delta X_i}\right)}{\sum_{i=1}^{N} \mathbf{1}(|\Delta X_i| > h_N)}, \tag{14}$$

where h_N is a deterministic bandwidth sequence tending to zero as N tends to infinity.[18]

The estimator $\widehat{\beta}(h_N)$ – which is consistent for β^M when X has finite support – has asymptotic properties similar to a standard (uniform) kernel regression estimator for a one-dimensional problem. In particular, it is straightforward to verify that

$$\mathbb{V}(\widehat{\beta}) = O\left(\frac{1}{Nh_N}\right) \gg O\left(\frac{1}{N}\right),$$

so the rate of convergence is necessarily slower than $1/N$ when $h_N \to 0$. Assuming in addition that the bias of $\widehat{\beta}(h_N)$ is geometric in the bandwidth parameter h_N – that is

$$\mathbb{E}\left[\mathbf{1}(|\Delta X| > h_N)\left(\frac{\Delta Y}{\Delta X}\right) - \beta(X)\right] = \mathbb{E}\left[\mathbf{1}(|\Delta X| \leq h_N)\beta(X)\right] = O(h_N^p)$$

for some $p > 0$ (typically $p = 2$) – the fastest rate of convergence of $\widehat{\beta}$ to β in quadratic mean will be achieved when the bandwidth sequence takes the form

$$h_N^* = h_0 \; N^{-1/(2p+1)},$$

which yields

$$\begin{aligned} \widehat{\beta}(h_N^*) - \beta &= O_p(N^{-p/(2p+1)}) \\ &\gg O_p(N^{-1/2}). \end{aligned}$$

While the bandwidth sequence h_N^* achieves the fastest rate of convergence for this estimator, the corresponding asymptotic normal distribution for $\widehat{\beta}(h_N^*)$ will be centered at a bias term involving the derivative of $\mathbb{E}[\beta(X)|\Delta X = d]$ at $d = 0$. The estimator $\widehat{\beta}$ will have an asymptotic (normal)

[18] An alternative consistent estimator would replace the denominator by the sample size N.

distribution centered at zero if the bandwidth h_N converges to zero faster than h_N^*; assuming

$$h_N = o(N^{-1/(2p+1)}),$$

routine application of Liapunov's CLT for triangular arrays yields the asymptotic distribution for $\widehat{\beta}$,

$$\sqrt{Nh_N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, 2\phi_0\sigma_0^2),$$

where

$$\phi_0 \equiv \lim_{h\downarrow 0} \frac{\Pr\{|\Delta X| \leq h\}}{2h}$$

is the density of ΔX at zero and

$$\sigma_0^2 \equiv \mathbb{V}\left(\Delta Y|\,\Delta X = 0\right) = \lim_{h\downarrow 0} \mathbb{V}\left(\Delta Y| - h < \Delta X < h\right).$$

Assuming $p = 2$, the asymptotic distribution of $\widehat{\beta}$ is similar as the asymptotic distribution of a (uniform) kernel regression estimator of $\mathbb{E}[\Delta Y|\Delta X = 0]$, except that the variance of the latter varies inversely with the density ϕ_0.

3.2.2 Aggregate time effect

When aggregate time effects are present, and the 'common trends' condition (Assumption 2.5) holds with $\delta\left(X_2\right) = \delta$, then (10) implies that the average partial effect β is identified by

$$\beta = \mathbb{E}\left[\frac{\mathbb{E}\left[\Delta Y|\,X\right] - \delta}{\Delta X}\right] = \mathbb{E}\left[\left.\frac{\mathbb{E}\left[\Delta Y|\,X\right] - \delta}{\Delta X}\right| \Delta X \neq 0\right],$$

recalling that $\delta \equiv \mathbb{E}[\Delta Y|\Delta X = 0]$. If δ were known, a straightforward modification of the estimator proposed in the preceding section would be

$$\widehat{\beta}_I = \frac{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)\left(\frac{\Delta Y_i - \delta}{\Delta X_i}\right)}{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)},$$

which would inherit the large sample properties of $\widehat{\beta}$ above.

When δ is unknown, a natural counterpart to the infeasible estimator $\widehat{\beta}_I$ replaces δ with the uniform kernel estimator,

$$\widehat{\delta} \equiv \frac{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| \leq h_N)\Delta Y_i}{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| \leq h_N)}, \tag{15}$$

whose asymptotic properties are well-known when ΔX is continuously distributed. Under standard regularity conditions a normalized version of $\widehat{\delta}$ has the asymptotic distribution,

$$\sqrt{Nh_N}(\widehat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2/2\phi_0),$$

where ϕ_0 and σ_0^2 are defined above. Furthermore, $\widehat{\beta}_I$ and $\widehat{\delta}$ will be asymptotically independent, as the product of their influence functions will be zero by construction.

Given this estimator of the common trend δ, a feasible estimator of the APE β would be

$$\widehat{\beta}_F = \frac{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)\left(\frac{\Delta Y_i - \widehat{\delta}}{\Delta X_i}\right)}{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)}. \tag{16}$$

Though simple in appearance, derivation of the large-sample properties of $\widehat{\beta}_F$ is difficult, as its rate of convergence depends in a delicate way on the distribution of the regressors X. Writing the normalized version of $\widehat{\beta}_F$ in terms of its infeasible counterpart $\widehat{\beta}_I$ yields

$$\sqrt{Nh_N}(\widehat{\beta}_F - \beta) = \sqrt{Nh_N}(\widehat{\beta}_I - \beta) - \sqrt{Nh_N}(\widehat{\delta} - \delta) \times \left[\frac{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)\left(\frac{1}{\Delta X_i}\right)}{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)}\right].$$

While the asymptotic behavior of the first two terms in this decomposition are straightforward, the rate of convergence of the third term,

$$\widehat{\xi} \equiv \frac{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)\left(\frac{1}{\Delta X_i}\right)}{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| > h_N)},$$

will crucially depend upon the behavior of

$$\tau(d) \equiv \mathbb{E}\left[sgn\{\Delta X\}|\ |\Delta X|\ = d\right]$$

for d in a neighborhood of zero.

If, for example, X_1 and X_2 are exchangeable, so that ΔX is symmetrically distributed about zero (at least for $|\Delta X|$ in a neighborhood of zero), then $\tau(d) \equiv 0$ and $\hat{\xi}$ will converge in probability to zero, ensuring the asymptotic equivalence of the feasible estimator $\widehat{\beta}_F$ and its infeasible counterpart $\widehat{\beta}_I$. Alternatively, if there is constant positive drift in the distribution of regressors, so that $\tau(0) > 0$, then the third term $\widehat{\xi}$ will diverge, with expectation of $O(\log(h_N^{-1}))$, which is $O(\log(N))$ if $h_N = O(N^{-r})$ for some $r > 0$. In the latter case, the asymptotic distribution of the feasible estimator $\widehat{\beta}_F$ will be dominated by the asymptotic distribution of the estimator $\widehat{\delta}$ of the common trend. An intermediate case could have $\tau(d) = O(d)$ in a neighborhood of zero, with the third term converging in probability to some nonzero limit.

In any event, an asymptotic variance estimator for $\widehat{\beta}_F$ can be constructed if consistent estimators of the density ϕ_0 and conditional variance σ_0^2 terms appearing in the asymptotic variances of $\widehat{\beta}_I$ and $\widehat{\delta}$ can be constructed. Under standard regularity conditions, the kernel estimators

$$\widehat{\phi} \equiv \frac{1}{2Nh_N}\sum_{i=1}^N \mathbf{1}(|\Delta X_i| \leq h_N), \qquad \widehat{\sigma}^2 \equiv \frac{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| \leq h_N)\,(\Delta Y_i)^2}{\sum_{i=1}^N \mathbf{1}(|\Delta X_i| \leq h_N)} - \widehat{\delta}^2$$

should converge in probability to ϕ_0 and σ_0^2; given these estimators, an estimator of the asymptotic variance of the feasible estimator $\widehat{\beta}_F$ can be constructed as

$$\widehat{AVar}(\widehat{\beta}_F) = \frac{\widehat{\sigma}^2}{Nh_N}\left(2\widehat{\phi} + \frac{\widehat{\xi}^2}{2\widehat{\phi}}\right),$$

for $\widehat{\xi}$ as defined above. This estimator will automatically adapt to divergence of $\widehat{\xi}$ or its convergence to a (possibly nonzero) constant in probability.

3.2.3 Mixed discrete-continuous regressors

In some applications the distribution of the regressors (X_1, X_2) may have mass points at a finite set of values, and will be continuously distributed elsewhere. If there is overlap in the mass points of X_1 and X_2, then the distribution of first differences ΔX will generally have a mass point at zero, and will otherwise be continuously distributed in a neighborhood of zero. In this setting, the average partial effect β will generally differ from its 'movers' counterpart β^M, due to the nonzero probability that $\Delta X = 0$; while this mass point simplifies estimation of a nonzero common trend component δ (and the conditional variance of ΔY given $\Delta X = 0$), it complicates estimation of the APE. This is because β typically differs from β^M, which is the implicit estimand of (14) and (16) above, when 'stayers' are a non-negligble portion of the population.

When $\pi_0 \equiv \Pr(\Delta X = 0) > 0$, the estimator

$$\widetilde{\delta} \equiv \frac{\sum_{i=1}^N \mathbf{1}(\Delta X_i = 0) \cdot \Delta Y_i}{\sum_{i=1}^N \mathbf{1}(\Delta X_i = 0)},$$

used for the discrete X_t case discussed above, is clearly $\sqrt{N}$-consistent and asymptotically normal estimator for δ, as would be the (asymptotically equivalent) estimator $\widehat{\delta}$, defined in the previous subsection (under standard regularity conditions). Using the decomposition for the feasible estimator $\widehat{\beta}_F$ of $\beta^M \equiv \mathbb{E}[\beta(X)|\Delta X \neq 0]$ in the previous section, it follows that

$$\begin{aligned} \sqrt{Nh_N}(\widehat{\beta}_F - \beta^M) &= \sqrt{Nh_N}(\widehat{\beta}_F - \beta^M) + O_p(\sqrt{h_N}) \cdot O_p(\log(h_n^{-1})) \\ &= \sqrt{Nh_N}(\widehat{\beta} - \beta^M) + o_p(1), \end{aligned}$$

so that preliminary estimation of the common trend component δ will not affect the asymptotic distribution of the feasible estimator $\widehat{\beta}_F$. If a consistent estimator of the stayers effect

$$\beta^S \equiv \mathbb{E}[\beta(X)|\ \ \Delta X \ = 0]$$

can be constructed, a corresponding consistent estimator of the APE $\beta = \pi_0\beta^S + (1-\pi_0)\beta^M$ would be

$$\widehat{\beta} \equiv \widehat{\pi}\widehat{\beta}^S + (1-\widehat{\pi})\widehat{\beta}^F,$$

where $\widehat{\pi} \equiv \sum_{\iota=1}^{N} \mathbf{1}(|\Delta X| \leq h_N)/N$ is a $\sqrt{N}$-consistent estimator for π_0.

Defining

$$\nu(d) \equiv \mathbb{E}\left[\Delta Y| \ |\Delta X| \ = d\right],$$

the results of Section 2 above imply that

$$\beta^S = \lim_{h \downarrow 0} \frac{\nu(h) - \nu(0)}{h};$$

thus, estimation of β^S amounts to estimation of a (left) derivative at zero of the conditional mean of ΔY given $\Delta X = 0$. One such consistent estimator would be the slope coefficient of a local linear regression of ΔY on a constant term and ΔX, i.e.,

$$\begin{pmatrix} \overline{\delta} \\ \widehat{\beta}^S \end{pmatrix} = \arg\min_{d, b^S} \sum_{i=1}^{N} \mathbf{1}(|\Delta X_i| \leq h_N) \cdot (\Delta Y_i - d - b^S \Delta X_i)^2, \tag{17}$$

with the intercept $\overline{\delta}$ being an alternative ($\sqrt{N}$-)consistent estimator of the common trend δ. Since the rate of convergence of a nonparametric estimator of the derivative of a regression function is lower than for its level, the rate of convergence the combined estimator $\widehat{\beta} \equiv \widehat{\pi}\widehat{\beta}^S + (1-\widehat{\pi})\widehat{\beta}^F$ of the APE β will be the same as for $\widehat{\beta}^S$, and the asymptotic distribution of the latter would dominate the asymptotic distribution of $\widehat{\beta}$ in this setting.

3.2.4 Local average response [Incomplete]

A consistent estimate of $\gamma_t(x_t)$ is

$$\widehat{\gamma}_t(x_t) = \frac{\sum_{i=1}^{N} \mathcal{K}\left(\frac{X_{it}-x_t}{h_{2N}}\right) \mathbf{1}(|\Delta X_i| > h_{1N}) \left(\frac{\Delta Y_i - \widehat{\delta}}{\Delta X_i}\right)}{\sum_{i=1}^{N} \mathcal{K}\left(\frac{X_{it}-x_t}{h_{2N}}\right) \mathbf{1}(|\Delta X_i| > h_{1N})}$$

where $\widehat{\delta}$ is given by (15) above.

3.2.5 Computation

APE For estimation of, and inference on, the APE we propose using a simple 'instrumental variables' procedure. Consider the instrumental variables fit associated with the linear regression of ΔY on a constant and the interactions $\mathbf{1}(|\Delta X| > h_N) \cdot \Delta X$ and $\mathbf{1}(|\Delta X| \leq h_N) \cdot \Delta X$ with $\mathbf{1}(|\Delta X| \leq h_N)$, $\mathbf{1}(|\Delta X| \leq h_N) \cdot \Delta X$ and $\frac{1(|\Delta X| > h_N)}{\Delta X}$ serving as excluded instruments. The coefficients on the first two regressors will equal those defined by (17) above, while the coefficient on the last regressors is equal to (16) (with $\overline{\delta}$ replacing $\widehat{\delta}$). The robust standard errors reported by most statistical packages will be asymptotically valid.[19]

[19] Note these standard errors will implicitly include estimates of asymptotically negligible terms. However, this may improve small sample coverage of the resulting confidence intervals (cf., Newey 1994b).

If the mixed discrete-continuous case discussed above is of relevance, then choosing $\widehat{\theta}\,(h_N) = (\widehat{\pi}, \overline{\delta}, \widehat{\beta}^S, \widehat{\beta}^M)'$ to solve

$$\sum_{i=1}^{N} \psi(Z, \widehat{\theta}\,(h_N))/N = 0,$$

with

$$\begin{aligned}
&\psi\left(Z, \theta(h_N)\right) \\
&= \begin{pmatrix}
\mathbf{1}(|\Delta X| \leq h_N) - \pi \\
\mathbf{1}(|\Delta X| \leq h_N)\left(\Delta Y - \delta - \beta^M\left\{\mathbf{1}(|\Delta X| > h_N)\cdot \Delta X\right\} - \beta^S\left\{\mathbf{1}(|\Delta X| \leq h_N)\cdot \Delta X\right\}\right) \\
\left\{\mathbf{1}(|\Delta X| \leq h_N)\cdot \Delta X\right\}\left(\Delta Y - \delta - \beta^M\left\{\mathbf{1}(|\Delta X| > h_N)\cdot \Delta X\right\} - \beta^S\left\{\mathbf{1}(|\Delta X| \leq h_N)\cdot \Delta X\right\}\right) \\
\frac{\mathbf{1}(|\Delta X| > h_N)}{\Delta X}\left(\Delta Y - \delta - \beta^M\left\{\mathbf{1}(|\Delta X| > h_N)\cdot \Delta X\right\} - \beta^S\left\{\mathbf{1}(|\Delta X| \leq h_N)\cdot \Delta X\right\}\right)
\end{pmatrix},
\end{aligned}$$

recovers all the components needed to form an estimate of the average partial effect

$$\widehat{\beta} \equiv \widehat{\pi}\widehat{\beta}^S + (1-\widehat{\pi})\widehat{\beta}^F.$$

In practice, a combination of the conventional GMM covariance matrix for $\widehat{\theta}\,(h_N)$ and the textbook delta method may be used to form standard errors for $\widehat{\beta}$.

4 Multiple regressors and time periods

In this section we extend our basic model to permit multiple non-constant regressors and panels of arbitrary length. We analyze the following correlated random coefficients model:

$$Y_t = \mathbf{W}_t'\mathbf{d}\,(A, U_t) + \mathbf{X}_t'\mathbf{b}\,(A, U_t)\,, \quad t = 1, ..., T,$$

where $\mathbf{W}_t$ and $\mathbf{X}_t$ are $q \times 1$ and $p \times 1$ vectors of observable regressors and $\mathbf{d}\,(U_t)$ and $\mathbf{b}\,(A, U_t)$ corresponding random coefficients (all with bounded moments).

Our marginal stationarity restriction is

$$U_t|\,\mathbf{W}, \mathbf{X}, A \sim U_s|\,\mathbf{W}, \mathbf{X}, A,$$

for $s \neq t$, $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_T)'$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)'$. This implies that

$$\mathbb{E}[\mathbf{d}\,(A, U_t)\,|\mathbf{W}, \mathbf{X}] = \boldsymbol{\delta}\,(\mathbf{W}, \mathbf{X})$$

and

$$\mathbb{E}[\mathbf{b}\,(A, U_t)\,|\mathbf{W}, \mathbf{X}] = \boldsymbol{\beta}\,(\mathbf{W}, \mathbf{X})\,.$$

To complete the model we make the additional restrictions that

$$\boldsymbol{\delta}\,(\mathbf{W}, \mathbf{X}) \equiv \boldsymbol{\delta}, \quad \boldsymbol{\beta}\,(\mathbf{W}, \mathbf{X}) = \boldsymbol{\beta}\,(\mathbf{X})\,.$$

In this model $\mathbf{W}$ is a $T \times q$ matrix of aggregate 'time shifters'. Typically we think of these regressors as varying deterministically with t, and hence the coefficients $\mathbf{d}\,(A, U_t)$ as capturing time- and individual-specific trends. The $T \times p$ matrix of regressors $\mathbf{X}$ includes the choice/policy variables of primary interest.

The two period model considered in the preceding sections is contained within the above family with $T = p = 2$ and $q = 1$. The matrix of time shifters and its corresponding coefficient vector parameterize the common intercept shift across periods:

$$\mathbf{W} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \boldsymbol{\delta} = \delta,$$

while the choice variable and the conditional means of the random coefficients are given by

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \end{pmatrix}, \quad \boldsymbol{\beta}\,(\mathbf{X}) = \begin{pmatrix} \alpha\,(X) \\ \beta\,(X) \end{pmatrix}.$$

As before, the parameters of interest are $\boldsymbol{\delta} \equiv \mathbb{E}[\mathbf{d}\,(A, U_t)]$, $\beta \equiv \mathbb{E}[\mathbf{b}\,(A, U_t)]$ and $\boldsymbol{\gamma}_t\,(\mathbf{x}_t) \equiv \mathbb{E}[\,\mathbf{b}\,(A, U_t)|\,\mathbf{X}_t = \mathbf{x}_t]$.

The above model is a special case of the CRC model proposed and analyzed by Chamberlain (1992a), who worked with a more general setup where regressors and trend coefficients were permitted to vary parametrically (i.e., $\mathbf{W} = \mathbf{W}(\boldsymbol{\theta})$, $\mathbf{X} = \mathbf{X}(\boldsymbol{\theta})$, and $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{\theta})$). Identification of $\boldsymbol{\delta}$ and β in the overidentified setup $T > p$ was considered in detail by Chamberlain (1992a), we begin with 'just identified' case $T = p$, which he did not consider, and return to the overidentified case subsequently.

4.1 Just identification

Writing $\mathbf{Y} = (Y_1, ..., Y_T)'$ we have

$$\mathbb{E}[\mathbf{Y}|\mathbf{W}, \mathbf{X}] = \mathbf{W}\boldsymbol{\delta} + \mathbf{X}\boldsymbol{\beta}(\mathbf{X}). \tag{18}$$

Define $\tilde{X}$ to be the (scalar) determinant of the matrix of regressors,

$$\tilde{X} = \det(\mathbf{X}),$$

and $\mathbf{X}^*$ to be the *adjoint* (or *adjunct*) matrix to $\mathbf{X}$, i.e., the transpose of the matrix of cofactors of $\mathbf{X}$,

$$\mathbf{X}^* \equiv adj(\mathbf{X}),$$

so that, $\mathbf{X}^*\mathbf{X} = \tilde{X} \cdot \mathbf{I}$, and, when $\tilde{X} \neq 0$, $\mathbf{X}^{-1} = (1/\tilde{X}) \cdot \mathbf{X}^*$ (recall that with $T = p$ that $\mathbf{X}$ is a square matrix). Premultiplication of the vector of conditional means of Y_t by the adjoint matrix $\mathbf{X}^*$ thus yields

$$\mathbb{E}[\mathbf{X}^*\mathbf{Y}|\mathbf{W}, \mathbf{X}] = \mathbf{X}^*\mathbf{W}\boldsymbol{\delta} + \tilde{X} \cdot \boldsymbol{\beta}(\mathbf{X}),$$

which implies that

$$\mathbb{E}[\mathbf{X}^*\mathbf{Y}|\mathbf{X},\mathbf{W},\tilde{X}=0]=\mathbf{X}^*\mathbf{W}\boldsymbol{\delta},$$

assuming $\mathbf{Y}$ and $\mathbf{X}$ have at least $T+1$ moments finite (ensuring $E[||\mathbf{X}^*\mathbf{Y}||]<\infty$).

Provided the random $(T\times q)$ matrix $\mathbf{X}^*\mathbf{W}$ has q-dimensional support conditional on $\tilde{X}=0$, the coefficient vector $\boldsymbol{\delta}$ is identified by a population regression of $\mathbf{X}^*\mathbf{Y}$ on $\mathbf{X}^*\mathbf{W}$ conditional on $\tilde{X}\equiv\det(\mathbf{X})=0$. By analogy with the estimation results for the scalar case presented above, a consistent estimator of $\boldsymbol{\delta}$ can be constructed using a weighted least-squares regression of $\mathbf{X}_i^*\mathbf{Y}_i$ on $\mathbf{X}_i^*\mathbf{W}_i$ across all observations $i=1,...,N$, with weights equal to $\mathbf{1}(|\tilde{X}_i|\le h_N)$. Thus, estimation of $\boldsymbol{\delta}$ still involves a one-dimensional nonparametric regression problem in the (scalar) conditioning variable $\tilde{X}_i$.

In the $T=2$ case considered in the preceding sections we have

$$\tilde{X}=\det\begin{pmatrix}1 & X_1\\ 1 & X_2\end{pmatrix}=\Delta X,$$

so that

$$\mathbf{X}^*\mathbf{W}\boldsymbol{\delta}=\begin{bmatrix}X_2 & -X_1\\ -1 & 1\end{bmatrix}\begin{pmatrix}0\\ 1\end{pmatrix}\delta=\begin{pmatrix}-X_1\delta\\ \delta\end{pmatrix},$$

and

$$\mathbf{X}^*\mathbf{Y}=\begin{pmatrix}X_2Y_1-X_1Y_2\\ \Delta Y\end{pmatrix}.$$

When $\tilde{X}=\Delta X=0$, the two rows of $\mathbf{X}^*\mathbf{Y}-\mathbf{X}^*\mathbf{W}\boldsymbol{\delta}$ are proportional to each other, and either could be used to define a nonparametric estimator of δ; in the preceding sections, the second row was used.

Returning to the general case $T=p\ge 2$, given identification of $\boldsymbol{\delta}$, identification of β^M follows from the equality

$$\mathbb{E}[\mathbf{Y}-\mathbf{W}\boldsymbol{\delta}|\mathbf{W},\mathbf{X}]=\mathbf{X}\boldsymbol{\beta}(\mathbf{X}).$$

When $\tilde{X}\equiv\det(\mathbf{X})\neq\mathbf{0}$, premultiplying both sides of this relation by $\mathbf{X}^{-1}$ yields

$$\mathbf{E}[\mathbf{X}^{-1}(\mathbf{Y}-\mathbf{W}\boldsymbol{\delta})\mid\mathbf{X}=\mathbf{x}]\equiv\boldsymbol{\beta}(\mathbf{x}),$$

so that, assuming $\Pr\left(\tilde{X}\neq 0\right)>0$

$$\mathbb{E}[\mathbf{X}^{-1}(\mathbf{Y}-\mathbf{W}\boldsymbol{\delta})\mid\tilde{X}\neq 0]=\mathbb{E}[\boldsymbol{\beta}(\mathbf{X})\mid\tilde{X}\neq 0]\equiv\boldsymbol{\beta}^M$$

by iterated expectations.

If $\tilde{X}\neq 0$ with probability one, then the movers average partial effect coincides with the overall or full average partial effect (i.e., $\beta^M=\beta=\mathbb{E}[\mathbf{b}\,(A,U_t)]$). Heuristically, β^M is identified as an average of a generalized least-squares regression of the detrended conditional mean $\mathbb{E}[\mathbf{Y}|\mathbf{W},\mathbf{X}]-\mathbf{W}\boldsymbol{\delta}$ on $\mathbf{X}$, averaging over those observations with $\tilde{X}=\det(\mathbf{X})\neq\mathbf{0}$.

Because the expectation of $\mathbf{X}^{-1}(\mathbf{Y}-\mathbf{W}\boldsymbol{\delta})$ will generally be undefined when $\tilde{X}$ is continuously distributed with positive density near zero, estimation of β^M will involve the same trimmed mean as discussed for the special case $T=p=2$ above. The extension of the feasible estimator $\widehat{\boldsymbol{\beta}}_F$ to this context is

$$\widehat{\boldsymbol{\beta}}_F = \frac{\sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| > h_N)\cdot \mathbf{X}_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\widehat{\boldsymbol{\delta}})}{\sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| > h_N)},$$

where $\widehat{\boldsymbol{\delta}}$ is the nonparametric estimator

$$\widehat{\boldsymbol{\delta}} = \left[\sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| \le h_N)\,(\mathbf{X}_i^{*}\mathbf{W}_i)'\,(\mathbf{X}_i^{*}\mathbf{W}_i)\right]^{-1} \times \sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| \le h_N)\,(\mathbf{X}_i^{*}\mathbf{W}_i)'\,(\mathbf{X}^{*}\mathbf{Y}).$$

This estimator will converge in probability to β^M at a one-dimensional nonparametric rate if $h_N \to 0$ at the appropriate rate, provided the term

$$\hat{\boldsymbol{\xi}} \equiv \frac{\sum_{i=1}^{N} 1(\left|\tilde{X}_i\right| > h_N)\cdot \mathbf{X}_i^{-1}\mathbf{W}_i}{\sum_{i=1}^{N} 1(\left|\tilde{X}_i\right| > h_N)}$$

does not diverge too quickly as $N \to \infty$.

In the mixed discrete-continuous case $\Pr\left(\tilde{X}=0\right) > 0$, and estimation of $\hat{\beta}$ requires estimation of

$$\boldsymbol{\beta}^S = \lim_{h \downarrow 0} \frac{\boldsymbol{\nu}(h) - \boldsymbol{\nu}(0)}{h},$$

where

$$\boldsymbol{\nu}(x) \equiv E[\mathbf{X}^{-1}\mathbf{Y}|\tilde{X}=x];$$

the resulting estimator converges at the rate for nonparametric estimation of the derivative of a one-dimensional regression function.

4.2 Overidentification

When $T>p$, the vector of common trend parameters $\boldsymbol{\delta}$ will satisfy some conditional moment restrictions, and, as Chamberlain (1992a) shows, these typically suffice for identification and construction of root-N-consistent and asymptotically-normal estimators of $\boldsymbol{\delta}$. In this overidentified setting, for each realized matrix of regressors $\mathbf{X}$ there will be a $T \times (T-p)$ matrix $\mathbf{Z} \equiv \boldsymbol{\zeta}(\mathbf{X})$ of functions of $\mathbf{X}$ for which

$$\mathbf{Z}'\mathbf{X} = \mathbf{0};$$

from the relation (18) above, it follows that

$$\begin{aligned}\mathbf{Z}'\mathbb{E}[\mathbf{Y}|\mathbf{W},\mathbf{X}] &\equiv \mathbf{Z}'\mathbf{W}\boldsymbol{\delta} + \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}(\mathbf{X}) \\ &= \mathbf{Z}'\mathbf{W}\boldsymbol{\delta},\end{aligned}$$

so that

$$\mathbb{E}[\mathbf{Z}'(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}) \mid \mathbf{W}, \mathbf{X}] = \mathbf{0},$$

which, depending upon the form of $\mathbf{W}$, will typically serve to identify the trend coefficients $\boldsymbol{\delta}$.

For example, in the $T = 2$ example considered above, suppose the restriction $\alpha(x) \equiv \alpha$ is imposed, so that

$$\boldsymbol{\delta} \equiv (\alpha, \delta)', \quad \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{X} \equiv (X_1, X_2)';$$

then, taking $\mathbf{Z} = (X_2, -X_1)'$, the parameters α and δ will satisfy

$$\mathbb{E}[\,X_2Y_1 - X_1Y_2 - \alpha(X_1 + X_2) - \delta X_1 |\, X_1, X_2] = 0,$$

which implies that α and δ will be identified as population least-squares regression coefficients of $X_2Y_1 - X_1Y_2$ on $(X_1 + X_2)$ and X_1, respectively. Alternatively, restricting $\beta(x) = \beta$ but leaving $a(x)$ unrestricted, δ and β will be identified by the population regression of ΔY on a constant and ΔX, that is, the population analogue of the usual fixed-effects regression estimator.

Even in the just-identified setting $(T = p)$, it may be possible to obtain consistent estimators of $\boldsymbol{\delta}$ that achieve the parametric rate of convergence. If

$$\tilde{\mathbf{W}} \equiv \mathbf{W} - \mathbb{E}[\mathbf{W} \mid \mathbf{X}]$$

has a covariance matrix of full rank, then $\boldsymbol{\delta}$ will be identified by

$$\boldsymbol{\delta} = \mathbb{V}(\tilde{\mathbf{W}})^{-1}\mathbb{C}(\tilde{\mathbf{W}}, \mathbf{Y}),$$

and as long as the rank of $\mathbb{V}(\tilde{\mathbf{W}})$ is nonzero, some linear combinations of $\boldsymbol{\delta}$ will be identified by a similar argument. For the special cases considered above, where $\mathbf{W} = \boldsymbol{\omega}(\mathbf{X})$, this is not applicable, but such restrictions may be useful when $\mathbf{W}$ includes regressors which are not deterministic functions of $\mathbf{X}$ even when $T = p$.

Overidentification also makes estimation of β less problematic. As Chamberlain (1992a) shows, defining

$$\widehat{\boldsymbol{\beta}}_i \equiv (\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\widehat{\boldsymbol{\delta}})$$

for $\widehat{\boldsymbol{\delta}}$ a root-N-consistent estimator of $\boldsymbol{\delta}$ and $\mathbf{V}_i \equiv \boldsymbol{\nu}(\mathbf{W}_i, \mathbf{X}_i)$ positive definite with probability one, the sample mean of $\widehat{\boldsymbol{\beta}}_i$ will be a root-N-consistent estimator of β when $\mathbf{V}_i = \mathbb{V}(\,\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta} |\, \mathbf{X}_i)$ and

$$\mathbb{E}\left[\frac{1}{\det(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})}\right] < \infty. \tag{19}$$

This estimator also attains the semiparametric efficiency bound for estimation of β. Chamberlain (1992a) shows that a feasible version, based upon an efficient estimator of $\boldsymbol{\delta}$ and consistent estimators of $\{\mathbf{V}_i\}_{i=1}^N$, will also be semiparametrically efficient.

As the order of overidentification $T - p$ increases, condition (19) becomes less restrictive even if the components of $\mathbf{X}$ are continuously distributed. For example, consider the $p = 2$ case with $\mathbf{X}_t = (1, X_t)'$ and suppose that $X_t \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\mathbf{V}_i \equiv \mathbf{I}$; then

$$\det(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i) = \sum_{t=1}^{T}(X_t - \bar{X})^2 \sim \chi^2_{T-1},$$

and (19) will hold as long as $T - 1 > 2$, i.e., $T \geq 4$ here. More generally, as $T - p$ increases, the density of $\det(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)$ should approach zero more rapidly as its argument approaches zero – ensuring (19) holds – provided the continuous components of $\mathbf{X}_i$ are weakly dependent across rows and the matrix $\mathbf{V}_i$ is well-behaved.

Nevertheless, the trimming scheme used to estimate β in the just-identified setting may still be helpful in the overidentified case, even when (19) holds. Defining the (infeasible) trimmed mean

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{N}\mathbf{1}(\det(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i) > h_N)\cdot(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta})}{\sum_{i=1}^{N}\mathbf{1}(\det(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i) > h_N)},$$

it is straightforward to show this will be asymptotically equivalent to the sample mean of $\hat{\beta}_i$ when $\mathbb{E}[\boldsymbol{\beta}(\mathbf{X})|\det(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i) \leq h]$ is smooth (Lipschitz-continuous) in h, condition (19) holds, and $h_N = o(1/\sqrt{N})$. Since $\hat{\beta}$ will still be consistent for β even when (19) fails, a feasible version of the trimmed mean $\hat{\beta}$ may be better behaved in finite samples if the design matrix $(\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)$ is nearly singular for some observations.

5 Empirical application: the demand for calories

5.1 Data description and overview

We use data collected in conjunction with an external evaluation of the Nicaraguan conditional cash transfer program Red de Protección Social (RPS) (see IFPRI 2005). The RPS evaluation sample is a panel of 1,581 households from 42 rural communities in the departments of Madriz and Matagalpa, located in the northern part of the Central Region of Nicaragua. Twenty one of the sampled communities were randomly assigned to participate in the RPS program. Each sampled household was first interviewed in August/September 2000 with follow-ups attempted in October of both 2001 and 2002. Here we analyze a balanced panel of 1,358 households from the 2001 and 2002 waves.[20]

The survey was fielded using an abbreviated version of the 1998 Nicaraguan Living Standards Measurement Survey (LSMS) instrument. As such it includes a detailed consumption module with information on household expenditure, both actual and in kind, on 59 specific foods and several dozen other common budget categories (e.g., housing and utilities, health, education, and household goods). The responses to these questions were combined to form an annualized consumption aggregate, C_{it}. In forming this variable we followed the algorithm outlined by Deaton and Zaidi

[20] A total of 1,359 households were successfully interviewed in all three waves. One of these households reports zero food expenditures (and hence caloric availability) in one wave and is dropped from our sample.

(2002).

In addition to recording food expenditures, actual quantities of foods acquired are available. Using conversion factors listed in the World Bank (1998) and INEC (2005) we converted all food quantities into grams. We then used the caloric content and edible percent information in the INCAP (2000) food composition tables to construct a measure of daily total calories available for each household.[21] The logarithm of this measure, Y_{it}, serves as the dependent variable in our analysis.

The combination of both expenditure and quantity information at the household-level also allowed us to estimate unit prices for foods. These unit values were used to form a Paasche cost-of-living index for the i^{th} household in year t of

$$I_{it} = \left[S_{it} \left\{ \sum_{f=1}^{F} W_{f,it} \left(P_f^b / P_{f,it} \right) \right\} + (1 - S_{it}) J_{it} \right]^{-1}, \tag{20}$$

where S_{it} is the fraction of household spending devoted to food, $W_{f,it}$ the share of overall food spending devoted to the f^{th} specific food, $P_{f,it}$ the year t unit price paid by the household for food f, and P_f^b its 'base' price (equal to the relevant 2001 sample median price). To avoid the introduction of spurious noise we, following the suggestion of Deaton and Zaidi (2002), replace household-level unit prices with village medians. In the absence of price information on nonfood goods we set J_{it} equal to one in 2001 and to the national consumer price index (CPI) in 2002. Our independent variable of interest is the logarithm of real consumption: $X_{it} = \ln(C_{it}/I_{it})$.

Table 1 summarizes some key features of our estimation sample. Panel A gives the share of total food spending devoted to each of eleven broad food categories as well as the share of total calories derived from each group. Spending on staples (cereals, roots and pulses) accounts for about half of the average household's food budget and over two thirds of its calories. Among the poorest quartile of households an average of around 55 percent of budgets are devoted to, and over three quarters of calories available derived from, staples. Spending on vegetables, fruit and meat accounts for less than 15 percent of the average household's food budget and less than 3 percent of calories available. That such a large fraction of calories are derived from staples, while not good dietary practice, is not uncommon in poor households elsewhere in the developing world (cf., Subramanian and Deaton, 1996; Smith and Subandoro, 2007).

Panel B of the table lists real annual expenditure in Cordobas per adult equivalent and per capita. Adult equivalents are defined in terms of age- and gender-specific FAO (2001) recommended energy intakes for individuals engaging in 'light activity' relative to prime-aged males. As a point of reference the 2001 average annual expenditure per capita across all of Nicaragua was a nominal C$7,781, while amongst rural households it was C$5,038 (World Bank. 2003). The 42 communities in our sample, consistent with their participation in an anti-poverty demonstration experiment, are considerably poorer than the average Nicaraguan rural community.[22]

[21] In forming our measure of calorie availability we followed the general recommendations of Smith and Subandoro (2007).

[22] In October of 2001 the Coroba-to-US$ exchange rate was 13.65. Therefore per capita consumption levels in our

Using the FAO (2001) energy intake recommendations for 'light activity' we categorized each household, on the basis of its demographic structure, as energy deficient, or not. By this criterion approximately 40 percent of households in our sample are energy deficient each period. Amongst the poorest quartile this fraction rises to over 75 percent. These figures are reported in Panel B of Table 1.

Table 2 reports the median amount of Cordobas paid per one thousand calories by food type and expenditure quartiles. As found in other parts of the developing world, 'rich' households spend more per calorie than poor households, however, these price differences are not especially large in our sample. If quality upgrading is an important feature of food demand, then the elasticity of calorie demand with respect to total expenditure may be quite low even if the elasticity of food expenditure is quite high (Behrman and Deolalikar, 1987; Subramanian and Deaton, 1996).

5.2 Estimation results

We assume that the logarithm of total calorie availability in period t, Y_t, varies according to

$$Y_t = W_{1t}'\delta_1 + a_t(A, U_t) + b_t(A, U_t)X_t,$$

where W_{1t} is a $q_1 \times 1$ vector of time-varying demographic controls capturing the age and gender composition of the household, X_t is the logarithm of real total outlay, and $a_t(A, U_t)$ and $b_t(A, U_t)$ are the household and time specific intercept and calorie elasticity. We allow for common intercept and elasticity drift of the form:

$$\mathbb{E}\left[a_2(A, U_t) - a_1(A, U_t)\right| W_1, X] = \delta_a, \quad \mathbb{E}\left[b_2(A, U_t) - b_1(A, U_t)\right| W_1, X] = \delta_b.$$

Defining $\mathbf{W}_1 = (W_{11}, W_{12})'$ with

$$\mathbf{W}_2 = \begin{pmatrix} 0 & 0 \\ 1 & X_2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \end{pmatrix}$$

and $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2)$ gives our general model

$$\mathbb{E}[\mathbf{Y}|\mathbf{W}, \mathbf{X}] = \mathbf{W}\boldsymbol{\delta} + \mathbf{X}\boldsymbol{\beta}(\mathbf{X}),$$

with $\boldsymbol{\delta} = \left(\delta_1', \delta_a, \delta_b\right)'$ and $\boldsymbol{\beta}(\mathbf{x}) = (\alpha(x), \beta(x))' = (\mathbb{E}[a_1(A, U_t)|\, X = x], \mathbb{E}[b_1(A, U_t)|\, X = x])'$.

We estimate the time shift coefficients and the stayers average partial effect by the solution to the local linear regression

$$\begin{pmatrix} \overline{\boldsymbol{\delta}} \\ \widehat{\boldsymbol{\beta}}^S \end{pmatrix} = \arg\min_{d, b^S} \sum_{i=1}^{N} \mathbf{1}\left(\left|\tilde{X}_i\right| \leq h_N\right) \cdot (\mathbf{X}_i^* \mathbf{Y}_i - \mathbf{X}_i^* \mathbf{W}_i d - \tilde{X}_i b^S)^2,$$

sample averaged less than US$ 300 per year.

the movers average partial effect is then calculated using

$$\widehat{\boldsymbol{\beta}}^{M} = \frac{\sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| > h_N) \cdot \mathbf{X}_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\overline{\boldsymbol{\delta}})}{\sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| > h_N)}.$$

The distribution of X_t does not appear to have an discrete components, therefore $\widehat{\boldsymbol{\beta}}^{M}$ should provide a consistent estimate of the overall average elasticity of calorie demand as well as the movers average elasticity. Nevertheless we also report estimates of β based on

$$\widehat{\boldsymbol{\beta}} = \widehat{\pi}\widehat{\boldsymbol{\beta}}^{S} + (1-\widehat{\pi})\,\widehat{\boldsymbol{\beta}}^{M},$$

with $\widehat{\pi} = \sum_{i=1}^{N} \mathbf{1}(\left|\tilde{X}_i\right| \le h_N)/N$.

Practically we compute 'instrumental variables' estimates of π, $\boldsymbol{\delta}$, $\boldsymbol{\beta}^S$ and $\boldsymbol{\beta}^M$ by solving the sample 'moment conditions'

$$0 = \sum_{i=1}^{N} \psi_i\left(\widehat{\theta}\left(h_N\right)\right),$$

for $\widehat{\theta}\left(h_N\right) = \left(\widehat{\pi}, \widehat{\boldsymbol{\delta}}', \widehat{\boldsymbol{\beta}}^{S\prime}, \widehat{\boldsymbol{\beta}}^{M\prime}\right)'$ where

$$\psi_i\left(\theta\left(h_N\right)\right) = \left(\begin{array}{c} \mathbf{1}(\left|\tilde{X}_i\right| \le h_N) - \pi \\ \mathbf{Z}_i'\left(\tilde{\mathbf{Y}}_i - \tilde{\mathbf{W}}_i\boldsymbol{\delta} - \tilde{X}_i\left(\begin{array}{cc} \mathbf{1}(\left|\tilde{X}_i\right| \le h) & 0 \\ 0 & \mathbf{1}(\left|\tilde{X}_i\right| > h) \end{array}\right)\left(\begin{array}{c} \boldsymbol{\beta}^S \\ \boldsymbol{\beta}^M \end{array}\right)\right) \end{array}\right)$$

with

$$\tilde{\mathbf{Y}}_i = \mathbf{X}_i^{*}\mathbf{Y}_i, \qquad \tilde{\mathbf{W}}_i = \mathbf{X}_i^{*}\mathbf{W}_i,$$

and a $T \times q \times 2p$ instrument matrix given by

$$\mathbf{Z}_i \equiv \left[\mathbf{1}(\left|\tilde{X}_i\right| \le h) \cdot \tilde{\mathbf{W}}, \mathbf{1}(\left|\tilde{X}_i\right| \le h) \cdot \tilde{X}_i \cdot \mathbf{I}_p, \frac{\mathbf{1}(\left|\tilde{X}_i\right| > h)}{\tilde{X}_i} \cdot \mathbf{I}_p\right].$$

We use several different values of the bandwidth h; our default choice is the value that minimizes the 'cross-validation' criterion

$$CV(h) \equiv \sum_{i=1}^{N} \psi_i\left(\widehat{\theta}_{-i}\left(h\right)\right)' \psi_i\left(\widehat{\theta}_{-i}\left(h\right)\right)/N$$

where $\widehat{\theta}_{-i}\left(h\right)$ is the IV estimate calculated by omitting the i^{th} observation from the sample. For comparison purposes, the estimated standard deviation of $\tilde{X}_i \equiv \det(\mathbf{X}_i)$ for this sample is $\hat{\sigma} = 0.5315$, so $\hat{\sigma} \cdot N^{-1/5} = 0.1256$, which gives a 'rule-of-thumb' benchmark against which the cross-validated choice can be compared.

5.3 Summary and extensions

[To be completed].

Panel A:	Expenditure Shares (%)						Calorie Shares (%)					
	All		Lower 25%		Upper 25%		All		Lower 25%		Upper 25%	
	2001	2002	2001	2002	2001	2002	2001	2002	2001	2002	2001	2002
Cereals	36.0	32.7	41.0	35.6	32.2	29.4	60.3	59.9	63.7	61.9	58.1	58.2
Roots	3.1	2.7	2.3	1.7	3.5	3.8	1.5	1.6	1.4	1.2	1.7	2.3
Pulses	12.5	13.6	14.6	16.6	10.1	10.9	11.3	12.8	11.8	13.6	10.5	11.7
Vegetables	4.9	4.5	4.0	3.2	5.8	5.6	0.7	0.6	0.5	0.4	0.9	0.8
Fruit	0.9	1.1	0.6	0.9	1.1	1.3	0.5	0.4	0.2	0.3	0.7	5.8
Meat	6.9	7.7	3.6	4.7	9.7	10.6	1.3	1.3	0.6	0.7	2.1	1.9
Dairy	14.7	17.3	11.5	15.5	16.2	19.5	4.3	3.7	1.7	2.2	4.6	5.0
Oil	5.0	5.0	5.4	5.2	4.0	4.5	7.6	7.1	6.2	5.9	7.6	7.6
Other foods	16.0	15.4	17.0	16.5	14.2	14.5	12.6	10.3	9.5	9.8	11.8	10.5
Staples$^{\diamond}$	51.6	49.0	57.8	54.0	45.7	44.0	73.1	74.3	76.9	76.8	70.3	72.1
Panel B:	Total Real Expenditure & Calories											
Expenditure per adult$^{\flat}$	4,679	4,510	2,225	2,061	7,766	7,773						
(Expenditure per capita)	(3,764)	(3,887)	(1,947)	(1,991)	(5,970)	(6,320)						
Food share	69.2	68.8	68.4	68.5	68.7	68.2						
Calories per adult$^{\flat}$	3,015	2,949	2,045	1,952	3,981	3,878						
(Calories per capita)	(2,435)	(2,530)	(1,776)	(1,825)	(3,050)	(3,121)						
Percent energy deficient$^{\natural}$	39.3	39.7	73.5	77.6	12.7	10.6						

Table 1: Real food expenditure and calorie shares of RPS households in 2001 and 2002

NOTES: Authors' calculations based on a balanced panel of 1,358 households from the RPS evaluation dataset (see IFPRI (2005)). Real household expenditure equals total annualized nominal outlay divided by a Paasche cost-of-living index. Base prices for the price index are 2001 sample medians. The nomial exchange rate in October of 2001 was 13.65 Cordobas per US dollar. Total calorie availability is calculated using the RPS food quantity data and the calorie content and edible portion information contained in INCAP (2000). The Data Appendix provides complete details. Lower and upper 25 percent refers to the bottom and top quartiles of households based on the sum of 2001 and 2002 real consumption per adult equivalent and thus contains the same set of households in both years.

$^{\diamond}$ Sum of cereal, roots and pulses.

$^{\flat}$ "Adults" correspond to adult equivalents based on FAO (2001) recommended energy requirements for light activity.

$^{\natural}$ Percentage of households with estimated calorie availability less than FAO (2001) recommendations for light activity given household demographics.

	Median Cordobas per 1,000 calories					
	All		Bottom 25%		Top 25%	
	2001	2002	2001	2002	2001	2002
Cereals	1.3	1.2	1.2	1.0	1.5	1.4
Roots	8.6	7.2	7.6	6.0	8.6	7.2
Pulses	2.6	2.3	2.6	2.4	2.6	2.6
Vegetables	23.2	22.7	22.7	19.6	24.2	23.8
Fruit	6.3	6.6	5.4	5.5	6.6	7.1
Meat	19.1	18.6	18.6	18.1	19.9	19.3
Dairy	10.1	10.0	10.9	10.4	10.3	10.1
Oil	1.5	1.5	1.6	1.5	1.5	1.5
Other foods	3.1	3.1	2.9	2.8	3.4	3.6
All foods	2.4	2.4	2.0	2.0	3.0	2.9

Table 2: Real Cordobas spent by RPS households in 2001 and 2002 per 1,000 calories available by food category
NOTES: Authors' calculations based on a balanced panel of 1,358 households from the RPS evaluation dataset (IFPRI, 2005). Reported calorie prices are the median among households with positive consumption in the relevant category. Lower and upper 25 percent refers to the bottom and top quartiles of households based on the sum of 2001 and 2002 real consumption per adult equivalent. See notes to Table 1 and the Data Appendix for additional details.

	Pooled OLS		FE-OLS	
	(1.a)	(1.b)	(2.a)	(2.b)
log(Expenditure)	0.7383	0.6089	0.6325	0.6255
	(0.0377)	(0.0377)	(0.0383)	(0.0382)
1(Year=2002)	0.0423	0.0385	0.0286	0.0274
	(0.0320)	(0.0300)	(0.0285)	(0.0286)
Demographics	No	Yes	No	Yes

Table 3: Conventional estimates of the elasticity of calorie demand with respect to household expenditure
NOTES: Estimates based on the balanced panel of 1,358 households described in the main text. "Pooled OLS" denotes least squares applied to the pooled 2001 and 2002 samples, "FE-OLS" denotes least squares estimates with household-specific intercepts. The standard errors are computed in a way that allows for arbitrary within-village correlation in disturbances across households and time.