Identification and estimation of average partial effects in 'irregular' correlated random coefficient panel data models[1]

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

Initial Draft: December 2007

This Draft: July 8, 2011

Abstract

In this paper we study identification and estimation of a correlated random coefficients (CRC) panel data model. The outcome of interest varies linearly with a vector of endogenous regressors. The coefficients on these regressors are heterogenous across units and may covary with them. We consider the average partial effect (APE) of a small change in the regressor vector on the outcome (cf., Chamberlain, 1984; Wooldridge, 2005a). Chamberlain (1992) calculates the semiparametric efficiency bound for the APE in our model and proposes a $\sqrt{N}$ consistent estimator. Nonsingularity of the APE's information bound, and hence the appropriateness of Chamberlain's (1992) estimator, requires (i) the time dimension of the panel (T) to strictly exceed the number of random coefficients (p) and (ii) strong conditions on the time series properties of the regressor vector. We demonstrate irregular identification of the APE when $T = p$ and for more persistent regressor processes. Our approach exploits the different identifying content of the subpopulations of 'stayers' – or units whose regressor values change little across periods – and 'movers' – or units whose regressor values change substantially across periods. We propose a feasible estimator based on our identification result and characterize its large sample properties. While irregularity precludes our estimator from attaining parametric rates of convergence, its limiting distribution is normal and inference is straightforward to conduct. Standard software may be used to compute point estimates and standard errors. 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, Semiparametric Efficiency, Irregularity, Calorie Demand

---

[1]We would like to thank seminar participants at UC - Berkeley, UCLA, USC, Harvard, Yale, NYU, Princeton, Rutgers, Syracuse, Penn State, University College London, University of Pennsylvania, members of the Berkeley Econometrics Reading Group and participants in the Conference in Economics and Statistics in honor of Theodore W. Anderson's 90th Birthday (Stanford University), the Copenhagen Microeconometrics Summer Workshop and the JAE Conference on Distributional Dynamics (CEMFI, Madrid) for comments and feedback. Discussions with Manuel Arellano, Stéphane Bonhomme, Gary Chamberlain, Iván Fernández-Val, Jinyong Hahn, Jerry Hausman, Bo Honoré, Michael Jansson, Roger Klein, Ulrich Müller, John Strauss, and Edward Vytlacil were helpful in numerous ways. This revision has also benefited from the detailed comments of a co-editor as well as three anonymous referees. Max Kasy and Alex Poirier provided excellent research assistance. Financial support from the National Science Foundation (SES #0921928) is gratefully acknowledged. All the usual disclaimers apply.

[†]Department of Economics, New York University, 19 West 4th Street 6FL, New York, NY 10012 and National Bureau of Economic Research. E-mail: bryan.graham@nyu.edu. Web: https://files.nyu.edu/bsg1/public/

[◇]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/.

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., Card, 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 $\mathbf{Y} = (Y_1, \ldots, Y_T)'$ be a $T \times 1$ vector of outcomes and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)'$ a $T \times p$ matrix of regressors with $\mathbf{X}_t \in \mathbb{X}_t \subset \mathbb{R}^p$ and $\mathbf{X} \in \mathbb{X}^T$ where $\mathbb{X}^T = \times_{t \in \{1,\ldots,T\}} \mathbb{X}_t$. We assume that $\mathbf{X}_t$ is strictly exogenous. This rules out feedback from the period t outcome Y_t to the period $s \geq t$ regressor $\mathbf{X}_s$. One implication of this assumption is that lags of the dependent variable may not be included in $\mathbf{X}_t$. Our model is a static one.

Available is a random sample $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^N$ from a distribution F_0. The t^{th} period outcome is given by

$$Y_t = \mathbf{X}_t' b_t(A, U_t), \tag{1}$$

where A is time-invariant unobserved unit-level heterogeneity and U_t a time-varying disturbance. Both A and U_t may be vector-valued. The $p \times 1$ vector of functions $b_t(A, U_t)$, which we allow to vary over time, map A and U_t into unit-by-period-specific slope coefficients. By 'random' coefficients we mean that $b_t(A, U_t)$ varies across units. By 'correlated', we mean that the entire path of regressor values, $\mathbf{X}$, may have predictive power for $b_t(A, U_t)$. This implies that an agent's incremental return to an additional unit of $\mathbf{X}_t$ may vary with $\mathbf{X}_t$. In this sense $\mathbf{X}_t$ may be endogenous.

Equation (1) is structural in the sense that the *unit-specific* function

$$Y_t(\mathbf{x}_t) = \mathbf{x}_t' b_t(A, U_t) \tag{2}$$

traces out a unit's period t potential outcome across different hypothetical values of $\mathbf{x}_t \in \mathbb{X}_t$.[3] Let $\mathbf{X}_t = (1, X_t')'$; setting $b_{1t}(A, U_t) = \beta_1 + A + U_t$ (with A and U_t scalar and mean zero) and $b_{kt}(A, U_t) = \beta_k$ for $k = 2, \ldots, p$ yields the textbook linear panel data model:

$$Y_t(\mathbf{x}_t) = \mathbf{x}_t' \boldsymbol{\beta} + A + U_t, \tag{3}$$

for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Equation (2), while preserving linearity in $\mathbf{X}_t$, is more flexible than (3) in

[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, lower case letters specific realizations of them, and blackboard bold letters to denote their support (e.g., X, x and $\mathbb{X}$).

that it allows for time-varying random coefficients on all of the regressors (not just the intercept). Furthermore these coefficients may nonlinearly depend on A and/or U_t.

Our goal is to characterize the effect of an exogenous change in $\mathbf{X}_t$ on the probability distribution of Y_t. By 'exogenous change' we mean an external manipulation of $\mathbf{X}_t$ in the sense described by Blundell and Powell (2003) or Imbens and Newey (2009). We begin by studying identification and estimation of the *average partial effect* (APE) of $\mathbf{X}_t$ on Y_t (cf., Chamberlain, 1984; Blundell and Powell, 2003; Wooldridge, 2005a). Under (1) the average partial effect is given by

$$\boldsymbol{\beta}_{0t} \overset{def}{\equiv} \mathbb{E}\left[\frac{\partial Y_t(\mathbf{x}_t)}{\partial \mathbf{x}_t}\right] = \mathbb{E}\left[b_t(A, U_t)\right]. \tag{4}$$

Identification and estimation of (4) is nontrivial because, in our setup, $\mathbf{X}_t$ may vary systematically with A and/or U_t. To see the consequences of such dependence observe that the derivative of the mean regression function of Y_t given $\mathbf{X} = \mathbf{x}$ does not identify a structural parameter. Differentiating through the integral we have

$$\frac{\partial \mathbb{E}\left[Y_t | \mathbf{X} = \mathbf{x}\right]}{\partial \mathbf{x}_t} = \boldsymbol{\beta}_{0t}(\mathbf{x}) + \mathbb{E}\left[Y_t(\mathbf{X}_t)\mathbb{S}_{X_t}(A, U_t | \mathbf{X}) | \mathbf{X} = \mathbf{x}\right], \tag{5}$$

with $\boldsymbol{\beta}_{0t}(\mathbf{x}) = \mathbb{E}\left[b_t(A, U_t) | \mathbf{X} = \mathbf{x}\right]$ and $\mathbb{S}_{X_t}(A, U_t | \mathbf{X}) = \nabla_{X_t} \log f(A, U_t | \mathbf{X})$. The second term is what Chamberlain (1982) calls heterogeneity bias. If the (log) density of the unobserved heterogeneity varies sharply with $\mathbf{x}_t$ – corresponding to 'selection bias' or 'endogeneity' in a unit's choice of $\mathbf{x}_t$ – then the second term in (5) can be quite large.

Chamberlain (1982) studies identification of $\boldsymbol{\beta}_0 \overset{def}{\equiv} \boldsymbol{\beta}_{00}$ using panel data (cf., Mundlak, 1961, 1978b). In a second paper, Chamberlain (1992, pp. 579 - 585) calculates the semiparametric variance bound for $\boldsymbol{\beta}_0$ and proposes an efficient method-of-moments estimator.[4] His approach is based on a generalized within-group transformation; naturally extending the idea that panel data allow the researcher to control for time-invariant heterogeneity by 'differencing it away'.[5] Under regularity conditions, which ensure nonsingularity of $\boldsymbol{\beta}_0$'s information bound, Chamberlain's estimator converges at the standard $\sqrt{N}$ rate.

Nonsingularity of $\mathcal{I}(\boldsymbol{\beta}_0)$, the information for $\boldsymbol{\beta}_0$, requires the time dimension of the panel to exceed the number of random coefficients ($T > p$). Depending on the time series properties of the regressors, T may need to substantially exceed p. In extreme cases $\mathcal{I}(\boldsymbol{\beta}_0)$ may be zero for all values of T. In such settings Chamberlain's method breaks down. We show that, under mild conditions, $\boldsymbol{\beta}_0$ nevertheless remains identified. Our method of identification is necessarily 'irregular': the information bound is singular and hence no regular $\sqrt{N}$ consistent estimator exists (Chamberlain, 1986). We develop a feasible analog estimator for $\boldsymbol{\beta}_0$ and characterize its large

[4] Despite its innovative nature, and contemporary relevance given the resurgence of interest in models with heterogenous marginal effects, Chamberlain's work on the CRC model is not widely known. The CRC specification is not discussed in Chamberlain's own *Handbook of Econometrics* chapter (Chamberlain, 1984), while the panel data portion of Chamberlain (1992) is only briefly reviewed in the more recent survey by Arellano and Honoré (2001).

[5] Bonhomme (2010) further generalizes this idea, introducing a notion of 'functional differencing'.

sample properties. Although its rate of convergence is slower than the standard parametric one, its limiting distributions is normal. Inference is straightforward.

Our work shares features with other studies of irregularly identified semiparametric models (e.g., Chamberlain, 1986; Manski, 1987; Heckman, 1990; Horowitz, 1992; Abrevaya, 2000; Honoré and Kyriazidou, 1997; Kyriazidou, 1997; Andrews and Schafgans, 1998; Khan and Tamer, 2010). A general feature of irregular identification is its dependence on the special properties of small subpopulations. These special properties are, in turn, generated by specific features of the semiparametric model. Consequently these types of identification arguments tend to highlight the importance, sometimes uncomfortably so, of maintained modelling assumptions (cf., Chamberlain, 1986, pp. 205 - 207; Khan and Tamer, 2010).

Our approach exploits the different properties, borrowing a terminology introduced by Chamberlain (1982), of 'movers' and 'stayers'. Loosely speaking these two subpopulations respectively correspond to those units whose regressors values, $\mathbf{X}_t$, change and do not change across periods (a precise definition in terms of singularity of a unit-specific design matrix is given below). We identify aggregate time effects using the variation in Y_t in the 'stayers' subpopulation. A common trends assumption allows us to extrapolate these estimated effects to the entire population. Having identified the aggregate time effects using stayers, we then identify the APE by the limit of a trimmed mean of a particular unit-specific vector of regression coefficients.

Connection to other work on panel data In order to connect our work to the wider panel data literature it is useful to consider the more general outcome response function:

$$Y_t(\mathbf{x}_t) = m(\mathbf{x}_t, A, U_t).$$

Identification of the APE in the above model may be achieved by one of two main classes of restrictions. The *correlated random effects* approach invokes assumptions on the joint distribution of $(\mathbf{U}, A)|\, \mathbf{X}$; with $\mathbf{U} = (U_1, \ldots, U_T)'$. Mundlak (1978a,b) and Chamberlain (1980, 1984) develop this approach for the case where $m(\mathbf{X}_t, A, U_t)$ and $F(\mathbf{U}, A|\, \mathbf{X})$ are parametrically specified. Newey (1994a) considers a semiparametric specification for $F(\mathbf{U}, A|\, \mathbf{X})$ (cf., Arellano and Carrasco, 2003). Recently, Altonji and Matzkin (2005) and Bester and Hansen (2009) have extended this idea to the case where $m(\mathbf{X}_t, A, U_t)$ is either semi- or non-parametric along with $F(\mathbf{U}, A|\, \mathbf{X})$.

The *fixed effects* approach imposes restrictions on $m(\mathbf{X}_t, A, U_t)$ and $F(\mathbf{U}|\, \mathbf{X}, A)$, while leaving $F(A|\, \mathbf{X})$, the distribution of the time-invariant heterogeneity, the so-called 'fixed effects', unrestricted. Chamberlain (1980, 1984, 1992), Manski (1987), Honoré (1992), Abrevaya (2000), and Bonhomme (2010) are examples of this approach. Depending on the form of $m(\mathbf{X}_t, A, U_t)$, the fixed effect approach may not allow for a complete characterization of the effect of exogenous changes in $\mathbf{X}_t$ 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).

Our methods are of the 'fixed effect' variety. In addition to assuming the CRC structure for $Y_t(\mathbf{x}_t)$ we impose a marginal stationarity restriction on $F(U_t|\, \mathbf{X}, A)$, a restriction also used by

Manski (1987), Honoré (1992) and Abrevaya (2000), however, other than some weak smoothness conditions, we leave $F(A|\mathbf{X})$ unrestricted.

Wooldridge (2005b) and Arellano and Bonhomme (2009) also analyze the CRC panel data model. Wooldridge focuses 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). Arellano and Bonhomme (2009) study the identification and estimation of higher-order moments of the distribution of the random coefficients. Unlike us, they maintain Chamberlain's (1992) regularity conditions as well as impose additional assumptions.

Chamberlain (1982) showed that when $\mathbf{X}_t$ is discretely valued the APE is generally not identified (p. 13). However, Chernozhukov, Fernández-Val, Hahn and Newey (2009), working with more general forms for $\mathbb{E}[Y_t|\mathbf{X}, A]$, show that when Y_t has bounded support the APE is partially identified and propose a method of estimating the identified set.[6] In contrast, in our setup we show that the APE is point identified when at least one component of $\mathbf{X}_t$ is continuously-valued.

Section 1 presents our identification results. We begin by (i) briefly reviewing the approach of Chamberlain (1992) and (ii) characterizing irregularity in the CRC model. We then present our method of irregular identification. Section 2 outlines our estimator as well as its large sample properties. Section 3 discusses various extensions of our basic approach.

In Section 4, we use our methods to estimate the average elasticity of calorie demand with respect to total household resources in a set 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 immediately prior to the start of the RPS program over half of households had less then the required number of calories needed for all their members to engage in 'light activity' on a daily basis.[7]

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.[8] 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.[9] While most estimates of the elasticity of calorie demand are significantly positive, several recent estimates are small in value and/or imprecisely estimated, casting doubt on the value of income-oriented anti-hunger programs (Behrman and

[6] They consider the probit and logit models with unit-specific intercepts (in the index) in detail. They show how to construct bounds on the APE despite the incidental parameters problem and provide conditions on the distribution of $\mathbf{X}_t$ such that these bounds shrink as T grows.

[7] 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'.

[8] 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).

[9] Another motivation for studying this elasticity has to do with its role in theoretical models of nutrition-based poverty traps (see Dasgupta (1993) for a survey).

Deolalikar, 1987).[10]

Disagreement about the size of the elasticity of calorie demand has prompted a vigorous methodological debate in development economics. Much of this debate has centered, appropriately so, on issues of measurement and measurement error (e.g., 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.

We compare our CRC estimates of the elasticity of calorie demand with those estimated using standard panel data estimators (e.g., Behrman and Deolalikar, 1987; Bouis and Haddad, 1992), as well as those derived from cross-sectional regression techniques as in Strauss and Thomas (1990, 1995), Subramanian and Deaton (1996), and others. Our preferred CRC elasticity estimates are 10 to 30 percent smaller than their corresponding textbook linear 'fixed effects' estimates (FE-OLS). Our results are consistent with the presence of modest 'correlated random coefficients bias'.

Section 5 summarizes and suggests areas for further research. Proofs are in the Appendix. The notation $\underline{0}_T$, ι_T, I_T and $\overset{D}{=}$ respectively denotes a $T \times 1$ vector of zeros, a $T \times 1$ vector of ones, the $T \times T$ identity matrix, and equality in distribution.

1 Identification

Our benchmark data generating process combines (1) with the following assumption.

Assumption 1.1 (Stationarity and Common Trends)
(i) $b_t(A, U_t) = b^*(A, U_t) + d_t(U_{2t})$ *for* $t = 1, \ldots, T$ *and* $U_t = (U_{1t}', U_{2t}')'$;
(ii) $U_t| \mathbf{X}, A \overset{D}{=} U_s| \mathbf{X}, A$ *for* $t = 1, \ldots, T, \quad t \neq s$;
(iii) $U_{2t}| \mathbf{X}, A \overset{D}{=} U_{2t}$ *for* $t = 1, \ldots, T$;
(iv) $\mathbb{E}[b_t(A, U_t)| \mathbf{X} = \mathbf{x}]$ *exists for all* $t = 1, \ldots, T$ *and* $\mathbf{x} \in \mathbb{X}^T$.

Part (i) of Assumption 1.1 implies that the random coefficient consists of a 'stationary' and 'nonstationary' component. The stationary part, $b^*(A, U_t)$, does not vary over time so that if $U_t = U_s$ we have $b^*(A, U_t) = b^*(A, U_s)$. The non-stationary part, which is a function of the subvector U_{2t} alone, may vary over time so that even if $U_{2t} = U_{2s}$ we may have $d_t(U_{2t}) \neq d_s(U_{2s})$.

Part (ii) imposes marginal stationarity of U_t given $\mathbf{X}$ and A (cf., Manski, 1987). Stationarity implies that the joint distribution of (U_t, A) given $\mathbf{X}$ does not depend on t. This implies that time may not be used to forecast values of the unobserved heterogeneity. While (ii) allows for

[10] Wolfe and Behrman (1983), using data from Somoza-era 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.

serial dependence in U_t, it rules out time-varying heteroscedasticity. Part (iii) requires that U_{2t} is independent of both $\mathbf{X}$ and A. Maintaining (ii) and (iii) is weaker than assuming that U_t is i.i.d. over time and independent of $\mathbf{X}$ and A as is often done in nonlinear panel data research (e.g., Chamberlain, 1980). Part (iv) is a technical condition. Note that Assumption 1.1 does not restrict the joint distribution of $\mathbf{X}$ and A. Our model is a 'fixed effects' one.

Under Assumption 1.1 we have

$$\begin{aligned}
\mathbb{E}\left[b_t\left(A,U_t\right)\middle|\,\mathbf{X}\right] &= \mathbb{E}\left[b^*\left(A,U_t\right)\middle|\,\mathbf{X}\right]+\mathbb{E}\left[d_t\left(U_{2t}\right)\middle|\,\mathbf{X}\right] \\
&= \mathbb{E}\left[b^*\left(A,U_1\right)\middle|\,\mathbf{X}\right]+\mathbb{E}\left[d_t\left(U_{21}\right)\right] \\
&= \beta_0\left(\mathbf{X}\right)+\boldsymbol{\delta}_{0t}, \quad t=1,\ldots,T,
\end{aligned} \tag{6}$$

where first equality uses part (i) of Assumption 1.1, the second parts (ii) and (iii), and the third establishes the notation $\beta_0\left(\mathbf{X}\right) = \mathbb{E}\left[b^*\left(A,U_1\right)\middle|\,\mathbf{X}\right]$ and $\boldsymbol{\delta}_{0t} = \mathbb{E}\left[d_t\left(U_{2t}\right)\right]$. In what follows we normalize $\boldsymbol{\delta}_{01} = \underline{0}$.

Equation (6) is a 'common trends' assumption. To see this consider two subpopulations with different regressor histories ($\mathbf{X}=\mathbf{x}$ and $\mathbf{X}=\mathbf{x}'$). Restriction (6) implies that

$$\begin{aligned}
\mathbb{E}\left[b_t\left(A,U_t\right)\middle|\,\mathbf{x}\right]-\mathbb{E}\left[b_s\left(A,U_s\right)\middle|\,\mathbf{x}\right] &= \mathbb{E}\left[b_t\left(A,U_t\right)\middle|\,\mathbf{x}'\right]-\mathbb{E}\left[b_s\left(A,U_s\right)\middle|\,\mathbf{x}'\right] \\
&= \boldsymbol{\delta}_{0t}-\boldsymbol{\delta}_{0s}.
\end{aligned}$$

Now recall that a unit's period t potential outcome function is $Y_t\left(\mathbf{x}_t\right)=\mathbf{x}_t' b_t\left(A,U_t\right)$. Let $\boldsymbol{\tau}$ be any point in the support of both $\mathbf{X}_t$ and $\mathbf{X}_s$, we have for all $\mathbf{x}\in\mathbb{X}^T$

$$\mathbb{E}\left[Y_t\left(\boldsymbol{\tau}\right)-Y_s\left(\boldsymbol{\tau}\right)\middle|\,\mathbf{X}=\mathbf{x}\right]=\mathbb{E}\left[Y_t\left(\boldsymbol{\tau}\right)-Y_s\left(\boldsymbol{\tau}\right)\right]=\boldsymbol{\tau}'\left(\boldsymbol{\delta}_{0t}-\boldsymbol{\delta}_{0s}\right). \tag{7}$$

Equation (7) implies that while the period t (linear) potential outcome functions may vary arbitrarily across subpopulations defined in terms of $\mathbf{X}=\mathbf{x}$, shifts in these functions over time are mean independent of $\mathbf{X}$. A variant of (7) is widely-employed in the program evaluation literature (e.g., Heckman, Ichimura, Smith and Todd, 1998; Angrist and Krueger, 1999). It is also satisfied by the linear panel data model featured in Chamberlain (1984).[11]

Let the $(T-1)\,p\times 1$ vector of aggregate shifts in the random coefficients $\left(\boldsymbol{\delta}_2',\ldots,\boldsymbol{\delta}_T'\right)'$ be denoted by $\boldsymbol{\delta}$ with the corresponding $T\times(T-1)\,p$ matrix of time shifters given by

$$\mathbf{W}=\begin{pmatrix} \underline{0}_p' & & \underline{0}_p' \\ \mathbf{X}_2' & & \underline{0}_p' \\ & \ddots & \\ \underline{0}_p' & & \mathbf{X}_T' \end{pmatrix}. \tag{8}$$

[11] In an NBER working paper we show how to weaken (6) while still getting positive identification results. As we do not use these additional results when considering estimation they are omitted. Formulating a specification test based on the overidentifying implications of (6) would be straightforward.

Under Assumption 1.1 we can write the conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$ as:

$$\mathbb{E}\left[\mathbf{Y}|\,\mathbf{X}\right] = \mathbf{W}\boldsymbol{\delta}_0 + \mathbf{X}\boldsymbol{\beta}_0\left(\mathbf{X}\right). \tag{9}$$

In some cases it will be convenient to impose a priori zero restrictions on $\boldsymbol{\delta}_0$ (which would imply restrictions on how $\mathbb{E}\left[Y_t\left(\mathbf{x}_t\right)\right]$ is allowed to vary over time). In order to accommodate such situations (without introducing additional notation) we can simply redefine $\mathbf{W}$ and $\boldsymbol{\delta}_0$ accordingly. For example a model which allows only the intercept of $\mathbb{E}\left[Y_t\left(\mathbf{x}_t\right)\right]$ to shift over time is given by (9) above with $\mathbf{W} = \left(\underline{0}_{T-1}, I_{T-1}\right)'$ and $\boldsymbol{\delta}_0$ equal to the $T-1$ vector of intercept shifts. To accommodate a range of options we hereon assume that $\mathbf{W}$ is a $T \times q$ function of $\mathbf{X}$.

Equation (9), which specifies a semiparametric mean regression function for $\mathbf{Y}$ given $\mathbf{X}$, is the fundamental building block of the results that follow. Our identification results are based solely on different implications of (9). The role of equation (1) and Assumption 1.1 is to provide primitive restrictions on F_0 which imply (9). We emphasize that our results neither hinge on, nor necessarily fully exploit, all of these assumptions. Rather they flow from just one of their implications.

1.1 Regular identification

The partially linear form of (9) suggests identifying $\boldsymbol{\delta}_0$ using the conditional variation in $\mathbf{W}$ given $\mathbf{X}$ as in, for example, Engle, Granger, Rice and Weiss (1986).[12] In our benchmark model, however, $\mathbf{W}$ is a $T \times q$ function of $\mathbf{X}$ and hence no such conditional variation is available. Nevertheless Chamberlain (1992) has shown that $\boldsymbol{\delta}_0$ may be identified using the panel structure.

Let $\Phi\left(\mathbf{X}\right)$ be some function of $\mathbf{X}$ mapping into $T \times T$ positive definite matrices (in practice $\Phi\left(\mathbf{X}\right) = I_T$ will often suffice) and define the $T \times T$ idempotent 'residual maker' matrix:

$$M_\Phi\left(\mathbf{X}\right) = I_T - \mathbf{X}\left[\mathbf{X}'\Phi^{-1}\left(\mathbf{X}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\Phi^{-1}\left(\mathbf{X}\right). \tag{10}$$

Using the fact that $M_\Phi\left(\mathbf{X}\right)\mathbf{X} = 0$ Chamberlain (1992) derived, for $T > p$, the pair of moment restrictions

$$\mathbb{E}\left[\begin{array}{c} \mathbf{W}'\Phi^{-1}\left(\mathbf{X}\right)M_\Phi\left(\mathbf{X}\right)\left(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0\right) \\ \left[\mathbf{X}'\Phi^{-1}\left(\mathbf{X}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\Phi^{-1}\left(\mathbf{X}\right)\left(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0\right) - \boldsymbol{\beta}_0 \end{array}\right] = 0,$$

which identify $\boldsymbol{\delta}_0$ and $\boldsymbol{\beta}_0$ by

$$\boldsymbol{\delta}_0 = \mathbb{E}\left[\overline{\mathbf{W}}_\Phi'\Phi^{-1}\left(\mathbf{X}\right)\overline{\mathbf{W}}_\Phi\right]^{-1} \times \mathbb{E}\left[\overline{\mathbf{W}}_\Phi'\Phi^{-1}\left(\mathbf{X}\right)\overline{\mathbf{Y}}_\Phi\right] \tag{11}$$

$$\boldsymbol{\beta}_0 = \mathbb{E}\left[\left(\mathbf{X}'\Phi^{-1}\left(\mathbf{X}\right)\mathbf{X}\right)^{-1}\mathbf{X}'\Phi^{-1}\left(\mathbf{X}\right)\left(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0\right)\right], \tag{12}$$

where $\overline{\mathbf{W}}_\Phi = M_\Phi\left(\mathbf{X}\right)\mathbf{W}$ and $\overline{\mathbf{Y}}_\Phi = M_\Phi\left(\mathbf{X}\right)\mathbf{Y}$.

Note that $M_\Phi\left(\mathbf{X}\right)$ may be viewed as a generalization of the within-group transform. To see this

[12] To be specific if $\widetilde{\mathbf{W}} = \mathbf{W} - \mathbb{E}\left[\mathbf{W}|\,\mathbf{X}\right]$ has a covariance matrix of full rank, then $\boldsymbol{\delta}_0 = \mathbb{E}\left[\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}\right]^{-1} \times \mathbb{E}\left[\widetilde{\mathbf{W}}'\mathbf{Y}\right]$.

note that premultiplying (9) by $M_\Phi(\mathbf{X})$ yields

$$\begin{aligned}\mathbb{E}\left[\overline{\mathbf{Y}}_\Phi \middle| \mathbf{X}\right] &= \overline{\mathbf{W}}_\Phi \boldsymbol{\delta}_0 + M_\Phi(\mathbf{X})\mathbf{X}\boldsymbol{\beta}_{\mathbf{0}}(\mathbf{X}) \\ &= \overline{\mathbf{W}}_\Phi \boldsymbol{\delta}_0,\end{aligned}$$

so that $M_\Phi(\mathbf{X})$ 'differences away' the unobserved correlated effects, $\boldsymbol{\beta}_0(\mathbf{X})$. Equation (11) shows that $\boldsymbol{\delta}_0$ is identified by the remaining 'within-group' variation in $\mathbf{W}_t$.

With $\boldsymbol{\delta}_0$ asymptotically known, the APE is then identified by the (population) mean of the unit-specific generalized least squares (GLS) fits

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\Phi^{-1}(\mathbf{X}_i)\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\Phi^{-1}(\mathbf{X}_i)(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta}_0). \tag{13}$$

Chamberlain (1992) showed that setting $\Phi(\mathbf{X}) = \Sigma(\mathbf{X}) = \mathbb{V}(\mathbf{Y}|\mathbf{X})$ is optimal; resulting in estimators with asymptotic sampling variances equal to the variance bounds:

$$\mathcal{I}(\boldsymbol{\delta}_0)^{-1} = \mathbb{E}\left[\overline{\mathbf{W}}_\Sigma'\Sigma^{-1}(\mathbf{X})\overline{\mathbf{W}}_\Sigma\right]^{-1} \tag{14}$$

$$\mathcal{I}(\boldsymbol{\beta}_0)^{-1} = \mathbb{V}(\boldsymbol{\beta}_0(\mathbf{X})) + \mathbb{E}\left[\left(\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{X}\right)^{-1}\right] + K\mathcal{I}(\boldsymbol{\delta}_0)^{-1}K', \tag{15}$$

where $K = \mathbb{E}\left[\left(\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{X}\right)^{-1}\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{W}\right]$.

1.2 Irregularity of the CRC panel data model

Chamberlain's approach requires nonsingularity of $\mathcal{I}(\boldsymbol{\delta}_0)$ and $\mathcal{I}(\boldsymbol{\beta}_0)$. In this section we discuss when this condition does not hold and, consequently, no regular $\sqrt{N}$ consistent estimator exists. We begin by noting that singularity $\mathcal{I}(\boldsymbol{\delta}_0)$ and $\mathcal{I}(\boldsymbol{\beta}_0)$ is generic if $T = p$. The following proposition specializes Proposition 1 of Chamberlain (1992) to our problem.

Proposition 1.1 (Zero Information) *Suppose that (i) $(F_0, \boldsymbol{\delta}_0, \boldsymbol{\beta}_0(\cdot))$ satisfies (9), (ii) $\Sigma(\mathbf{x})$ is positive definite for all $\mathbf{x} \in \mathbb{X}^T$, (iii) $\mathbb{E}\left[\mathbf{W}'\Sigma^{-1}(\mathbf{X})\mathbf{W}\right] < \infty$ and (iv) $T = p$, then $\mathcal{I}(\boldsymbol{\delta}_0) = 0$.*

Proof. From Chamberlain (1992) the information bound for $\boldsymbol{\delta}_0$ is given by

$$\mathcal{I}(\boldsymbol{\delta}_0) = \mathbb{E}\left[\overline{\mathbf{W}}_\Sigma'\Sigma^{-1}(\mathbf{X})\overline{\mathbf{W}}_\Sigma\right]$$

so that $\alpha'\mathcal{I}(\boldsymbol{\delta}_0)\alpha = 0$ is equivalent to $\overline{\mathbf{W}}_\Sigma\alpha = 0$ with probability one. If $T = p$, then $\mathbf{X}$ is square so that

$$\overline{\mathbf{W}}_\Sigma = \mathbf{W}\left(I_T - \mathbf{X}\left[\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{X}\right]^{-1}\mathbf{X}'\Sigma^{-1}(\mathbf{X})\right) = 0$$

such that $\overline{\mathbf{W}}_\Sigma\alpha = 0$. ■

An intuition for Proposition 1.1 is that when $T = p$ Chamberlain's generalized within-group transform of $\mathbf{W}$ eliminates *all* residual variation in $\mathbf{W}_t$ over time. This is because the p predictors $\mathbf{X}_t$ perfectly (linearly) predict each element of $\mathbf{W}_t$ when $T = p$. Consequently the deviation of

$\mathbf{W}_t$ from its 'within-group mean' is identically equal to zero; any approach based on within-unit variation will necessarily fail.

As a simple example consider the one period ($T = p = 1$) 'panel data' model where, suppressing the t subscript,

$$Y = \boldsymbol{\delta}_0 + Xb(A, U), \tag{16}$$

with X scalar. Under Assumption 1.1 this gives (9) with $\mathbf{W} = 1$ and $\mathbf{X} = X$. The generalized within-group operator for this model is $M_\Phi(\mathbf{X}) = 1 - X\left(\frac{X^2}{\Phi(\mathbf{X})}\right)^{-1}\frac{X}{\Phi(\mathbf{X})} = 0$. Consequently $\overline{\mathbf{Y}}_I = \overline{\mathbf{W}}_I = 0$ and (11) does not identify $\boldsymbol{\delta}_0$. By Proposition 1.1 $\mathcal{I}(\boldsymbol{\delta}_0) = 0$. We show that $\boldsymbol{\delta}_0$ and $\boldsymbol{\beta}_0$ are irregularly identified in this model below.

We do not provide a general result on when regular $\sqrt{N}$ estimation of $\boldsymbol{\beta}_0$ is possible. However some insight into this question can be gleaned from a few examples. First, when $T = p$, it appears as though $\boldsymbol{\beta}_0$ will not be regularly identifiable unless $\boldsymbol{\delta}_0$ is known. This can be conjectured by the form of (15) which will generally be infinite if $\mathcal{I}(\boldsymbol{\delta}_0)^{-1}$ is. Even if $\boldsymbol{\delta}_0$ is known regular identification can be delicate. Consider the $T = p = 1$ model given above. In this model the right-hand-side of (12) above specializes to $\mathbb{E}[(Y - \boldsymbol{\delta}_0)/X]$, which will be undefined if X has positive density in the neighborhood of zero.

Less obviously, regular estimation may be impossible in heavily overidentified models (i.e., those where T substantially exceeds p).[13] To illustrate again consider (16) with $\boldsymbol{\delta}_0$ known, but with $T \geq 2$. Assume further that $\Sigma(\mathbf{X}) = I_T$ and $X_t = S \cdot Z_t$ where

$$S \sim \mathcal{U}[a, b], \quad Z_t \overset{iid}{\sim} \mathcal{N}(0, 1).$$

Variation in X_t over time in this model is governed by S, which varies across units. For those units with S close to zero, X_t will vary little across periods. The unit specific design matrix in this model is given by $\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{X} = \sum Z_t{}^2 \cdot S \sim \chi^2_T \cdot \mathcal{U}[a, b]$. If $0 < a < b$ then

$$\mathbb{E}\left[\left(\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{X}\right)^{-1}\right] = \begin{cases} \frac{\ln(b) - \ln(a)}{T-2} & T \geq 3 \\ \infty & T < 3 \end{cases},$$

so the right-hand-side of (12) will be well-defined if $T \geq 3$. If $a \leq 0$, then it is undefined *regardless of the number of time periods*. If $a \leq 0$ the support of S will contain zero, ensuring a positive density of units whose values of $\mathbf{X}_t$ do not change over time. These 'stayers' will have singular design matrices in (13), causing the variance bound for $\boldsymbol{\beta}_0$ to be infinite.

To summarize regular identification of $\boldsymbol{\beta}_0$ requires sufficient within-unit variation in $\mathbf{X}_t$ *for all units*. This is a very strong condition. Many microeconometric applications are characterized by a preponderance of stayers.[14] While time series variation in $\mathbf{X}_t$ is essential for identification,

[13] In contrast the variance bound for $\boldsymbol{\delta}_0$ will be finite when $T > p$ as long as there is *some* variation in $\mathbf{X}_t$ over time.

[14] In Card's (1996) analysis of the union wage premium, for example, less than 10 percent of workers switch between collective bargaining coverage and non-coverage across periods (Table V, p. 971).

persistence in its process is common in practice. This persistence may imply that the right-hand-side of (12) is undefined.

1.3 Irregular identification

In this section we show that, under weak conditions, $\boldsymbol{\delta}_0$ and $\boldsymbol{\beta}_0$ are irregularly identified when $T = p$. We show how to extend our methods to the irregular $T > p$ (overidentified) case in Section 3 below. Let $D = \det(\mathbf{X})$ and $\mathbf{X}^* = \operatorname{adj}(\mathbf{X})$ respectively denote the determinant and adjoint of $\mathbf{X}$ such that $\mathbf{X}^{-1} = \frac{1}{D}\mathbf{X}^*$ when the former exists.[15] In what follows we will often refer to units where $D = 0$ as *stayers*. To motivate this terminology consider the case where $T = p = 2$ with $\mathbf{W}$ and $\mathbf{X}$ in (9) equal to

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

with X_t scalar. This corresponds to a model with (i) a random intercept and slope coefficient and (ii) a common intercept shift between periods one and two. In this model

$$D = X_2 - X_1 = \Delta X;$$

hence $D = 0$ corresponds to $\Delta X = 0$, or a unit's value of X_t staying fixed across the two periods. More generally $D = 0$ if two or more rows of $\mathbf{X}$ coincide, which occurs if $\mathbf{X}_t$ does not change across adjacent periods or reverts to an earlier value in a later period. Loosely-speaking, we may think of stayers as units whose value of $\mathbf{X}_t$ changes little across periods.

Let $\mathbf{Y}^* = \mathbf{X}^*\mathbf{Y}$ and $\mathbf{W}^* = \mathbf{X}^*\mathbf{W}$ equal $\mathbf{Y}$ and $\mathbf{W}$ after premultiplication by the adjoint of $\mathbf{X}$. In the $T = p = 2$ example introduced above we have

$$\mathbf{X}^* = \begin{pmatrix} X_2 & -X_1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{Y}^* = \begin{pmatrix} X_2 Y_1 - X_1 Y_2 \\ \Delta Y \end{pmatrix}, \quad \mathbf{W}^* = \begin{pmatrix} -X_1 \\ 1 \end{pmatrix}.$$

In an abuse of notation let $\boldsymbol{\beta}_0(d) = \mathbb{E}\left[\boldsymbol{\beta}_0(\mathbf{X})\middle| D = d\right].$ Our identification result, in addition to (9), requires the following assumption.

Assumption 1.2 (Smoothness and Continuity)
(i) For some $u_0 > 0$, $D = \det(\mathbf{X})$ has $\Pr(|D| < h) = \int_{-h}^{h} \phi(u)\,\mathrm{d}u$ with $\phi(u) > 0$ for all $h \leq u_0$;
(ii) $\mathbb{E}\left[\|\mathbf{W}^*\|^2\right] < \infty$ and $\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^*\middle| D = 0\right]$ is nonsingular; and
(iii) the functions $\boldsymbol{\beta}_0(u)$, $\phi(u)$, $\mathbb{E}\left[\mathbf{W}^*\middle| D = u\right]$, and $\mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^*\middle| D = u\right]$ are all twice continuously differentiable in u for $-u_0 \leq u \leq u_0$.

Part (i) of Assumption 1.2 is essential as our approach involves conditioning on different values of D. While the requirement that D has positive density near zero is indispensable, the implication that $\Pr(D = 0) = 0$ can be relaxed. In Section 3 we show how to deal with the case where the

[15] The adjoint matrix of A is the transpose of its cofactor matrix.

distribution D has a point mass at zero. This may occur if the distribution of $\mathbf{X}_t$ has mass points at a finite set of values, while being continuously distributed elsewhere. If there is overlap in the mass points of $\mathbf{X}_t$ and $\mathbf{X}_s$ ($t \neq s$), then the distribution of D will have a mass point at zero.

Part (ii) of Assumption 1.2 is required for identification of $\boldsymbol{\delta}_0$. It will typically hold in well-specified models and is straightforward to verify. Part (iii) is a smoothness assumption which, in conjunction with (i), allows us to trim without changing the estimand.

Identification of the aggregate time effects, $\boldsymbol{\delta}_0$: We begin by premultiplying (9) by $\mathbf{X}^*$ to get

$$\mathbb{E}\left[\mathbf{Y}^* \middle| \mathbf{X}\right] = \mathbf{W}^* \boldsymbol{\delta}_0 + D \boldsymbol{\beta}_0 \left(\mathbf{X}\right),$$

where we use the fact that $DI_T = \mathbf{X}^* \mathbf{X}$. Conditioning on the subpopulation of 'stayers' yields

$$\mathbb{E}\left[\mathbf{Y}^* \middle| \mathbf{X}, D = 0\right] = \mathbf{W}^* \boldsymbol{\delta}_0. \tag{17}$$

Under Assumption 1.2 equation (17) implies that $\boldsymbol{\delta}_0$ is identified by the conditional linear predictor (CLP)

$$\boldsymbol{\delta}_0 = \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^* \middle| D = 0\right]^{-1} \times \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{Y}^* \middle| D = 0\right]. \tag{18}$$

Equation (18) shows that the subpopulation of stayers, or 'within-stayer' variation, is used to tie down the aggregate time effects, $\boldsymbol{\delta}_0$. Since stayer's correspond to units whose values of $\mathbf{X}_t$ change little over time, the evolution of Y_t among these units is driven solely by the aggregate time effects. This approach to identifying $\boldsymbol{\delta}_0$ is reminiscent of Chamberlain's (1986) 'identification at infinity' result for the intercept of the censored regression model (p. 205). Both approaches use a *small subpopulation* to tie down a feature of the *entire population*. An important difference is that our result does not require $\mathbf{X}_t$ to have unbounded support. Consequently, our identification result is not sensitive to the 'tail properties' of the distribution of $\mathbf{X}$. Our key requirement, that D have positive density in a neighborhood about zero, is straightforward to verify. We do this in the empirical application by plotting a kernel density estimate of $\phi(d)$, the density of D (see Figure 1 below).

In the $T = p = 2$ example we have, conditional on $D = 0$, the equality $\mathbf{Y}^* = \mathbf{W}^* \Delta Y$ so that (18) simplifies to, recalling that $D = \Delta X$,

$$\boldsymbol{\delta}_0 = \mathbb{E}\left[\Delta Y \middle| \Delta X = 0\right]. \tag{19}$$

The common intercept shift is identified by the average change in Y_t in the subpopulation of stayers. Identification of $\boldsymbol{\delta}_0$ is irregular since $\Pr(D = 0) = 0$; $\boldsymbol{\delta}_0$ corresponds to the value of the nonparametric mean regression of ΔY given D *at* $D = 0$. Note the importance of the (verifiable) requirement that $\phi_0 \equiv \phi(0) > 0$ for this result.

As a second example of (18) consider the one period 'panel data' model introduced above. From (16) we have

$$\mathbb{E}\left[Y \middle| X = 0\right] = \boldsymbol{\delta}_0,$$

or 'identification at zero'.

Identification of the average partial effects, $\boldsymbol{\beta}_0$: Treating $\boldsymbol{\delta}_0$ as known we identify $\boldsymbol{\beta}_0(\mathbf{x})$ for all $\mathbf{x}$ such that d is non-zero by

$$\boldsymbol{\beta}_0(\mathbf{x}) = \mathbb{E}\left[\mathbf{X}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0)\middle|\, \mathbf{X} = \mathbf{x}\right]. \tag{20}$$

It is instructive to consider the $T = p = 2$ case introduced above. In that model the second component of the right-hand side of (20), corresponding to the slope coefficient on X_t, evaluates to

$$\begin{aligned} \boldsymbol{\beta}_{20}(\mathbf{x}) &= \frac{\mathbb{E}\left[\Delta Y|\, \mathbf{X} = \mathbf{x}\right] - \boldsymbol{\delta}_0}{x_2 - x_1} \\ &= \frac{\mathbb{E}\left[\Delta Y|\, \mathbf{X} = \mathbf{x}\right] - \mathbb{E}\left[\Delta Y|\, \Delta X = 0\right]}{x_2 - x_1} \end{aligned} \tag{21}$$

where the second, difference-in-differences, equality follows by substituting in (19) above. Equation (21) indicates that the average slope coefficient, in a subpopulation homogenous in $\mathbf{X} = \mathbf{x}$, is equal to the *average* 'rise' – $\mathbb{E}\left[\Delta Y|\, \mathbf{X} = \mathbf{x}\right]$ – over the *common* 'run' – $x_2 - x_1$. The evolution of Y_t amongst stayers is used to eliminate the aggregate time effect from the average rise (i.e., to control for 'common trends') in this computation. Stayers serve as a control group.

Using (21) we then might try, by appealing to the law of iterated expectations, to identify $\boldsymbol{\beta}_{20}$ by

$$\mathbb{E}\left[\frac{\Delta Y - \boldsymbol{\delta}_0}{\Delta X}\right]. \tag{22}$$

An approach based on (22) was informally suggested by Mundlak (1961, p. 45). Chamberlain (1982) considered (22) with $\boldsymbol{\delta}_0 = 0$, showing that it identifies $\boldsymbol{\beta}_{20}$ if $\mathbb{E}\left[|\Delta Y/\Delta X|\right] < \infty$. However, if ΔX has a positive, continuous density at zero – and if $\mathbb{E}[|\Delta Y| \mid \Delta X = d] - \boldsymbol{\delta}_0$ does not vanish at $d = 0$ – then (22) will not be finite. For example, if ΔY and ΔX are independently and identically distributed according to the standard normal distribution, then $\Delta Y/\Delta X$ will be distributed according to the Cauchy distribution, whose expectation does not exist.

More generally the expectation

$$\mathbb{E}\left[\mathbf{X}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0)\right]$$

will be generally undefined if the distribution of $\mathbf{X}$ is such that D has a positive density in the neighborhood of $D = 0$ (i.e., there is a positive density of 'stayers'). This will occur when, for example, at least two rows of $\mathbf{X}$ 'nearly' coincide for 'enough' units (i.e., when part (i) of Assumption 1.2 holds).

To deal with the small denominator effects of stayers we trim. Under parts (i) and (iii) of

Assumption 1.2 we have the equalities (see equation (41) in the Appendix)

$$
\begin{aligned}
\boldsymbol{\beta}_0 &= \mathbb{E}\left[\boldsymbol{\beta}_0\left(\mathbf{X}\right)\right] \\
&= \lim_{h\downarrow 0} \mathbb{E}\left[\boldsymbol{\beta}_0\left(\mathbf{X}\right)\cdot \mathbf{1}\left(|D|>h\right)\right] \\
&= \lim_{h\downarrow 0} \mathbb{E}\left[\mathbf{X}^{-1}\left(\mathbf{Y}-\mathbf{W}\boldsymbol{\delta}_0\right)\cdot \mathbf{1}\left(|D|>h\right)\right], \qquad (23)
\end{aligned}
$$

so that $\boldsymbol{\beta}_0$ is identified by the limit of the trimmed mean of $\mathbf{X}^{-1}\left(\mathbf{Y}-\mathbf{W}\boldsymbol{\delta}_0\right)$. Trimming eliminates those units with near-singular design matrices (i.e., stayers); by taking limits and exploiting continuity we avoid changing the estimand.

Note that if there is a point mass of stayers such that $\Pr\left(D=0\right)=\pi_0>0$, then (23) does not equal $\boldsymbol{\beta}_0$, instead it equals

$$\boldsymbol{\beta}_0^M = \mathbb{E}\left[\left.\boldsymbol{\beta}_0\left(\mathbf{X}\right)\right| D\neq 0\right],$$

or the movers average partial effect (MAPE). Let $\boldsymbol{\beta}_0^S = \mathbb{E}\left[\left.\boldsymbol{\beta}_0\left(\mathbf{X}\right)\right| D=0\right]$ equal the corresponding stayers average partial effect (SAPE). In Section 3 we show how to extend our results to identify the full average partial effect $\boldsymbol{\beta}_0 = \pi_0\boldsymbol{\beta}_0^S + \left(1-\pi_0\right)\boldsymbol{\beta}_0^M$ in this case.

The following proposition, which is proven in the Appendix as a by-product of the consistency part of Theorem 2.1 below, summarizes our main identification result.

Proposition 1.2 (Irregular Identification) *Suppose that (i) $\left(F_0, \boldsymbol{\delta}_0, \boldsymbol{\beta}_0\left(\cdot\right)\right)$ satisfies (9), (ii) $\Sigma\left(\mathbf{x}\right)$ is positive definite for all $\mathbf{x}\in\mathbb{X}^T$, (iii) $T=p$, and (iv) Assumption 1.2 holds, then $\boldsymbol{\delta}_0$ and $\boldsymbol{\beta}_0$ are identified by, respectively, (18) and (23).*

2 Estimation

Our approach to estimation is to replace (18) and (23) with their sample analogs. We begin by discussing our estimator for the common parameters $\boldsymbol{\delta}_0$. Let h_N denote some bandwidth sequence such that $h_N\to 0$ as $N\to\infty$. We estimate $\boldsymbol{\delta}_0$ by the nonparametric conditional linear predictor fit:

$$\widehat{\boldsymbol{\delta}} = \left[\frac{1}{Nh_N}\sum\nolimits_{i=1}^{N}\mathbf{1}\left(|D_i|\leq h_N\right)\mathbf{W}_i^{*\prime}\mathbf{W}_i^{*}\right]^{-1}\times\left[\frac{1}{Nh_N}\sum\nolimits_{i=1}^{N}\mathbf{1}\left(|D_i|\leq h_N\right)\mathbf{W}_i^{*\prime}\mathbf{Y}_i^{*}\right]. \qquad (24)$$

Observe that $\widehat{\boldsymbol{\delta}}$ may be computed by a least squares fit of $\mathbf{Y}_i^*$ onto $\mathbf{W}_i^*$ using the subsample of units for which $|D_i|\leq h_N$. This estimator has asymptotic properties similar to a standard (uniform) kernel regression fit for a one-dimensional problem. In particular, in the proof to Lemma A.2 in the Appendix we show that

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

so that its mean squared error (MSE) rate of convergence is slower than $1/N$ when $h_N\to 0$. We also show that the leading bias term in $\widehat{\boldsymbol{\delta}}$ is quadratic in the bandwidth so that the fastest rate of convergence of $\widehat{\boldsymbol{\delta}}$ to $\boldsymbol{\delta}_0$ will be achieved when the bandwidth sequence is of the form

$$h_N^* \propto N^{-1/5}.$$

To center the limiting distribution of $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\right)$ at zero we use a bandwidth sequence that approaches zero faster than the MSE-optimal one. We discuss our chosen bandwidth sequence in more detail below.

With $\widehat{\boldsymbol{\delta}}$ in hand we then estimate $\boldsymbol{\beta}_0$ using the trimmed mean[16]

$$\widehat{\boldsymbol{\beta}} = \frac{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right) \mathbf{X}_i^{-1}\left(\mathbf{Y}_i - \mathbf{W}_i \widehat{\boldsymbol{\delta}}\right)}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right)}. \tag{25}$$

To derive the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ we begin by considering those of the infeasible estimator based on the true value of the time effects, $\boldsymbol{\delta}_0$:

$$\widehat{\boldsymbol{\beta}}_I = \frac{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right) \mathbf{X}_i^{-1}\left(\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\delta}_0\right)}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right)}. \tag{26}$$

Like $\widehat{\boldsymbol{\delta}}$ the variance of $\widehat{\boldsymbol{\beta}}_I$ is of order $1/Nh_N$, however its asymptotic bias is linear, not quadratic, in h_N. The fastest feasible rate of convergence of $\widehat{\boldsymbol{\beta}}_I$ to $\boldsymbol{\beta}_0$ is consequently slower than the that of $\widehat{\boldsymbol{\delta}}$ to $\boldsymbol{\delta}_0$ ($N^{-2/3}$ versus $N^{-4/5}$). In order to center the limiting distribution of $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\right)$ at zero we assume that $(Nh_N)^{1/2} h_N \rightarrow 0$ as as $N \rightarrow \infty$. This is stronger than what is needed to appropriately center the distribution of the aggregate time effects, where assuming $(Nh_N)^{1/2} h_N^2 \rightarrow 0$ as as $N \rightarrow \infty$ would suffice.

The value of studying the large sample properties of $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\right)$ is that our feasible estimator is a linear combination of $\widehat{\boldsymbol{\beta}}_I$ and $\widehat{\boldsymbol{\delta}}$:

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_I + \widehat{\Xi}_N\left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\right), \tag{27}$$

with

$$\widehat{\Xi}_N = \frac{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right) D_i^{-1} \mathbf{W}_i^*}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right)}. \tag{28}$$

Note that $\widehat{\boldsymbol{\beta}}_I$ and $\widehat{\boldsymbol{\delta}}$ are respectively computed using the $|D_i| > h_N$ and $|D_i| \leq h_N$ subsamples, so they are conditionally independent given the $\{\mathbf{X}_i\}$. This independence exploits the fact that the same bandwidth sequence is used to estimate $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$; it also results from our choice of the uniform kernel, which has bounded support. We proceed under these maintained assumption, acknowledging that it means that the rate of convergence of $\widehat{\boldsymbol{\delta}}$ to $\boldsymbol{\delta}_0$ is well below its optimal one.

[16] The denominator in (25) could be replaced by 1.

We view the gains from using the same bandwidth sequence for both $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$ – in terms of simplicity and transparency of asymptotic analysis – as worth the cost in generality. This approach has the further advantage in that it allows for the effect of sampling error in $\widehat{\boldsymbol{\delta}}$ on that of $\widehat{\boldsymbol{\beta}}$ to be easily characterized.

Lemma A.3 in the Appendix shows that

$$\widehat{\Xi}_N \xrightarrow{p} \Xi_0 \equiv \lim_{h \downarrow 0} \mathbb{E}\left[\mathbf{1}\left(|D_i| > h\right) D_i^{-1} \mathbf{W}_i^*\right].$$

We therefore recover the limiting distribution of the feasible estimator $\widehat{\boldsymbol{\beta}}$ from our results on $\widehat{\boldsymbol{\beta}}_I$ and $\widehat{\boldsymbol{\delta}}$ using a delta method type argument based on (27).

To formalize the above discussion and provide a precise result we require the following additional assumptions.

Assumption 2.1 (Random Sampling) *$\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^N$ are i.i.d. draws from a distribution F_0 which satisfies condition (9) above.*

Assumption 2.2 (Bounded Moments) $\mathbb{E}\left[\|\mathbf{X}_i^* \mathbf{Y}_i\|^4 + \|\mathbf{X}_i^* \mathbf{W}_i\|^4\right] < \infty$.

Assumption 2.3 (Smoothness) *The conditional expectations $\boldsymbol{\beta}_0(u)$, $\mathbb{E}\left[\mathbf{X}^* \Sigma\left(\mathbf{X}\right) \mathbf{X}^{*\prime} \middle| D = u\right]$, and $m_r\left(u\right) = \mathbb{E}\left[\|\mathbf{X}_i^* \mathbf{Y}_i\|^r + \|\mathbf{X}_i^* \mathbf{W}_i\|^r \middle| D = u\right]$ exist and are twice continuously differentiable for u in a neighborhood of zero and $0 \leq r \leq 4$.*

Assumption 2.4 (Local Identification) $\mathbb{E}\left[\mathbf{X}^* \Sigma\left(\mathbf{X}\right) \mathbf{X}^{*\prime} \middle| D = 0\right]$ *is positive definite.*

Assumption 2.5 (Bandwidth) *As $N \rightarrow \infty$ we have $h_N \rightarrow 0$ such that $Nh_N \rightarrow \infty$ and $\left(Nh_N\right)^{1/2} h_N \rightarrow 0$.*

Assumption 2.1 is a standard random sampling assumption. Our methods could be extended to consider other sampling schemes in the usual way. Assumptions 2.2 and 2.3 are regularity conditions that allow for the application of Liapunov's central limit theorem for triangular arrays (e.g., Serfling, 1980). Assumption 2.5 is a bandwidth condition which ensures that $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ is asymptotically centered at zero with a finite variance as discussed below.

The smoothness imposed by Assumption 2.3 can be restrictive. For example if $T = p = 2$ with $\mathbf{X}_t = (1, X_t)'$ and X_1 and X_2 independent exponential random variables with parameter $1/\lambda$, then $D = \Delta X$ will be a Laplace$(0, \lambda)$ random variable (the density of which is non-differentiable at zero). Non-differentiablility of the density of D at $D = 0$ will prevent us from consistently estimating the common time effects, $\boldsymbol{\delta}$ (and, consequently, also $\boldsymbol{\beta}_0$).[17] To gauge the restrictiveness of Assumption 2.3 note that twice continuous differentiability is required for nonparametric kernel estimation of, for example, $\phi(u)$ and $\mathbb{E}\left[\mathbf{W}^* \middle| D = u\right]$, and is, consequently, a standard assumption in the literature on nonparametric density and conditional moment estimation (e.g., Pagan and Ullah, 1999; Chapters 2, 3).

[17] More precisely it will invalidate our proof.

Theorem 2.1 (Large Sample Distribution) *Suppose that (i) $(F_0, \boldsymbol{\delta}_0, \boldsymbol{\beta}_0(\cdot))$ satisfies (9), (ii) $\Sigma(\mathbf{x})$ is positive definite for all $\mathbf{x} \in \mathbb{X}^T$, (iii) $T = p$, and (iv) Assumptions 1.2 to 2.5 hold, then $\widehat{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\delta}_0$ and $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ with the normal limiting distribution*

$$\sqrt{Nh_N}\begin{pmatrix} \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \Omega_0), \qquad \Omega_0 = \begin{pmatrix} \frac{\Lambda_0}{2\phi_0} & \frac{\Lambda_0 \Xi_0'}{2\phi_0} \\ \frac{\Xi_0 \Lambda_0}{2\phi_0} & 2\Upsilon_0\phi_0 + \frac{\Xi_0\Lambda_0\Xi_0'}{2\phi_0} \end{pmatrix}$$

where

$$\Lambda_0 = \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^* \middle| D = 0\right]^{-1} \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{X}^*\Sigma(\mathbf{X})\mathbf{X}^{*\prime}\mathbf{W}^* \middle| D = 0\right] \mathbb{E}\left[\mathbf{W}^{*\prime}\mathbf{W}^* \middle| D = 0\right]^{-1}$$
$$\Upsilon_0 = \mathbb{E}\left[\mathbf{X}^*\Sigma(\mathbf{X})\mathbf{X}^{*\prime} \middle| D = 0\right].$$

We comment that, in contrast to the irregularly identified semiparametric models discussed in Heckman (1990), Andrews and Schafgans (1998), and Khan and Tamer (2010), the rate of convergence for our estimator does not depend on delicate 'relative tail conditions'. Our identification approach is distinct from the type of 'identification and infinity' arguments introduced by Chamberlain (1986) and leads to a somewhat simpler asymptotic analysis.

Ex ante, that the rate of convergence of $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}_I$ coincide, might be considered surprising. While $\widehat{\boldsymbol{\delta}}$ is based on an increasingly smaller, $\widehat{\boldsymbol{\beta}}_I$ is based on an increasingly larger, fraction of the sample as $N \to \infty$. However the latter estimate increasingly includes high variance observations (i.e., units with D close to zero) as $N \to \infty$. The sampling variability induced by the inclusion of these units ensures that, in large enough samples, $\widehat{\boldsymbol{\beta}}_I$'s variance is of order $1/Nh_N$.

It is instructive to compare the asymptotic variances given in Theorem 2.1 with Chamberlain's regular counterparts (given in (14) and (15) above). First consider the asymptotic variance of $\widehat{\boldsymbol{\delta}}$. In our setup $\mathbf{W}^*$ plays a role analogous to the generalized within-group transformation of $\mathbf{W}$ used by Chamberlain (i.e., $\overline{\mathbf{W}}_\Phi = M_\Phi(\mathbf{X})\mathbf{W}$). Viewed in this light the form of Λ_0 is similar to that of $\mathcal{I}(\boldsymbol{\delta}_0)^{-1}$ in the regular case. The key difference is that (i) the expectations in Λ_0 are conditional on $D = 0$ (i.e., averages over the subpopulation of stayers) and (ii) the variance of $\widehat{\boldsymbol{\delta}}$ varies inversely with ϕ_0. The greater the density of stayers, the easier it is to estimate $\widehat{\boldsymbol{\delta}}$. We comment that we could estimate $\widehat{\boldsymbol{\delta}}$ more precisely if we replaced (24) with a weighted least squares estimator. We do not pursue this idea here as it would require pilot estimation of $\Sigma(\mathbf{X})$, a high dimensional object, and hence is unlikely to be useful in practice.

The asymptotic variance of $\widehat{\boldsymbol{\beta}}$ also parallels the form of $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$. The first term, $2\Upsilon_0\phi_0$, plays the role of $\mathbb{E}\left[\left(\mathbf{X}'\Sigma^{-1}(\mathbf{X})\mathbf{X}\right)^{-1}\right]$ in (15). This term corresponds to the average of the conditional sampling variances of the unit specific slope estimates. The 'better' the typical unit-specific design matrix, the greater the precision of the average $\widehat{\boldsymbol{\beta}}$. In the irregular case $2\Upsilon_0\phi_0$ captures a similar effect. There the average is conditional on $D = 0$. In contrast to the aggregate time effects, the first term in the variance of $\widehat{\boldsymbol{\beta}}$ varies linearly with ϕ_0; suggesting that a small density of stayers is better for estimation of $\boldsymbol{\beta}_0$.

The second term in $\widehat{\boldsymbol{\beta}}$'s variance is analogous to the $K\mathcal{I}(\boldsymbol{\delta}_0)^{-1}K'$ term in (15). This term captures the effect of sampling variation in $\widehat{\boldsymbol{\delta}}$ on that of $\widehat{\boldsymbol{\beta}}$. Note that K is equal to the average of the $p \times q$ matrix of coefficients associated with the unit-specific GLS fit of the $q \times 1$ vector $\mathbf{W}_t$ given the $p \times 1$ vector of regressors $\mathbf{X}_t$. It is instructive to consider an example where there is no asymptotic penalty associated with not knowing $\boldsymbol{\delta}_0$. Let $\mathbf{W} = \left(\underline{0}_{T-1}, I_{T-1}\right)'$ and $\mathbf{X}_t = (1, X_t)'$ with X_t scalar such that $p = 2$ and $\boldsymbol{\delta}_0$ corresponds to a $q = T - 1$ vector of time-specific intercept shifts. If the distribution of X_t is stationary over time, then realizations of X_t cannot be used to predict the time period dummies. In that case each column of K will consist of a vector of zeros with the exception of the first element (which will equal $1/T$) . The lower-right-hand element of $K\mathcal{I}\left(\boldsymbol{\delta}_0\right)^{-1} K'$ will equal zero so that ignorance of $\boldsymbol{\delta}_0$ does not affect the precision which which the second component of $\boldsymbol{\beta}_0$, corresponding to the average slope, may be estimated.[18]

Now consider the irregular case where $T = p = 2$. We have

$$\Xi_0 = \lim_{h \downarrow 0} \mathbb{E}\left[\mathbf{1}\left(|\Delta X| > h\right) \frac{1}{\Delta X} \begin{pmatrix} -X_1 \\ 1 \end{pmatrix}\right],$$

so that the lower-right-hand element of $\Xi_0 \Lambda_0 \Xi_0'$ will equal zero if $\lim_{h \downarrow 0} \mathbb{E}\left[\mathbf{1}\left(|\Delta X| > h\right) / \Delta X\right] = 0$. This condition will hold 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). This will ensure the asymptotic equivalence of the feasible estimator $\widehat{\beta}$ and its infeasible counterpart $\widehat{\beta}_I$.

Chamberlain's variance bound for $\boldsymbol{\beta}_0$ contains a third term the analog of which is not present in the irregular case. This term, $\mathbb{V}\left(\boldsymbol{\beta}_0\left(\mathbf{X}\right)\right)$, captures the effect of heterogeneity in the conditional average of the random coefficients on the asymptotic variance of $\widehat{\beta}$. In the irregular case a term equal to $\mathbb{V}\left(\boldsymbol{\beta}_0\left(\mathbf{X}\right)\right)/N$ also enters the expression for the sampling variance of $\widehat{\beta}$ (see the calculations immediately prior to Equation (44) in the Appendix). However this term is asymptotically dominated by the two terms listed in Theorem 2.1 (which are of order $1/Nh_N$). The variance estimator described in Theorem 2.2 below implicitly accounts for this asymptotically dominated component.

The conditions of Theorem 2.1 place only weak restrictions on the bandwidth sequence. As is common in the semiparametric literature we deal with bias by undersmoothing. Let a be a $p \times 1$ vector of known constants, the appendix shows that the fastest rate of convergence of $a'\widehat{\boldsymbol{\beta}}$ for $a'\boldsymbol{\beta}_0$ in mean square is achieved by bandwidth sequences of the form,

$$h_N^* = C_0 N^{-1/3},$$

where the mean squared error minimizing choice of constant is

$$C_0 = \frac{1}{2}\left(\frac{1}{\phi_0}\right)^{1/3} \left\{ \frac{a'\left(2\Upsilon_0 + \frac{\Xi_0 \Lambda_0 \Xi_0'}{2\phi_0^2}\right) a}{a'\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)' a} \right\}^{1/3}, \tag{29}$$

[18]Sampling error in the estimated time effects does affect the precision with which the common intercept, the first component of $\boldsymbol{\beta}_0$, may be estimated.

and $\boldsymbol{\beta}_0^S = \mathbb{E}\left[\boldsymbol{\beta}_0(\mathbf{X})\right| D = 0]$ equals the average of the random coefficients in the subpopulation of stayers. While the bandwidth sequence h_N^* achieves the fastest rate of convergence for our estimator, the corresponding asymptotic normal distribution for $a'\widehat{\beta}(h_N^*)$ will be centered at a bias term of $2a'\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)\phi_0$. To eliminate this bias Assumption 2.5 requires that $h_N \to 0$ fast enough such that $(Nh_N)^{1/2} h_N \to 0$ as $N \to \infty$, but slow enough such that $(Nh_N)^{1/2} \to \infty$. A bandwidth sequence which converges to zero slightly faster than h_N^* is sufficient for this purpose. In particular if

$$h_N = o(N^{-1/3}),$$

then $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ will be asymptotically centered at zero.

An alternative to undersmoothing would be to use a plug in bandwidth based on a consistent estimate of (29), say $\widehat{C}$. Such an approach is taken by Horowitz (1992) in the context of smoothed maximum score estimation. Denote the resulting estimate by $a'\widehat{\boldsymbol{\beta}}_{\mathrm{PI}}$ (PI for 'plug in'). Let $a'\widehat{\boldsymbol{\beta}}$ be the consistent undersmoothed estimate of Theorem 2.1 and $\widehat{\boldsymbol{\beta}}^S$ and $\widehat{\phi}_0$ estimates of $\boldsymbol{\beta}_0^S$ and ϕ_0. The bias corrected estimate is then

$$a'\widehat{\boldsymbol{\beta}}_{\mathrm{BC}} = a'\widehat{\boldsymbol{\beta}}_{\mathrm{PI}} - 2a'\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^S\right)\widehat{\phi}_0\widehat{C}N^{-1/3}.$$

Unlike undersmoothing, this does not slow down the rate of convergence of $\widehat{\boldsymbol{\beta}}_{\mathrm{BC}}$ to $\boldsymbol{\beta}_0$. A disadvantage is that it is more computationally demanding. In the empirical application below we experiment with a number of bandwidth values. A more systematic analysis of bandwidth selection, while beyond the scope of this paper, would be an interesting topic for further research.

Computation and consistent variance estimation: The computation of $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\beta}}$ is facilitated by observing that the solutions to (24) and (25) above coincide with those of the linear instrumental variables fit

$$\widehat{\boldsymbol{\theta}} = \left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{Q}_i'\mathbf{R}_i\right]^{-1} \times \left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{Q}_i'\mathbf{Y}_i^*\right],$$

for $\boldsymbol{\theta} = \left(\boldsymbol{\delta}', \boldsymbol{\beta}'\right)'$ and

$$\underset{T\times q+p}{\mathbf{Q}_i} = \left(h_N^{-1}\mathbf{1}\left(|D_i| \le h_N\right)\mathbf{W}_i^*, \frac{\mathbf{1}\left(|D_i| > h_N\right)}{D_i}I_p\right), \quad \underset{T\times q+p}{\mathbf{R}_i} = \left(\mathbf{W}_i^*, \mathbf{1}\left(|D_i| > h_N\right)D_iI_p\right),$$

where the dependence of $\mathbf{Q}_i$ and $\mathbf{R}_i$ on h_N is suppressed.

Let $\boldsymbol{\theta}(h)$ denote the probability limit of $\widehat{\boldsymbol{\theta}}$ when the bandwidth is held fixed at h; then, by standard GMM arguments,

$$\widehat{V}(h) = \left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{Q}_i'\mathbf{R}_i\right]^{-1} \times \left[\frac{h}{N}\sum\nolimits_{i=1}^{N}\mathbf{Q}_i'\widehat{\mathbf{U}}_i^{+}\widehat{\mathbf{U}}_i^{+\prime}\mathbf{Q}_i\right] \times \left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{Q}_i'\mathbf{R}_i\right]^{-1} \tag{30}$$

is a consistent estimate of the asymptotic covariance of $\sqrt{Nh}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}(h)\right)$ with

$$\widehat{\mathbf{U}}_i^{+} = \mathbf{Y}_i^{*} - \mathbf{R}_i \widehat{\boldsymbol{\theta}}.$$

Conveniently, this covariance estimator remains valid when, as is required by Theorem 2.1, the bandwidth shrinks with N.

Theorem 2.2 *Suppose the hypotheses of Theorem 2.1 hold, and that* $\mathbb{E}\left[\left\|\mathbf{X}_i^{*}\mathbf{Y}_i\right\|^8 + \left\|\mathbf{X}_i^{*}\mathbf{W}_i\right\|^8\right] < \infty$ *and that Assumption 2.3 holds for* $r \leq 8$*. Then* $\widehat{V}_N \equiv \widehat{V}(h_N) \xrightarrow{p} \Omega_0$.

Relative to a direct estimate of Ω_0, (30) implicitly includes estimates of terms that, while asymptotically negligible, may be sizeable in small samples. Consequently confidence intervals constructed using it may have superior properties (cf., Newey, 1994b; Graham, Imbens and Ridder, 2009).

Operationally, estimation and inference may proceed as follows. Let $\mathbf{Y}_{it}^{*}$, $\mathbf{R}_{it}$ and $\mathbf{Q}_{it}$ denote the t^{th} rows of their corresponding matrices. Using standard software compute the linear instrumental variables fit of $\mathbf{Y}_{it}^{*}$ onto $\mathbf{R}_{it}$ using $\mathbf{Q}_{it}$ as the instrument (exclude the default constant term from this calculation). By Theorem 2.2 the 'robust/clustered' (at the unit-level) standard errors reported by the program will be asymptotically valid under the conditions of Theorem 2.1.

3 Extensions

In this section we briefly develop four direct extensions of our basic results. In Section 5 we discuss other possible generalizations and avenues for future research.

3.1 Linear functions of $\boldsymbol{\beta}_0(\mathbf{X})$

In some applications the elements of $\mathbf{X}_t$ may be functionally related. For example

$$\mathbf{X}_t = \left(1, R_t, R_t^2, \ldots, R_t^{p-1}\right)'. \tag{31}$$

In such settings $\boldsymbol{\beta}_0$ indexes the average structural function (ASF) of Blundell and Powell (2003). To emphasize the functional dependence write $\mathbf{X}_t = \mathbf{x}_t(R_t)$, then

$$g_t(\tau) = \mathbf{x}_t(\tau)'\left(\boldsymbol{\beta}_0 + \boldsymbol{\delta}_{0t}\right),$$

gives the expected period t outcome if (i) a unit is drawn at random from the (cross sectional) population and (ii) she is exogenously assigned input level $R_t = \tau$. Similarly, differences of the form

$$g_t\left(\tau'\right) - g_t(\tau),$$

give the average period t outcome difference across two counterfactual policies: one where all units are exogenously assigned input level $R_t = \tau'$ and another where they are assigned $R_t = \tau$. Since it is a linear function of $\boldsymbol{\beta}_0$ and $\boldsymbol{\delta}_{0t}$ Theorem 2.1 can be used to conduct inference on $g_t(\tau)$.

In the presence of functional dependence across the elements of $\mathbf{X}_t$ the derivative of $g_t(\tau)$ with respect to τ does not correspond to an average partial effect (APE).[19] Instead such derivatives characterize the local curvature of the ASF. In such settings the average effect of a population-wide unit increase in R_t (i.e., the APE) is instead given by

$$\begin{aligned}\boldsymbol{\gamma}_{0t} \overset{def}{\equiv} \mathbb{E}\left[\frac{\partial Y_t(\mathbf{x}_t(r_t))}{\partial r_t}\right] &= \mathbb{E}\left[\left(\frac{\partial \mathbf{x}_t(r_t)}{\partial r_t}\right)' b_t(A, U_t)\right] \\ &= \mathbb{E}\left[\left(\frac{\partial \mathbf{x}_t(r_t)}{\partial r_t}\right)' (\boldsymbol{\beta}_0(\mathbf{x}_t(R_t)) + \boldsymbol{\delta}_{0t})\right],\end{aligned} \tag{32}$$

where the second equality follows from iterated expectations and Assumption 1.1. Because $\partial \mathbf{x}_t(R_t)/\partial r_t$ may covary with $\boldsymbol{\beta}_0(\mathbf{x}_t(R_t))$ Theorem 2.1 cannot be directly applied to conduct inference on $\boldsymbol{\gamma}_{0t}$. Fortunately it is straightforward to extend our methods to identify and consistently estimate parameters of the form

$$\begin{aligned}\boldsymbol{\gamma}_{0t} &= \mathbb{E}\left[\Pi(\mathbf{X})(\boldsymbol{\beta}_0(\mathbf{X}) + \boldsymbol{\delta}_{0t})\right] \\ &= \boldsymbol{\gamma}_0 + \mathbb{E}\left[\Pi(\mathbf{X})\right]\boldsymbol{\delta}_{0t},\end{aligned}$$

where $\Pi(\mathbf{x})$ is a known function of $\mathbf{x}$ and $\boldsymbol{\gamma}_0 = \mathbb{E}\left[\Pi(\mathbf{X})\boldsymbol{\beta}_0(\mathbf{X})\right]$. If $\mathbf{X}_t$ is given by (31), for example, then to estimate the APE we would choose

$$\Pi(\mathbf{x}) = \frac{\partial \mathbf{x}_t(r_t)}{\partial r_t} = \left(0, 1, 2r_t, 3r_t^2 \ldots, (p-1)\, r_t^{p-2}\right).$$

In order to estimate $\boldsymbol{\gamma}_{0t}$ we proceed as follows.[20] First, identification and estimation of $\boldsymbol{\delta}_0$ is unaffected. Second, using (20) gives for any $\mathbf{x}$ with $d \neq 0$

$$\boldsymbol{\gamma}_0(\mathbf{x}) = \Pi(\mathbf{x})\,\mathbb{E}\left[\mathbf{X}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0)\middle|\,\mathbf{X} = \mathbf{x}\right],$$

so that

$$\boldsymbol{\gamma}_0 = \lim_{h \downarrow 0} \mathbb{E}\left[\Pi(\mathbf{X})\,\mathbf{X}^{-1}(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0) \cdot \mathbf{1}(|D| > h)\right].$$

This suggests the analog estimator

$$\widehat{\boldsymbol{\gamma}} = \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| > h_N)\,\Pi(\mathbf{X}_i)\,\mathbf{X}_i^{-1}\left(\mathbf{Y}_i - \mathbf{W}_i\widehat{\boldsymbol{\delta}}\right)}{\frac{1}{N}\sum_{i=1}^N \mathbf{1}(|D_i| > h_N)}.$$

[19] We thank a referee for several helpful comments on this point.

[20] It is possible that while $\boldsymbol{\beta}_0$ is only irregularly identified, $\mathbb{E}\left[\Pi(\mathbf{X})\boldsymbol{\beta}_0(\mathbf{X})\right]$ is regularly identified. Consider the $T = p = 2$, $q = 1$ example introduced above. If $\Pi(\mathbf{X}) = (1, X_1)$, then $\mathbb{E}\left[\Pi(\mathbf{X})\boldsymbol{\beta}_0(\mathbf{X})\right] = \mathbb{E}\left[\mathbb{E}\left[Y_1|\,\mathbf{X}\right]\right] = \mathbb{E}\left[Y_1\right]$ is clearly regularly identified. What follows, for simplicity, assumes that $\mathbb{E}\left[\Pi(\mathbf{X})\boldsymbol{\beta}_0(\mathbf{X})\right]$ is not regularly identified.

An argument essentially identical to that justifying Theorem 2.1 then gives

$$\sqrt{Nh_N}\left(\begin{array}{c}\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\\ \widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\end{array}\right)\xrightarrow{D}\mathcal{N}\left(0,\left(\begin{array}{cc}\frac{\Lambda_0}{2\phi_0} & \frac{\Lambda_0\Xi_{\Pi,0}'}{2\phi_0}\\ \frac{\Xi_{\Pi,0}\Lambda_0}{2\phi_0} & 2\Upsilon_{\Pi,0}\phi_0+\frac{\Xi_{\Pi,0}\Lambda_0\Xi_{\Pi,0}'}{2\phi_0}\end{array}\right)\right), \tag{33}$$

with

$$\begin{aligned}\Upsilon_{\Pi,0} &= \mathbb{E}\left[\Pi\left(\mathbf{X}\right)\mathbf{X}^{*}\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\Pi\left(\mathbf{X}\right)'\middle|\, D=0\right]\\ \Xi_{\Pi,0} &= \lim_{h\downarrow 0}\mathbb{E}\left[\mathbf{1}\left(\left|D\right|>h\right)\Pi\left(\mathbf{X}\right)\mathbf{X}^{-1}\mathbf{W}\right].\end{aligned}$$

We can then estimate $\boldsymbol{\gamma}_{0t}$ by

$$\widehat{\boldsymbol{\gamma}}_t=\widehat{\boldsymbol{\gamma}}+\widehat{\Pi}\widehat{\boldsymbol{\delta}},$$

with $\widehat{\Pi}=\sum_{i=1}^{N}\Pi\left(\mathbf{X}_i\right)/N$. To conduct inference on $\boldsymbol{\gamma}_{0t}$ we use the delta method treating $\widehat{\Pi}$ as known. We may ignore the effects of sampling variability in $\widehat{\Pi}$ since its rate of convergence to $\mathbb{E}\left[\Pi\left(\mathbf{X}\right)\right]$ is $1/N$.

3.2 Density of D has a point mass at $D=0$

In some settings a positive fraction of the population may be stayers such that $\pi_0\overset{def}{\equiv}\Pr\left(D=0\right)>0$. This may occur even if all elements of X_t are continuously-valued. If the only continuous component of $\mathbf{X}_t$ is the logarithm of annual earnings, for example, a positive fraction of individuals may have the same earnings level in each sampled period. This may be especially true if many workers are salaried.

A point mass at $D=0$ simplifies estimation of $\boldsymbol{\delta}_0$ and complicates that of $\boldsymbol{\beta}_0$. When $\pi_0>0$ the estimator

$$\widehat{\boldsymbol{\delta}}=\left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{1}\left(D_i=0\right)\mathbf{W}_i^{*\prime}\mathbf{W}_i^{*}\right]^{-1}\times\left[\frac{1}{N}\sum\nolimits_{i=1}^{N}\mathbf{1}\left(D_i=0\right)\mathbf{W}_i^{*\prime}\mathbf{Y}_i^{*}\right],$$

will be $\sqrt{N}$-consistent and asymptotically normal for $\boldsymbol{\delta}_0$, as would be the (asymptotically equivalent) estimator described in Section 2 above.

The large sample properties of the infeasible estimator $\widehat{\boldsymbol{\beta}}^{I}$ – see equation (26) – are unaffected by the point mass at $D=0$ with two important exceptions. First its probability limit is no longer $\boldsymbol{\beta}_0$, the (full population) average partial effect, but $\boldsymbol{\beta}_0^{M}=\mathbb{E}\left[\boldsymbol{\beta}_0\left(\mathbf{X}\right)\middle|\, D\neq 0\right]$, the movers' APE introduced in Section 1. Second, its asymptotic variance is scaled up by $1-\pi_0$, the population proportion of movers. This gives $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I-\boldsymbol{\beta}_0^{M}\right)\xrightarrow{D}\mathcal{N}\left(0,\frac{2\Upsilon_0\phi_0}{1-\pi_0}\right)$.

Reflecting the change of plims let $\widehat{\boldsymbol{\beta}}^{M}$ equal the feasible estimator defined by (25). Using decomposition (27) we have

$$\begin{aligned}\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}^M-\boldsymbol{\beta}_0^M\right) &= \sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I-\boldsymbol{\beta}_0^M\right)+\widehat{\Xi}_N\sqrt{Nh_N}\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\right)\\ &= \sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I-\boldsymbol{\beta}_0^M\right)+\Xi_0 O_p\left(\sqrt{h_N}\right)\\ &= \sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I-\boldsymbol{\beta}_0^M\right)+o_p(1),\end{aligned}$$

so that the sampling properties of $\widehat{\boldsymbol{\beta}}^M$ are unaffected by those of $\widehat{\boldsymbol{\delta}}$. In particular, adapting the argument used to show Theorem 2.1 yields

$$\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}^M-\boldsymbol{\beta}_0^M\right)\xrightarrow{D}\mathcal{N}\left(0,\frac{2\Upsilon_0\phi_0}{1-\pi_0}\right).$$

If a consistent estimator of the stayers effect

$$\boldsymbol{\beta}_0^S\equiv\mathbb{E}[\boldsymbol{\beta}_0(\mathbf{X})|\,D=0]$$

can be constructed, a corresponding consistent estimator of the APE $\boldsymbol{\beta}_0=\pi_0\boldsymbol{\beta}_0^S+(1-\pi_0)\boldsymbol{\beta}_0^M$ would be

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

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

Inspection of the equation immediately preceding (17) in Section 1 suggests one possible estimator for $\boldsymbol{\beta}_0^S$. We have

$$\mathbb{E}\left[\mathbf{Y}^*|\,\mathbf{X}\right]=\mathbf{W}^*\boldsymbol{\delta}_0+D\boldsymbol{\beta}_0(\mathbf{X}),$$

so that

$$\boldsymbol{\beta}_0^S=\lim_{h\downarrow 0}\frac{\mathbb{E}\left[\mathbf{Y}^*|\,D=h\right]-\mathbb{E}\left[\mathbf{Y}^*|\,D=0\right]}{h},$$

which suggests the estimator

$$\begin{pmatrix}\overline{\boldsymbol{\delta}}\\ \widehat{\boldsymbol{\beta}}^S\end{pmatrix}=\underset{\boldsymbol{\delta},\boldsymbol{\beta}^S}{\arg\min}\sum_{i=1}^N\mathbf{1}\left(|D_i|\leq h_N\right)\left(\mathbf{Y}_i^*-\mathbf{W}_i^*\boldsymbol{\delta}-D_i\boldsymbol{\beta}^S\right)'\left(\mathbf{Y}_i^*-\mathbf{W}_i^*\boldsymbol{\delta}-D_i\boldsymbol{\beta}^S\right),$$

with $\overline{\boldsymbol{\delta}}$ an alternative $\sqrt{N}$-consistent estimator for $\boldsymbol{\delta}_0$. 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 of the combined estimator $\widehat{\boldsymbol{\beta}}\equiv\widehat{\pi}\widehat{\boldsymbol{\beta}}^S+(1-\widehat{\pi})\widehat{\boldsymbol{\beta}}^M$ will coincide with that of $\widehat{\boldsymbol{\beta}}^S$, and the asymptotic distribution of the latter would dominate the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ in this setting. We comment that part (iii) of assumption 1.2 may be less plausible in settings where $\Pr(D=0)>0$.[21] It such settings 'stayers' may be very different from 'near stayers' such that a local linear regression approach to estimating $\widehat{\boldsymbol{\beta}}^S$ would be problematic.

[21] We thank a referee for this observation.

3.3 Overidentification $(T > p)$

When $T > p$ the vector of common parameters $\boldsymbol{\delta}_0$ may be $\sqrt{N}$ consistently estimated, as first suggested by Chamberlain (1992), by the sample counterpart of (11) above:

$$\widehat{\boldsymbol{\delta}} = \left[\frac{1}{N}\sum_{i=1}^{N}\overline{\mathbf{W}}'_{\Phi i}\Phi_i\overline{\mathbf{W}}_{\Phi i}^{-1}\right]^{-1} \times \left[\frac{1}{N}\sum_{i=1}^{N}\overline{\mathbf{W}}'_{\Phi i}\Phi_i\overline{\mathbf{Y}}_{\Phi i}\right],$$

with $\Phi_i \equiv \Phi(\mathbf{X}_i)$ positive definite with probability one.

The discussion in Section 1, however, suggests that Chamberlain's (1992) proposed estimate of $\boldsymbol{\beta}_0$, the sample average of

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

for $\widehat{\boldsymbol{\delta}}$ a $\sqrt{N}$-consistent estimator of $\boldsymbol{\delta}_0$, may behave poorly and will be formally inconsistent when $\mathcal{I}(\boldsymbol{\beta}_0) = 0$.

Adapting the trimming scheme introduced for the just identified $T = p$ case, a natural modification of Chamberlain's (1992) estimator is

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

If

$$\mathbb{E}\left[\frac{1}{\det(\mathbf{X}_i'\Phi_i^{-1}\mathbf{X}_i)}\right] < \infty, \tag{34}$$

then the introduction of trimming is formally unnecessary but may still be helpful in practice. It is straightforward to show asymptotic equivalence of the (infeasible) trimmed mean

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{N}\mathbf{1}(\det(\mathbf{X}_i'\Phi_i^{-1}\mathbf{X}_i) > h_N)\cdot(\mathbf{X}_i'\Phi_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\Phi_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta}_0)}{\sum_{i=1}^{N}\mathbf{1}(\det(\mathbf{X}_i'\Phi_i^{-1}\mathbf{X}_i) > h_N)},$$

with Chamberlain's (1992) proposal when $\mathbb{E}[\boldsymbol{\beta}(\mathbf{X})|\det(\mathbf{X}_i'\Phi_i^{-1}\mathbf{X}_i) \leq h]$ is smooth (Lipschitz-continuous) in h, condition (34) holds, and $h_N = o(1/\sqrt{N})$. Since $\hat{\beta}$ will still be consistent for β even when (34) fails, a feasible version of the trimmed mean $\hat{\beta}$ may be better behaved in finite samples if the design matrix $(\mathbf{X}_i'\Phi_i^{-1}\mathbf{X}_i)$ is nearly singular for some observations.

3.4 Additional regressors

Our benchmark model assumes that $\mathbf{W}$ is a known function of $\mathbf{X}$. Let $\mathbf{V}$ be a $T \times r$ matrix of additional regressors and assume that, in place of (9), we have the following conditional moment restriction.

$$\mathbb{E}[\mathbf{Y}|\mathbf{X},\mathbf{V}] = \mathbf{V}\boldsymbol{\zeta}_0 + \mathbf{W}\boldsymbol{\delta}_0 + \mathbf{X}\boldsymbol{\beta}_0(\mathbf{X}). \tag{35}$$

Such a model might arise if, instead of (1), we have $Y_t = \mathbf{V}_t'\boldsymbol{\zeta}_0 + \mathbf{X}_t' b_t(A, U_t)$ with $\mathbf{V}$ varying independently of $(A, \mathbf{U})$.

Assume that $\widetilde{\mathbf{V}} = \mathbf{V} - \mathbb{E}[\mathbf{V}|\,\mathbf{X}]$ has a covariance matrix of full rank. Following Engle, Granger, Rice and Weiss (1986) we have

$$\boldsymbol{\zeta}_0 = \mathbb{E}\left[\widetilde{\mathbf{V}}'\widetilde{\mathbf{V}}\right]^{-1} \times \mathbb{E}\left[\widetilde{\mathbf{V}}'\mathbf{Y}\right],$$

which, under regularity conditions, is also $\sqrt{N}$ estimable (e.g., Robinson, 1988). Letting $\widehat{\boldsymbol{\zeta}}$ be such a consistent estimate we may proceed as described in Section 2 after replacing $\mathbf{Y}$ with $\mathbf{Y} - \mathbf{V}\widehat{\boldsymbol{\zeta}}$. Since the rate of convergence of $\widehat{\boldsymbol{\zeta}}$ to $\boldsymbol{\zeta}_0$ is $1/N$, we conjecture that Theorem 2.1 will remain valid with $\Sigma(\mathbf{X})$ redefined to equal $\mathbb{V}(\mathbf{Y} - \mathbf{V}\boldsymbol{\zeta}_0|\,\mathbf{X})$. Model (35) indicates that while the feasible number of random coefficients is restricted by the length of the available panel, the overall number of regressors need not be.

4 Application

In this section we use our methods to estimate the elasticity of calorie demand using the panel dataset described in the introduction. Our goal is to provide a concrete illustration of our methods, to compare them with alternatives which presume the absence of any nonseparable correlated heterogeneity, and to highlight the practical importance of trimming.

Model specification: We assume that the logarithm of total household calorie availability per capita in period t, $\ln(\mathrm{Cal}_t)$, varies according to

$$\ln(\mathrm{Cal}_t) = b_{0t}(A, U_t) + b_{1t}(A, U_t)\ln(\mathrm{Exp}_t), \tag{36}$$

where Exp_t denotes real household expenditure per capita (in thousands of 2001 Cordobas) in year t and $b_{0t}(A, U_t)$ and $b_{1t}(A, U_t)$ are random coefficients; the latter equals the household-by-period-specific elasticity of calorie demand. Let $b_t(A, U_t) = (b_{0t}(A, U_t), b_{0t}(A, U_t))'$, $\mathbf{X}_t = (1, \ln(\mathrm{Exp}_t))'$, and $Y_t = \ln(\mathrm{Cal}_t)$ with $\mathbf{X}$ and $\mathbf{Y}$ as defined above. We allow for common intercept and slope shifts over time (i.e., we maintain Assumption 1.1).

Relative to prior work, the distinguishing feature of our model is that it allows for the elasticity of calorie demand to vary across households in a way that may co-vary with total outlay. This allows household expenditures to co-vary with the unobserved determinants of calorie demand. For example both expenditures and calorie consumption are likely to depend on labor supply decisions (cf., Strauss and Thomas, 1990). Allowing the calorie demand curve to vary across households also provides a nonparametric way to control for differences in household composition; a delicate modelling decision in this context (e.g., Subramanian and Deaton, 1996).[22]

[22] A limitation of our model is its presumption of linearity at the household level. Strauss and Thomas (1990) argue that the elasticity of demand should *structurally* decline with household income. As we have three periods of data

Data descriptions: 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). Here we analyze a balanced panel of 1,358 households interviewed in the fall of 2000, 2001 and 2002. We focus on the latter two years of data (see below). The Supplemental Web Appendix describes the construction of our dataset in detail.

Tables 1 and 2 summarize some key features of our estimation sample. Panel A of Tables 1 give the share of total food spending devoted to each of eleven broad food categories. 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 (Tables 1 and 2). 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., 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.[23]

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.

Assessment of required assumptions: Assumption 1.1 allows us to tie down common time effects using the subpopulation of stayers alone. The appropriateness of using stayers in this way depends on their comparability with movers. Assumption 1.1 will also be more plausible under conditions of relative macroeconomic stability.[24]

we could, in principal, include an additional function of Exp_t in the $\mathbf{X}_t$ vector. We briefly explore this possibility in the Supplemental Web Appendix.

[23] In October of 2001 the Cordoba-to-US$ exchange rate was 13.65. Therefore per capita consumption levels in our sample averaged less than US$ 300 per year.

[24] Heuristically the hope is that the impacts of any misspecification of aggregate time effects will be muted when such effects are small in magnitude.

Panel A:	Expenditure Shares (%)								
	All			Lower 25%			Upper 25%		
	2000	2001	2002	2000	2001	2002	2000	2001	2002
Cereals	49.1	36.0	32.7	53.3	40.9	35.7	45.7	31.6	29.4
Roots	1.3	3.1	2.7	1.3	2.6	2.0	1.5	3.6	3.6
Pulses	11.6	12.5	13.6	11.2	13.8	16.5	10.6	10.7	11.3
Vegetables	3.2	4.9	4.5	2.8	4.3	3.4	3.8	5.8	5.3
Fruit	0.6	0.9	1.1	0.5	0.7	0.9	0.8	1.2	1.2
Meat	3.1	6.9	7.7	2.2	4.0	5.1	5.3	9.9	10.4
Dairy	11.2	14.7	17.3	9.0	12.0	15.0	13.1	16.8	19.2
Oil	4.0	5.0	5.0	3.5	5.2	5.0	3.9	4.7	4.7
Other foods	15.8	16.0	15.4	16.2	16.7	16.5	15.4	15.7	14.9
Staples$^{\Diamond}$	62.1	51.6	49.0	65.8	57.3	54.1	57.8	45.9	44.3
Panel B:	Total Real Expenditure & Calories								
Expenditure per adult$^{\flat}$	5,506	4,679	4,510	2,503	2,397	2,200	9,481	7,578	7,460
(Expenditure per capita)	(4,277)	(3,764)	(3,887)	(2,016)	(2,131)	(2,102)	(7,302)	(5,845)	(6,114)
Food share	71.2	69.2	68.8	73.8	69.1	68.6	67.0	67.9	67.6
Calories per adult$^{\flat}$	2,701	3,015	2,948	1,706	2,127	2,013	3,737	3,849	3,758
(Calories per capita)	(2,086)	(2,435)	(2,529)	(1,351)	(1,854)	(1,873)	(2,842)	(2,962)	(3,041)
Percent energy deficient$^{\natural}$	51.0	39.3	39.7	85.0	69.7	76.2	19.8	14.5	13.0

Table 1: Real expenditure food budget shares of RPS households from 2000 to 2002

NOTES: Authors' calculations based on a balanced panel of 1,358 households from the RPS evaluation dataset (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 nominal 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). Lower and upper 25 percent refers to the bottom and top quartiles of households based on the average of year 2000, 2001 and 2002 real consumption per adult equivalent and thus contains the same set of households in all three 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.

With these considerations in mind we use the 2001 and 2002 waves of RPS data for our core analysis. Coffee production is important in the regions from which our data were collected. While coffee prices fell sharply between the 2000 and 2001 waves of data collection, they were more stable between the 2001 and 2002 waves (see Figure 2 in the Supplemental Web Appendix). Panel B of Table 1 indicates that while per capita household expenditures fell, on average, over 10 percent between 2000 and 2001, they were roughly constant, again on average, between 2001 and 2002.

We also informally compared 'stayers' and 'near-stayers' in terms of observables. Such comparisons provide a heuristic way of assessing the plausibility of the assumption that $\boldsymbol{\beta}(d)$ is smooth in d in the neighborhood of $d = 0$.[25] Using the bandwidth value underlying our preferred estimates (see Column 5 of Table 3 and Figure 1), we define stayers as units for which $D \in [-h_N, h_N]$ and near-stayers as units for which $D \in [-1.5h_N, h_N)$ or $D \in (h_N, 1.5h_N]$. We find that average expenditures across these two sets of households were nearly identical in 2001. We also compared their demographic structures. Across 16 age-by-gender categories we found significant differences (at the 10 percent level) in 1 categories in 2001 and 0 category in 2002. We conclude that our stayer and near-stayer subsamples are broadly comparable, although we acknowledge that our 'tests', in addition to being heuristic, are likely to have low power given our available sample size.

Panel B of Figure 1 plots a kernel density estimate of the change in $\ln(\mathrm{Exp}_t)$ between 2001 and 2002. As required by Assumption 1.2 there is substantial density in the neighborhood of zero. Furthermore there is no obvious evidence of a point mass at zero so that, in large samples, the mover and full average partial effects will coincide.

While we took great care in construction our expenditure and calories available variables, measurement error in each of them cannot be ruled out. We nevertheless proceed under the maintained assumption of no measurement error. An extension of our methods to accommodate measurement error would be an interesting topic for future research.

Results: Table 3 reports our point estimates. Our first estimate corresponds to the pooled ordinary least squares (OLS) fit of $\ln(\mathrm{Cal}_t)$ onto $\ln(\mathrm{Exp}_t)$ using all three waves of the RPS data. Aggregate shifts in the intercept and slope coefficient are included (throughout we use 2001 as the base year). Also included in the model, to control for variation in food prices across markets, is a vector of 42 village-specific intercepts. Variants of this specification are widely employed in empirical work (e.g, Subramanian and Deaton, 1996; Table 2). The pooled OLS calorie elasticities are reported in Column 1. The elasticity approximately equals 0.7 in 2000 and 0.6 in both 2001 and 2002. All three elasticities are precisely determined. The estimates are high relative to others in the literature, but realistic given the extreme poverty of the households in our sample.

Column 2 augments the first model by allowing the intercept to vary across households. This 'fixed effects' estimator (FE-OLS) is also widely used in empirical work when panel data are available (e.g., Behrman and Deolalikar, 1987; Bouis and Haddad, 1992). Allowing for household-specific

[25] The intuition is that if stayers and near-stayers are observationally very different, then it is plausible that the distribution of calorie demand schedules across the two subpopulations also differs.

	Calorie Shares (%)								
	All			Lower 25%			Upper 25%		
	2000	2001	2002	2000	2001	2002	2000	2001	2002
Cereals	57.7	60.3	59.9	60.7	63.9	62.0	55.5	57.1	57.4
Roots	1.5	1.5	1.6	1.9	1.5	1.2	1.6	1.8	2.1
Pulses	13.1	11.3	12.8	12.1	11.3	13.3	13.1	11.0	12.1
Vegetables	0.7	0.7	0.6	0.6	0.6	0.4	0.8	0.9	0.8
Fruit	0.3	0.5	0.4	0.3	0.3	0.4	0.5	0.7	0.6
Meat	0.7	1.3	1.3	0.5	0.7	0.7	1.3	1.9	1.9
Dairy	4.1	4.3	4.5	3.4	3.0	3.4	4.7	5.2	5.5
Oil	6.9	7.6	7.5	5.8	6.9	6.7	7.4	8.1	8.0
Other foods	15.0	12.6	11.4	14.7	11.9	11.9	15.2	13.2	11.5
Staples$^{\Diamond}$	72.3	73.1	74.3	74.7	76.7	76.6	70.2	69.9	71.7

Table 2: Calorie shares of RPS households from 2000 to 2002
NOTES: Authors' calculations based on a balanced panel of 1,358 households from the RPS evaluation dataset (see IFPRI (2005)). Total calorie availability is calculated using the RPS food quantity data and the calorie content and edible portion information contained in INCAP (2000). Lower and upper 25 percent refers to the bottom and top quartiles of households based on the average of year 2000, 2001 and 2002 real consumption per adult equivalent and thus contains the same set of households in all three years.
$^{\Diamond}$ Sum of cereal, roots and pulses.

intercepts increases the elasticity by about 10 percent in all three years. The standard errors almost double in size.

In Column 3 we use Chamberlain's (1992) regular correlated random coefficients (R-CRC) estimator with an identity weight matrix. Since we have three years of data and only two random coefficients his methods, at least in principle, apply. The top panel of Figure 1 plots a histogram of $\det(\mathbf{X}'\mathbf{X})$, which shows a reasonable amount of density in the neighborhood of zero. This suggests that the right-hand-side of (12) may be undefined in the population. In practice the R-CRC estimator generates 'sensible' point estimates with estimated standard errors approximately equal to those of the corresponding FE-OLS estimates. The R-CRC point estimates are smaller than both the OLS and FE-OLS ones. Column 3 of Panel B implements the trimmed version of Chamberlain's procedure described in Section 3 above. In this case trimming leaves the point estimates more or less unchanged, with a slight increase in their measured precision.

Columns 4 and 5 are based on only the 2001 and 2002 waves of data. By dropping the first wave of data we artificially impose that $T = p = 2$; this ensures irregularity (Proposition 1.1). Column 4 reports the 'Mundlak' estimate of the demand elasticity

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\ln(\mathrm{Cal}_{2002})-\ln(\mathrm{Cal}_{2001})}{\ln(\mathrm{Exp}_{2002})-\ln(\mathrm{Exp}_{2001})}.$$

This average, as expected, is poorly behaved. It generates a much lower elasticity estimate with a very large standard error.

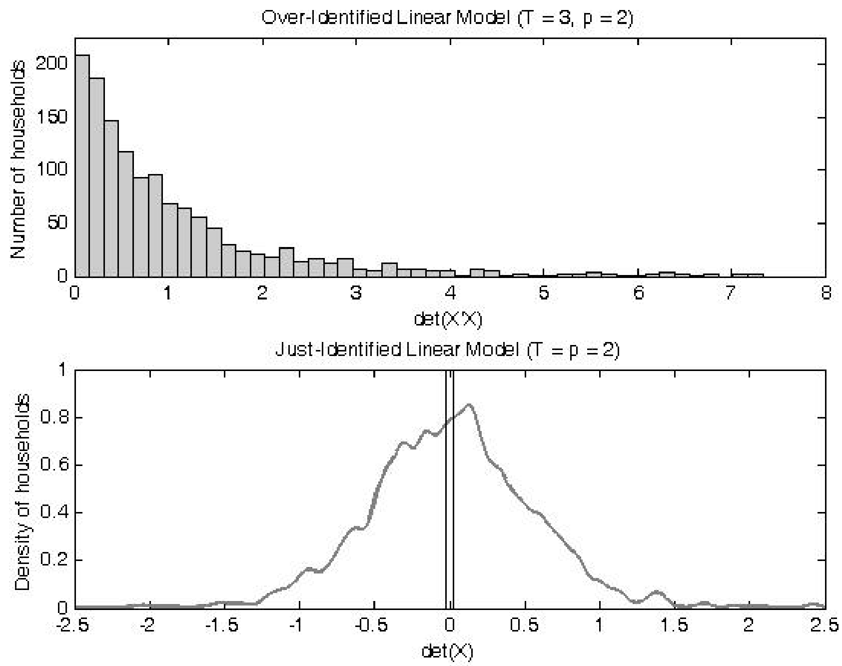

Figure 1: Histogram of the distribution of det $(\mathbf{X}'\mathbf{X})$ (top panel, $T = 3$, $p = 2$) and kernel density estimate of the distribution of D (bottom panel, $T = p = 2$)
NOTES: The two vertical lines in the lower panel correspond to the portion of the sample that is trimmed in our preferred estimates (Table 3, Column 5). A normal kernel and bandwidth of $h_N = c_D N^{-1/3}$ where $c_D = \min(s_D, r_D/1.34)$ is a robust estimate of the sample standard deviation of D are used to construct the density estimate (s_D is the sample standard deviation and r_D the interquartile range).

Column 5 implements our estimator (I-CRC) with a bandwidth of $h_N = \frac{c_D}{2} N^{-1/3}$ where $c_D = \min(s_D, r_D/1.34)$ is a robust estimate of the sample standard deviation of D (s_D is the sample standard deviation and r_D the interquartile range).[26] This implies that we trim, or categorize as 'stayers', about 4 percent of our sample. In contrast to its untrimmed counterpart, the I-CRC point estimate is sensible and well-determined. The estimated year 2001 and 2002 elasticities are over 25 percent small than their FE counterparts (Column 2). Panel B of the table explores the sensitivity of our I-CRC point estimates to trimming. We find that doubling the fraction of the sample categorized as stayers substantially improves estimated precision, but also shifts the point estimates upwards. Halving the fraction of stayers substantially reduces estimated precision (Columns 1 & 2 of Panel B). Overall we find that while the Column 5 point estimates are somewhat sensitive to modest variations in the bandwidth, they consistently lie below their FE-OLS counterparts.

[26] This bandwidth value corresponds to Silverman's well-known normal reference rule-of-thumb bandwidth for density estimation. We divide by 2 to adjust for the fact that our uniform kernel integrates to 2 instead of 1.

5 Conclusion

In this paper we have outlined a new estimator for the correlated random coefficients panel data model. Our estimator is designed for situations where the regularity conditions required for the method-of-moments procedure of Chamberlain (1992) do not hold. We illustrate the use of our methods in a study of the elasticity of demand for calories in a population of poor Nicaraguan households. This application is highly irregular, with many 'near stayers' in the sample. This implies that elasticity estimates based on the textbook FE-OLS estimator may be far from the relevant population average. We find that our methods work well in this setting, generating point estimates that are as much as 25 percent smaller in magnitude that their FE-OLS counterparts (Table 3, Columns 5 versus 2).

While our procedure is simple to implement, it does require choosing a smoothing parameter. As in other areas of semiparametric econometrics, our theory places only weak restrictions on this choice. Developing an automatic, data-based, method of bandwidth selection would be useful.

Irregularity arises in other fixed effects panel data models (e.g., Manski, 1987; Chamberlain, 2010; Honoré and Kyriazidou, 1997; Kyriazidou, 1997; Hoderlein and White, 2009). It is an open question as to whether features of our approach could be extended to more complex nonlinear and/or dynamic panel data models. In ongoing work we are studying how to extend our methods to estimate quantile partial effects (e.g., unconditional quantiles of the distribution of the random coefficients) and to accommodate additional 'triangular endogeneity'.

	Panel A: Calorie Demand Elasticities					Panel B: Sensitivity to Trimming		
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)
	OLS	FE	R-CRC	Naive	I-CRC	I-CRC	I-CRC	R-CRC
2000 Elasticity	0.6837 (0.0305)	0.7550 (0.0441)	0.6617 (0.0424)	–	–	–	–	0.6913 (0.0425)
2001 Elasticity	0.6105 (0.0383)	0.6635 (0.0608)	0.5861 (0.0565)	0.2444 (0.9491)	0.4800 (0.1202)	0.6040 (0.2023)	0.6087 (0.0745)	0.6157 (0.0501)
2002 Elasticity	0.5959 (0.0245)	0.6466 (0.0416)	0.5521 (0.0476)	–	0.4543 (0.1130)	0.5491 (0.1232)	0.5477 (0.0607)	0.5816 (0.0397)
Percent trimmed	–	–	–	0	3.8	2	8	4
Intercept shifters?	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Slope shifters?	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes

Table 3: Estimates of the calorie Engel curve: linear case

NOTES: Estimates based on the balanced panel of 1,358 households described in the main text. "OLS" denotes least squares applied to the pooled 2000, 2001, and 2002 samples, "FE-OLS" least squares with household-specific intercepts, "R-CRC" Chamberlain's (1992) estimator with identity weight matrix, "MDLK" the Mundlak (1961)/Chamberlain (1982) estimator described in the main text, and "I-CRC" our irregular correlated random coefficients estimator (using the 2001 and 2002 waves only). All models, with the exception of "MDLK", include common intercept and slope shifts across periods. The standard errors are computed in a way that allows for arbitrary within-village correlation in disturbances across households and time.

Appendix

This appendix contains a proof of Theorem 2.1. Some auxiliary Lemmas, as well as a proof of Theorem 2.2, may be found in the Supplemental Web Appendix.

A Proof of Theorem 2.1

As noted in the main text our derivation of the limiting distribution of $\widehat{\boldsymbol{\beta}}$ utilizes the decomposition

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_I + \widehat{\Xi}_N \left(\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 \right). \tag{37}$$

with $\widehat{\boldsymbol{\beta}}_I$, $\widehat{\boldsymbol{\delta}}$, and $\widehat{\Xi}_N$ respectively equal to (26), (24), and (28) of the main text. The proof proceeds in three steps. First, we derive the limiting distribution of the infeasible estimator $\widehat{\boldsymbol{\beta}}_I$. Second, that of the common parameters $\widehat{\boldsymbol{\delta}}$. Third, we show that $\widehat{\Xi}_N$ has a well-defined probability limit. The limiting distribution of $\widehat{\boldsymbol{\beta}}$ then follows from the delta method and the independence of $\widehat{\boldsymbol{\beta}}_I$ and $\widehat{\boldsymbol{\delta}}$.

Large sample properties of $\widehat{\boldsymbol{\beta}}_I$: We begin with the infeasible estimator (26) which treats $\boldsymbol{\delta}_0$ as known. Recentering (26) yields

$$\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0 = \frac{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right)\left(\mathbf{X}_i^{-1}\left(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta}_0\right) - \boldsymbol{\beta}_0\right)}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|D_i| > h_N\right)}. \tag{38}$$

First consider the expected value of the term entering the summation in the denominator of (38):

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}\left(|D_i| > h\right)\right] &= 1 - \Pr\{|D_i| \le h\} \\ &= 1 - h\int_{-1}^{1} \phi\left(uh\right) \mathrm{d}u \\ &= 1 - 2h\phi_0 + o(h), \end{aligned} \tag{39}$$

where the second equality follows from Assumption 1.2 and the change of variables $u = t/h$ (with Jacobian $\mathrm{d}t/\mathrm{d}u = h$).

Define $Z_{N,i}$ to be the term entering the summation in the numerator of (38):

$$Z_{N,i} \equiv \mathbf{1}\left(|D_i| > h\right)\left(\mathbf{X}_i^{-1}\left(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta}_0\right) - \boldsymbol{\beta}_0\right). \tag{40}$$

Taking its expectation yields

$$\begin{aligned} \mathbb{E}\left[Z_{N,i}\right] &= \mathbb{E}\left[\mathbf{1}\left(|D_i| > h\right) \cdot \left(\boldsymbol{\beta}_0\left(\mathbf{X}_i\right) - \boldsymbol{\beta}_0\right)\right] \\ &= \mathbb{E}\left[\mathbf{1}\left(|D_i| \le h\right)\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0\left(D_i\right)\right)\right] \\ &= \int_{-h}^{h} (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0\left(t\right))\phi\left(t\right) \mathrm{d}t \\ &= 2\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)\phi_0 h + o\left(h\right), \end{aligned} \tag{41}$$

where $\boldsymbol{\beta}_0^S \equiv \boldsymbol{\beta}_0\left(0\right)$, again using Assumption 1.2.

Turning to the variance of $Z_{N,i}$ we use the ANOVA decomposition

$$\mathbb{V}\left(Z_{N,i}\right) = \mathbb{V}\left(\mathbb{E}\left[\left.Z_{N,i}\right| D_i\right]\right) + \mathbb{E}\left[\mathbb{V}\left(\left.Z_{N,i}\right| D_i\right)\right]. \tag{42}$$

The first term in (42) equals

$$\begin{aligned}
\mathbb{V}\left(\mathbb{E}\left[\left.Z_{N,i}\right| D_i\right]\right) &= \mathbb{V}\left(\mathbf{1}\left(|D_i| > h\right)\mathbb{E}\left[\left.\left(\boldsymbol{\beta}_0\left(\mathbf{X}_i\right) - \boldsymbol{\beta}_0\right)\right| D_i\right]\right) \\
&= \mathbb{V}\left(\mathbf{1}\left(|D_i| > h\right)\left(\boldsymbol{\beta}_0\left(D_i\right) - \boldsymbol{\beta}_0\right)\right) \\
&= \mathbb{V}\left(\boldsymbol{\beta}_0\left(D_i\right)\right) + o\left(1\right).
\end{aligned}$$

Now consider $\mathbb{V}\left(\left.Z_{N,i}\right| D_i\right)$; using (41) above and recalling the equality $\mathbf{X}^{-1} = \frac{1}{D}\mathbf{X}^*$ when $|D| > 0$, we have

$$\begin{aligned}
Z_{N,i} - \mathbb{E}\left[\left.Z_{N,i}\right| D_i\right] &= \mathbf{1}\left(|D_i| > h\right)\left\{\mathbf{X}_i^{-1}\left(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta}\right) - \boldsymbol{\beta}_0 - \left(\boldsymbol{\beta}_0\left(D_i\right) - \boldsymbol{\beta}_0\right)\right\} \\
&= \frac{\mathbf{1}\left(|D_i| > h\right)}{D_i}\left\{\mathbf{Y}_i^* - \mathbf{W}_i^*\boldsymbol{\delta} - D_i\boldsymbol{\beta}_0\left(D_i\right)\right\} \\
&= \frac{\mathbf{1}\left(|D_i| > h\right)}{D_i}\mathbf{X}_i^*\left(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\beta}_0\left(D_i\right)\right).
\end{aligned}$$

Again defining

$$\begin{aligned}
\mathbf{U}_i &\equiv \mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\beta}_0\left(\mathbf{X}_i\right) \\
&= \mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\beta}_0\left(D_i\right) + \mathbf{X}_i(\boldsymbol{\beta}_0\left(\mathbf{X}_i\right) - \boldsymbol{\beta}_0\left(D_i\right)),
\end{aligned}$$

it follows from iterated expectations that

$$\begin{aligned}
\mathbb{V}\left(\left.Z_{N,i}\right| D_i\right) &= \frac{\mathbf{1}\left(|D_i| > h_N\right)}{D_i^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\left(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\beta}_0\left(D_i\right)\right)\left(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\delta} - \mathbf{X}_i\boldsymbol{\beta}_0\left(D_i\right)\right)'\mathbf{X}_i^{*\prime}\right| D_i\right] \\
&= \frac{\mathbf{1}\left(|D_i| > h_N\right)}{D_i^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i\right] + \mathbf{1}\left(|D_i| > h_N\right)\mathbb{E}\left[\left.\left(\boldsymbol{\beta}\left(\mathbf{X}_i\right) - \boldsymbol{\beta}_0\left(D_i\right)\right)\left(\boldsymbol{\beta}\left(\mathbf{X}_i\right) - \boldsymbol{\beta}_0\left(D_i\right)\right)'\right| D_i\right] \\
&= \frac{\mathbf{1}\left(|D_i| > h_N\right)}{D_i^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i\right] + \mathbf{1}\left(|D_i| > h_N\right)\mathbb{V}\left(\left.\boldsymbol{\beta}\left(\mathbf{X}_i\right)\right| D_i\right),
\end{aligned}$$

where $\boldsymbol{\Sigma}(\mathbf{X}_i) \equiv \mathbb{V}\left(\left.\mathbf{U}_i\right| \mathbf{X}_i\right)$. Averaging the first term in $\mathbb{V}\left(\left.Z_{N,i}\right| D_i\right)$ over the distribution of D_i gives

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbf{1}\left(|D_i| > h_N\right)}{D_i^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i\right]\right] &= \int_{-\infty}^{-h}\frac{1}{t^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = t\right]\phi\left(t\right)\mathrm{d}t \\
&\quad + \int_{h}^{\infty}\frac{1}{t^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = t\right]\phi\left(t\right)\mathrm{d}t \\
&= \frac{1}{h}\int_{-\infty}^{-1}\frac{1}{u^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = uh\right]\phi\left(uh\right)\mathrm{d}u \\
&\quad + \frac{1}{h}\int_{1}^{\infty}\frac{1}{u^2}\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = uh\right]\phi\left(uh\right)\mathrm{d}u \\
&= \frac{2\mathbb{E}\left[\left.\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = 0\right]\phi_0}{h} + O\left(1\right) \\
&= O(h^{-1}),
\end{aligned} \tag{43}$$

where the third equality exploits Assumptions 1.2 and 2.3.

Averaging the second term over the distribution of D_i yields

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left(|D_i|>h_N\right)\mathbb{V}\left(\boldsymbol{\beta}\left(\mathbf{X}_i\right)|\,D_i\right)\right] &\leq \mathbb{E}\left[\mathbb{V}\left(\boldsymbol{\beta}\left(\mathbf{X}_i\right)|\,D_i\right)\right] \\
&= \mathbb{V}\left(\boldsymbol{\beta}\left(\mathbf{X}_i\right)\right) \\
&= O\left(1\right).
\end{aligned}$$

Thus, combing terms,

$$\mathbb{E}\left[\mathbb{V}\left(Z_{N,i}|\,D_i\right)\right]=\frac{2\mathbb{E}\left[\mathbf{X}_i^{*}\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}|\,D_i=0\right]\phi_0}{h}+O\left(1\right).$$

Combing this result with the expression for $\mathbb{V}\left(\mathbb{E}\left[Z_{N,i}|\,D_i\right]\right)$ derived above yields a variance term of

$$\begin{aligned}
\mathbb{V}\left(Z_{N,i}\right) &= \frac{2\mathbb{E}\left[\mathbf{X}_i^{*}\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}|\,D_i=0\right]\phi_0}{h}+O\left(1\right) \\
&= O\left(h^{-1}\right).
\end{aligned} \tag{44}$$

Together (41), (44), and the independence generated by random sampling (Assumption 2.1) imply that

$$\mathbb{E}\left[\frac{1}{N}\sum\nolimits_{i=1}^{N}Z_{N,i}\right]=o_p\left(1\right),\qquad \mathbb{V}\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}Z_{N,i}\right)=O_p\left(\frac{1}{Nh_N}\right)=o_p\left(1\right), \tag{45}$$

under the bandwidth assumption (Assumption 2.5). This implies weak consistency of $\widehat{\boldsymbol{\beta}}_I$ for $\boldsymbol{\beta}_0$ (and indirectly Proposition 1.2).

To show asymptotic normality we need to check the conditions for Liapunov's CLT for triangular arrays. Repeated use of the inequality $||x+y||^3\leq 8(||x||^3+||y||^3)$ yields

$$\begin{aligned}
\mathbb{E}\left[\|Z_{N,i}\|^3\right] &= \mathbb{E}\left[\mathbf{1}\left(|D_i|>h_N\right)\left\|\left(\mathbf{X}_i^{-1}\left(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\delta}_0\right)-\boldsymbol{\beta}_0\right)\right\|^3\right] \\
&= \mathbb{E}\left[\mathbf{1}\left(|D_i|>h_N\right)\left\|\frac{1}{D_i}\mathbf{X}_i^{*}\left(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\delta}_0\right)-\boldsymbol{\beta}_0\right\|^3\right] \\
&\leq 8\mathbb{E}\left[\mathbf{1}\left(|D_i|>h_N\right)\left\|\frac{1}{D_i}\mathbf{X}_i^{*}\left(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\delta}_0\right)\right\|^3\right]+O(1) \\
&\leq 64(1+||\delta_0||^3)\mathbb{E}\left[\frac{\mathbf{1}\left(|D_i|>h_N\right)}{|D_i|^3}\mathbb{E}\left(\|\mathbf{X}_i^{*}\mathbf{Y}_i\|^3+||\mathbf{X}_i^{*}\mathbf{W}_i\|^3\,\Big|\,D_i\right)\right]+O(1) \\
&\leq 64(1+||\delta_0||^3)\mathbb{E}\left[\frac{\mathbf{1}\left(|D_i|>h_N\right)}{|D_i|^3}m_3(D_i)\right]+O(1),
\end{aligned}$$

where $m_3\left(D_i\right)$ is defined in Assumption 2.3. Choosing $\bar{u}>h$ sufficiently small that $m_3(u)\leq\bar{m}_3$ and $\phi(u)\leq\bar{\phi}$ when $|u|\leq\bar{u}$,

$$\begin{aligned}
\mathbb{E}\left[\|Z_{N,i}\|^3\right] &\leq 64(1+||\delta_0||^3)\int_h^{\infty}\frac{1}{t^3}\left(m_3\left(t\right)\phi(t)+m_3(-t)\,\phi\left(-t\right)\right)\mathrm{d}t+O(1) \\
&\leq 128(1+||\delta_0||^3)\bar{m}_3\bar{\phi}\int_{h_N}^{\bar{u}}\left(\frac{1}{t^3}\right)\mathrm{d}t+O(1) \\
&= O(h_N^{-2}).
\end{aligned}$$

Using the above result we can verify the Liapunov condition. Let $a_N = \left(\frac{N}{h_N}\right)$, then

$$\frac{1}{a_N}\sum\nolimits_{i=1}^{N}\mathbb{V}\left(Z_{N,i}\right) \to 2\mathbb{E}\left[\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = 0\left.\right]\phi_0,$$

and also

$$\begin{aligned}\frac{\left(\sum_{i=1}^{N}\mathbb{E}\left[\left\|Z_{N,i}-\mathbb{E}\left[Z_{N,i}\right]\right\|^3\right]\right)^{1/3}}{\left(a_N\right)^{1/2}} &\leq \frac{\left(8\sum_{i=1}^{N}\mathbb{E}\left[\left\|Z_{N,i}\right\|^3\right]\right)^{1/3}}{\left(a_N\right)^{1/2}}\\ &= O\left(\left(Nh\right)^{-1/6}\right)\\ &= o_p\left(1\right).\end{aligned}$$

Application of the Liapunov CLT for triangular arrays, equation (39) above, and Slutsky's Theorem, then yields the following Lemma.

Lemma A.1 *Suppose that (i)* $(F_0, \boldsymbol{\delta}_0, \boldsymbol{\beta}_0(\cdot))$ *satisfies (9), (ii)* $\Sigma(\mathbf{x})$ *is positive definite for all* $\mathbf{x} \in \mathbb{X}^T$*, (iii)* $T = p$*, and (iv) Assumptions 1.2 to 2.5 hold, then* $\widehat{\boldsymbol{\beta}}_I \xrightarrow{p} \boldsymbol{\beta}_0$ *with the normal limiting distribution*

$$\sqrt{Nh_N}\left(\widehat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\right) \xrightarrow{D} \mathcal{N}\left(0, 2\Upsilon_0\phi_0\right),$$

for $\Upsilon_0 = \mathbb{E}\left[\mathbf{X}_i^*\Sigma\left(\mathbf{X}_i\right)\mathbf{X}_i^{*\prime}\right| D_i = 0\left.\right]$.

Large sample properties of $\widehat{\boldsymbol{\delta}}$: Recall that the non-random coefficients $\boldsymbol{\delta}_0$ are estimated by a uniform conditional linear predictor (CLP) estimator. Recentering (24) yields

$$\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 = \left[\frac{1}{Nh}\sum\nolimits_{i=1}^{N}\mathbf{1}\left(\left|D_i\right| \leq h\right)\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right]^{-1} \times \left[\frac{1}{Nh}\sum\nolimits_{i=1}^{N}\mathbf{1}\left(\left|D_i\right| \leq h\right)\mathbf{W}_i^{*\prime}\left(D_i\boldsymbol{\beta}_0\left(\mathbf{X}_i\right) + \mathbf{U}_i^*\right)\right], \tag{46}$$

where

$$\begin{aligned}\mathbf{U}^* &= \mathbf{Y}^* - \mathbf{W}^*\boldsymbol{\delta}_0 - D\boldsymbol{\beta}_0\left(\mathbf{X}\right) &(47)\\ &= \mathbf{X}^*\left(\mathbf{Y} - \mathbf{W}\boldsymbol{\delta}_0 - \mathbf{X}\boldsymbol{\beta}_0\left(\mathbf{X}\right)\right)\\ &= \mathbf{X}^*\mathbf{U}.\end{aligned}$$

First consider the expected value of the matrix being inverted in (46). Manipulations similar to those used to analyze $\widehat{\boldsymbol{\beta}}_I$ above yield

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left(\left|D_i\right| \leq h\right)\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right] &= \mathbb{E}\left[\mathbf{1}\left(\left|D_i\right| \leq h\right)\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right| D_i\left.\right]\right]\\ &= \int_{-h}^{h}\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right| D_i = t\left.\right]\phi\left(t\right)\mathrm{d}t\\ &= h\int_{-1}^{1}\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right| D_i = uh\left.\right]\phi\left(uh\right)\mathrm{d}u\\ &= 2\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right| D_i = 0\left.\right]\phi_0 h + o\left(h\right), &(48)\end{aligned}$$

while for any fixed q-dimensional vector $\boldsymbol{\lambda}$ the variance of a quadratic form in that matrix satisfies

$$
\begin{aligned}
\mathbb{V}\left[\frac{1}{Nh}\sum_{i=1}^{N}\mathbf{1}\left(|D_i|\leq h\right)\left(\boldsymbol{\lambda}'\mathbf{W}_i^{*\prime}\mathbf{W}_i^{*}\boldsymbol{\lambda}\right)\right] &\leq \frac{1}{Nh^2}\mathbb{E}\left[\mathbf{1}\left(|D_i|\leq h\right)\mathbb{E}\left[\left\|\mathbf{W}_i^{*}\right\|^4\right| D_i\right]\right]\left\|\boldsymbol{\lambda}\right\|^4 \qquad (49)\\
&= \frac{2\mathbb{E}\left[\left\|\mathbf{W}_i^{*}\right\|^4\right| D_i=0\right]\phi_0\left\|\boldsymbol{\lambda}\right\|^4}{Nh}+o\left(\frac{1}{Nh}\right)\\
&= O\left(\frac{1}{Nh}\right),
\end{aligned}
$$

under Assumptions 2.3 and 2.5 so that

$$
\frac{1}{Nh}\sum_{i=1}^{N}\mathbf{1}\left(|D_i|\leq h\right)\mathbf{W}_i^{*\prime}\mathbf{W}_i^{*}=2\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^{*}\right| D_i=0\right]\phi_0+o_p\left(1\right). \qquad (50)
$$

Now redefine $Z_{N,i}$ to equal the term entering the summation in the numerator of (46):

$$
Z_{N,i}\equiv\mathbf{1}\left(|D_i|\leq h\right)\mathbf{W}_i^{*\prime}\left(D_i\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)+\mathbf{U}_i^{*}\right).
$$

Using the fact that $\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{U}_i^{*}\right| D_i]=\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^{*}\mathbb{E}\left[\mathbf{U}|\,\mathbf{X}\right]\right| D_i]=0$ yields an expected value of $Z_{N,i}$ equal to

$$
\begin{aligned}
\mathbb{E}\left[Z_{N,i}\right] &= \mathbb{E}\left[\mathbf{1}\left(|D_i|\leq h\right)\mathbf{W}_i^{*\prime}\left(D_i\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)+\mathbf{U}_i^{*}\right)\right]\\
&= \mathbb{E}\left[\mathbf{1}\left(|D_i|\leq h\right)D_i\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i\right]\right]\\
&= \int_{-h}^{h}t\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=t\right]\phi\left(t\right)\mathrm{d}t\\
&= h\int_{-1}^{1}uh\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=uh\right]\phi\left(uh\right)\mathrm{d}u\\
&= \mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=0\right]\phi_0 h^2\int_{-1}^{1}u\mathrm{d}u\\
&\quad+\left\{\left.\frac{\partial\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=d]}{\partial d}\right|_{d=0}\phi_0+\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=0\right]\phi_0'\right\}h^3\int_{-1}^{1}u^2\mathrm{d}u+o\left(h^3\right)\\
&= \frac{2}{3}\left\{\left.\frac{\partial\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=d]}{\partial d}\right|_{d=0}\phi_0+\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=0\right]\phi_0'\right\}h^3+o\left(h^3\right),
\end{aligned}
$$

where we use the following Taylor approximation and Assumption 1.2 and 2.3 in deriving the second to last equality above:

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=uh\right]uh\phi\left(uh\right) &= 0+\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=0\right]\phi_0 uh\\
&\quad+\left\{\left.\frac{\partial\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=d]}{\partial d}\right|_{d=0}\phi_0+\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=0\right]\phi_0'\right\}(uh)^2+o\left(h^2\right).
\end{aligned}
$$

The numerator (46) therefore equals

$$
\begin{aligned}
&\frac{1}{Nh}\sum_{i=1}^{N}\mathbf{1}\left(|D_i|\leq h\right)\mathbf{W}_i^{*\prime}\left(D_i\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)+\mathbf{U}_i^{*}\right) \qquad (51)\\
=\ &\frac{2}{3}\left\{\left.\frac{\partial\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=d]}{\partial d}\right|_{d=0}\phi_0+\mathbb{E}\left[\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right| D_i=0\right]\phi_0'\right\}h^2+o_p\left(h^2\right).
\end{aligned}
$$

Using the ratio of (51) and (50) yields a bias expression for $\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0$ of

$$\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0=\frac{1}{3}\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\right|D_i=0\right]^{-1}\times\left\{\left.\frac{\partial\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=d\right]}{\partial d}\right|_{d=0}+\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=0\right]\frac{\phi_0'}{\phi_0}\right\}h^2+o_p\left(h^2\right). \tag{52}$$

This implies that we can center the asymptotic distribution of $\sqrt{Nh_N}\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\right)$ at zero by choosing h_N such that $\left(Nh_N\right)^{1/2}h_N^2\rightarrow 0$ (Assumption 2.5).

Now consider the variance of $Z_{N,i}$. As before we proceed by evaluating the two terms in in the variance decomposition (42) separately. The first of the two terms evaluates to

$$\begin{aligned}
\mathbb{V}\left(\mathbb{E}\left[\left.Z_{N,i}\right|D_i\right]\right) &= \mathbb{V}\left(\mathbf{1}\left(\left|D_i\right|\leq h\right)D_i\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]\right)\\
&= \mathbb{E}\left[\mathbf{1}\left(\left|D_i\right|\leq h\right)D_i^2\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]'\right]\\
&\quad -\mathbb{E}\left[\mathbf{1}\left(\left|D_i\right|\leq h\right)D_i\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]\right]\mathbb{E}\left[\mathbf{1}\left(\left|D_i\right|\leq h\right)D_i\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]\right]'.
\end{aligned}$$

Evaluating the two expectations entering the above expressions yields

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left(\left|D_i\right|\leq h\right)D_i\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]\right] &= \int_{-h}^{h}t\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=t\right]\phi\left(t\right)\mathrm{d}t\\
&= h^2\int_{-1}^{1}u\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=uh\right]\phi\left(uh\right)\mathrm{d}u\\
&= o\left(h^2\right),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}\left(\left|D_i\right|\leq h\right)D_i^2\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right]'\right]\\
&=\int_{-h}^{h}t^2\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=t\right]\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=t\right]'\phi\left(t\right)\mathrm{d}t\\
&=h\int_{-1}^{1}\left(uh\right)^2\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=uh\right]\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=uh\right]'\phi\left(uh\right)\mathrm{d}u\\
&=\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=0\right]\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=0\right]'\phi_0h^3\int_{-1}^{1}u^2\mathrm{d}u+o\left(h^3\right)\\
&=\frac{2}{3}\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=0\right]\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i=0\right]'\phi_0h^3+o\left(h^3\right).
\end{aligned}$$

We conclude that $\mathbb{V}\left(\mathbb{E}\left[\left.Z_{N,i}\right|D_i\right]\right)=o\left(h^3\right)$.

Now consider the second term in (42). The conditional variance, using the conditional moment restriction (9), is

$$\begin{aligned}
\mathbb{V}\left(\left.Z_{N,i}\right|D_i\right) &= \mathbf{1}\left(\left|D_i\right|\leq h\right)\mathbb{V}\left(\left.D_i\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)+\mathbf{W}_i^{*\prime}\mathbf{U}_i^*\right|D_i\right)\\
&= \mathbf{1}\left(\left|D_i\right|\leq h\right)D_i^2\mathbb{V}\left(\left.\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\right|D_i\right)+\mathbf{1}\left(\left|D_i\right|\leq h\right)\mathbb{V}\left(\left.\mathbf{W}_i^{*\prime}\mathbf{U}_i^*\right|D_i\right).
\end{aligned}$$

Using an ANOVA decomposition to evaluate $\mathbb{V}\left(\left.\mathbf{W}_i^{*\prime}\mathbf{U}_i^*\right|D_i\right)$ gives

$$\begin{aligned}
\mathbb{V}\left(\left.\mathbf{W}_i^{*\prime}\mathbf{U}_i^*\right|D_i\right) &= \mathbb{E}\left[\left.\mathbb{V}\left(\left.\mathbf{W}_i^{*\prime}\mathbf{U}_i^*\right|\mathbf{X}_i\right)\right|D_i\right]+\mathbb{V}\left(\left.\mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\mathbf{U}_i^*\right|\mathbf{X}_i\right]\right|D_i\right)\\
&= \mathbb{E}\left[\left.\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\right|D_i\right]+0,
\end{aligned}$$

and hence

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left(|D_i|\leq h\right)\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i\right]\right] &= \int_{-h}^{h}\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i=t\right]\phi\left(t\right)\mathrm{d}t \\ &= h\int_{-1}^{1}\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i=uh\right]\phi\left(uh\right)\mathrm{d}u \\ &= 2\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i=0\right]\phi_0 h+o\left(h\right).\end{aligned}$$

Similarly

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\left(|D_i|\leq h\right)D_i^2\mathbb{V}\left(\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\middle|\,D_i\right)\right] &= \int_{-h}^{h}t^2\mathbb{V}\left(\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\middle|\,D_i=t\right)\phi\left(t\right)\mathrm{d}t \\ &= h\int_{-1}^{1}\left(uh\right)^2\mathbb{V}\left(\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\middle|\,D_i=uh\right)\phi\left(uh\right)\mathrm{d}u \\ &= \mathbb{V}\left(\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\middle|\,D_i=0\right)\phi_0 h^3\int_{-1}^{1}u^2\mathrm{d}u+o\left(h^3\right) \\ &= \frac{2}{3}\mathbb{V}\left(\mathbf{W}_i^{*\prime}\boldsymbol{\beta}_0\left(\mathbf{X}_i\right)\middle|\,D_i=0\right)\phi_0 h^3+o\left(h^3\right).\end{aligned}$$

Collecting terms we conclude that

$$\mathbb{V}\left(Z_{N,i}\right)=2\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i=0\right]\phi_0 h+o\left(h\right). \tag{53}$$

Applying Liapunov's Central Limit Theorem for triangular arrays, we have

$$\frac{1}{\sqrt{Nh_N}}\sum_{i=1}^{N}Z_{N,i}\xrightarrow{D}\mathcal{N}\left(0,2\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i=0\right]\phi_0\right).$$

Slutsky's Theorem and (50) above then give the following Lemma.

Lemma A.2 *Suppose that (i) $\left(F_0,\boldsymbol{\delta}_0,\boldsymbol{\beta}_0\left(\cdot\right)\right)$ satisfies (9), (ii) $\Sigma\left(\mathbf{x}\right)$ is positive definite for all $\mathbf{x}\in\mathbb{X}^T$, (iii) $T=p$, and (iv) Assumptions 1.2 to 2.5 hold, then $\widehat{\boldsymbol{\delta}}\xrightarrow{p}\boldsymbol{\delta}_0$ with the normal limiting distribution*

$$\sqrt{Nh_N}\left(\widehat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0\right)\xrightarrow{D}\mathcal{N}\left(0,\frac{\Lambda_0}{2\phi_0}\right),$$

where

$$\Lambda_0=\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\middle|\,D_i=0\right]^{-1}\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{X}^*\Sigma\left(\mathbf{X}\right)\mathbf{X}^{*\prime}\mathbf{W}_i^*\middle|\,D_i=0\right]\mathbb{E}\left[\mathbf{W}_i^{*\prime}\mathbf{W}_i^*\middle|\,D_i=0\right]^{-1}.$$

Large sample properties of $\widehat{\boldsymbol{\beta}}$: The following lemma characterizes the probability limit of $\widehat{\Xi}_N$.

Lemma A.3 *If $\left(F_0,\boldsymbol{\delta}_0,\boldsymbol{\beta}_0\left(\cdot\right)\right)$ satisfies (9) and Assumptions 1.2 to 2.5 hold we have $\widehat{\Xi}_N\xrightarrow{p}\Xi_0$, where*

$$\begin{aligned}\Xi_0 &= \lim_{h_N\downarrow 0}\mathbb{E}\left[\mathbf{1}\left(|D_i|>h_N\right)\mathbf{X}_i^{-1}\mathbf{W}_i\right]. \\ &\equiv \lim_{N\to\infty}\Xi_N.\end{aligned}$$

Proof. See the Supplemental Web Appendix. ■

Lemmas A.1, A.2, and A.3 as well as the decomposition (37) then give Theorem 2.1.

MSE-optimal bandwidth sequence: The MSE-optimal bandwidth sequence given in equation (29) of the main text may be derived as follows. Let a be a $p \times 1$ vector of constants. Using (41), (52) and Lemma A.3 yields a leading asymptotic bias term for $a'\boldsymbol{\beta}_0$ of $2a'\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)\phi_0 h$. Using the asymptotic variance expression given in the statement of Theorem 2.1 we get an asymptotic MSE for $a'\widehat{\boldsymbol{\beta}}$ of

$$4a'\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^S\right)' a\phi_0^2 h^2 + \frac{a'\left(2\Upsilon_0\phi_0 + \frac{\Xi_0\Lambda_0\Xi_0'}{2\phi_0}\right)a}{Nh}.$$

Minimizing this object with respect to h gives the result in the main text.

References

[1] Abrevaya, Jason. (2000). "Rank estimation of a generalized fixed-effects regression model," *Journal of Econometrics* 95 (1): 1 - 23.

[2] Altonji, Joseph G. and Rosa L. Matzkin (2005). "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica* 73 (4): 1053 - 1102.

[3] Angrist, Joshua D. and Alan B. Krueger. (1999). "Empirical strategies in labor economics," *Handbook of Labor Economics* 3 (1): 1277 - 1366 (O. C. Ashenfelter & D. Card, Eds.). Amsterdam: North-Holland.

[4] Andrews, Donald W. K. and Marcia M. A. Schafgans. (1998). "Semiparametric estimation of the intercept of a sample selection model," *Review of Economic Studies* 65 (3): 497 - 517.

[5] Arellano, Manuel and Stephanne Bonhomme. (2009). "Identifying distributional characteristics in random coefficients panel data models," *Mimeo.*

[6] Arellano, Manuel and Raquel Carrasco. (2003). "Binary choice panel data models with predetermined variables," *Journal of Econometrics* 115 (1): 125 - 157.

[7] Arellano, Manuel and Bo Honoré. (2001). "Panel data models: some recent developments," *Handbook of Econometrics* 5: 3229 - 3298 (J. Heckman & E. Leamer, Eds.). Amsterdam: North-Holland.

[8] Behrman, Jere R. and Anil B. Deolalikar. (1987)."Will developing country nutrition improve with income? A case study for rural south India," *Journal of Political Economy* 95 (3): 492 - 507.

[9] Bester, C. Alan and Christian Hansen. (2009). "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model," *Journal of Business and Economic Statistics* 27 (2): 235 - 250.

[10] Blundell, Richard W. and James L. Powell. (2003). "Endogeneity in nonparametric and semiparametric regression models," *Advances in Economics and Econometrics: Theory and Applications II*: 312 - 357. (M. Dewatripont, L.P. Hansen, S. J. Turnovsky, Eds.). Cambridge: Cambridge University Press.

[11] Bonhomme, Stephane. (2010). "Functional differencing," *Mimeo.*

[12] Bouis, Howarth E. (1994). "The effect of income on demand for food in poor countries: are our food consumption databases giving us reliable estimates? *Journal of Development Economics* 44(1): 199-226.

[13] Bouis, Howarth E. and Lawrence J. Haddad. (1992)."Are estimates of calorie-income elasticities too high? A recalibration of the plausible range," *Journal of Development Economics* 39 (2): 333 - 364.

[14] Browning, Martin and Jesus Carro. (2007). "Heterogeneity and microeconometrics modelling," *Advances in Economics and Econometrics: Theory and Applications III*: 47 - 74. (R. Blundell, W. Newey & T. Persson, Eds.). Cambridge: Cambridge University Press.

[15] Card, David. (1996). "The effect of unions on the structure of wages: a longitudinal analysis," *Econometrica* 64 (4): 957 - 979.

[16] Chamberlain, Gary. (1980). "Analysis of covariance with qualitative data," *Review of Economic Studies* 47 (1): 225 - 238.

[17] Chamberlain, Gary. (1982). "Multivariate regression models for panel data," *Journal of Econometrics* 18 (1): 5 - 46.

[18] Chamberlain, Gary. (1984). "Panel data," *Handbook of Econometrics 2:* 1247 - 1318 (Z. Griliches & M.D. Intriligator, Eds.). Amsterdam: North-Holland.

[19] Chamberlain, Gary. (1986). "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics* 32 (2): 189 - 218.

[20] Chamberlain, Gary. (1992). "Efficiency bounds for semiparametric regression," *Econometrica* 60 (3): 567 - 596.

[21] Chamberlain, Gary. (2010). "Binary response models for panel data: identification and information," *Econometrica* 78 (1): 159 - 168.

[22] Chernozhukov, Victor, Iván Fernández-Val, Jinyong Hahn and Whitney Newey. (2008). "Identification and estimation of marginal effects in nonlinear panel data models," *CEMMAP Working Paper* CWP25/08.

[23] Dasgupta, Partha. (1993). *An Inquiry into Well-Being and Destitution.* Oxford: Oxford University Press.

[24] Engle, Robert F., C. W. J. Granger, John Rice, and Andrew Weiss. (1986). "Semiparametric estimates of the relation between weather and electricity sales," *Journal of the American Statistical Association* 81 (394): 310 - 320.

[25] Food and Agricultural Organization (FAO). (2001)."Human energy requirements: report of a joint FAO/WHO/UNU expert consultation," *FAO Food and Nutrition Technical Report Series 1.*

[26] Food and Agricultural Organization (FAO). (2006). *The State of Food Insecurity in the World 2006.* Rome: Food and Agricultural Organization.

[27] Graham, Bryan S., Guido W. Imbens, Geert Ridder. (2009). "Complementarity and aggregate implications of assortative matching: a nonparametric analysis," *Mimeo.*

[28] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. (1998). "Characterizing selection bias using experimental data," *Econometrica* 66 (5): 1017 - 1098.

[29] Heckman, James J. (1990). "Varieties of selection bias," *American Economic Review* 80 (2): 313 - 18.

[30] Hoderlein, Stefan and Halbert White. (2009). "Nonparametric identification in nonseparable panel data models with generalized fixed effects," *Mimeo.*

[31] Honoré, Bo E. (1992). "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects," *Econometrica* 60 (3): 533 - 565.

[32] Honoré, Bo E. and Ekaterini Kyriazidou. (1997). "Panel data discrete choice models with lagged dependent variables," *Econometrica* 68 (4): 839 - 874.

[33] Horowitz, Joel L. (1992). "A smoothed maximum score estimator for the binary response model," *Econometrica* 60 (3): 505 - 531.

[34] Imbens, Guido W. (2007). "Nonadditive models with endogenous regressors," *Advances in Economics and Econometrics: Theory and Applications III*: 17 - 46. (R. Blundell, W. Newey & T. Persson, Eds.). Cambridge: Cambridge University Press.

[35] Imbens, Guido W. and Whitney K. Newey. (2009). "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica* 77 (5): 1481 - 1512.

[36] International Food Policy Research Institute (IFPRI). (2005). *Nicaraguan RPS evaluation data (2000-02): overview and description of data files (April 2005 Release).* Washington D.C.: International Food Policy Research Institute.

[37] Khan, Shakeeb and Elie Tamer. (2010). "Irregular identification, support conditions, and inverse weight estimation," *Econometrica* 78 (6): 2021 – 2042.

[38] Kyriazidou, Ekaterini. (1997). “Estimation of a panel sample selection model,” *Econometrica* 65 (6): 1335 - 1364.

[39] Manski, Charles F. (1987). “Semiparametric analysis of random effects linear models from binary panel data,” *Econometrica* 55 (2): 357 - 362.

[40] Mundlak, Yair. (1961). “Empirical production function free of management bias,” *Journal of Farm Economics* 43 (1): 44 - 56.

[41] Mundlak, Yair. (1978a). “On the pooling of time series and cross section data,” *Econometrica* 46 (1): 69 - 85.

[42] Mundlak, Yair. (1978b). “Models with variable coefficients: integration and extension,” *Annales de l'Insee* 30-31: 483 - 510.

[43] Newey, Whitney K. (1994a). “The asymptotic variance of semiparametric estimators,” *Econometrica* 62 (6): 1349 - 1382.

[44] Newey, Whitney K. (1994b). “Kernel estimation of partial means and a general variance estimator,” *Econometric Theory* 10 (2): 233 - 253.

[45] Pagan, Adrian and Aman Ullah. (1999). *Nonparametric Econometrics.* Cambridge: Cambridge University Press.

[46] Robinson, P. M. (1988). “Root-N-consistent semiparametric regression,” *Econometrica* 56 (4): 931 - 954.

[47] Serfling, Robert J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: John Wiley & Sons, Inc.

[48] Smith, Lisa C. and Ali Subandoro. (2007). *Measuring food security using household expenditure surveys.* Washington D.C.: International Food Policy Research Institute.

[49] Strauss, John and Duncan Thomas. (1990). “The shape of the calorie-expenditure curve,” *Yale University Economic Growth Center Discussion Paper No. 595.*

[50] Strauss, John and Duncan Thomas. (1995). “Human resources: empirical modeling of household and family decisions,” *Handbook of Development Economics* 3 (1): 1883 - 2023. (J. Behrman & T.N. Srinivasan). Amsterdam: North-Holland.

[51] Subramanian, Shankar and Angus Deaton. (1996). “The demand for food and calories,” *Journal of Political Economy* 104 (1): 133 - 162.

[52] Wolfe, Barbara L. and Jere R. Behrman. (1983). “Is income overrated in determining adequate nutrition?” *Economic Development and Cultural Change* 31 (3): 525 - 549.

[53] Wooldridge, Jeffrey M. (2005a). "Unobserved heterogeneity and estimation of average partial effects," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 27 - 55 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

[54] Wooldridge, Jeffrey M. (2005b). "Fixed-effects and related estimators for correlated-random coefficient and treatment-effect panel data models," *Review of Economics and Statistics* 87 (2): 385 - 390.

[55] World Bank. (2003). *Nicaragua Poverty Assessment: Raising Welfare and Reducing Vulnerability.* Washington D.C.: The World Bank.