Average Treatment Effects for Stayers with Correlated Random Coefficient Models of Panel Data

Valentin Verdier* Andrew Castro†

March 12, 2019

Abstract

Correlated random coefficient (CRC) models provide a useful framework for estimating average treatment effects (ATE) with panel data by accommodating heterogeneous treatment effects and flexible patterns of selection. In their simplest form, they lead to the well-known difference-in-differences estimator. Estimators for CRC models yield estimates of ATE for "movers", i.e. cross-sectional units whose treatment status changed over time, while ATE for "stayers", i.e. cross-sectional units who retained the same treatment status over time, are not identified. In some cases ATE for stayers are of first order importance, for instance when predicting the effect of expanding access to treatment. We study how to identify ATE for stayers in a CRC framework. We show that imposing particular restrictions on selection into treatment leads to the identification of ATE for stayers by a linear extrapolation using quantities identified by the CRC model. We show that this extrapolation is implicitly used by methods found in existing empirical work that rely on panel data models where a single term of unobserved heterogeneity determines both treated and untreated outcomes. We propose a test for the validity of the extrapolation as well as a more robust extrapolation that relaxes the restrictions imposed on selection into treatment. We also propose an easy to implement two-step estimation procedure. We use our results to estimate the returns to agricultural technology adoption among maize farmers in Kenya.

Keywords: Panel data, Correlated Random Coefficient Models, Difference-in-differences, Agricultural technology adoption. JEL codes: C23

Online appendix: `https://goo.gl/Swk3h7`

*Corresponding author. Department of Economics, University of North Carolina - Chapel Hill.
†Department of Economics, Michigan State University.

1 Introduction

When outcomes and treatment status are observed repeatedly over time, many methods frequently used in empirical work rely on comparisons both across time and across cross-sectional units to obtain estimates of treatment effects that are robust to patterns of selection into treatment that could lead to biases with more naive estimators. With selection on cross-sectional unobserved heterogeneity, first difference or fixed effects estimation yield valid estimates when this heterogeneity is constant over time and treatment effect is homogenous. If selection into treatment varies across time, for instance with more treated observations in later time periods, estimated treatment effects could be biased by the presence of aggregate shocks to outcomes, but estimators that include time fixed effects are valid when these aggregate shocks are common to all cross-sectional units.

The use of these methods in empirical work is too widespread to provide a comprehensive list of studies where they are implemented, but select examples are found in Freeman (1984), Jakubson (1991), Card (1996), and Lemieux (1998) who estimate the effect of union membership on wages, Ashenfelter and Card (1985) and Card and Sullivan (1988) who estimate the effect of job training programs on earnings and employment probabilities, Arcidiacono et al. (2008) who estimate returns to completing a MBA, Kowaleski-Jones and Duncan (2002), Behrman and Hoddinott (2005), and Alderman (2007) who estimate the effect of different social aid programs on nutrition, Rouse (1998) who estimates the effect of attending private school on student achievement, Hanushek et al. (1998) who estimate the effect of enrolling in a special education program on student achievement, Clotfelter et al. (2010) who estimate the effect of teacher certification on student achievement, Suri (2011) who estimates the effect of technology adoption on agricultural yields.

In many empirical applications, treatment effect is likely to be heterogenous across cross-sectional units.[1] This heterogeneity leads to biases in the fixed effects or first difference

[1]See Browning and Carro (2007) for empirical evidence on panel data models with more than additive unobserved heterogeneity.

estimators discussed above if they are perceived as estimating average treatment effects.[2] In addition estimating features of the heterogeneity in treatment effects might be of first-order interest in many empirical studies.

Correlated random coefficient (CRC) models take treatment effects to be specific to each cross-sectional unit. Noisy measures of treatment effects can be obtained for cross-sectional units whose treatment status has changed over time ("movers") by comparing the changes over time in their outcomes following a change in treatment status with the change over time in the outcome of cross-sectional units whose treatment status has remained constant over time ("stayers"). These measures can be used to obtain consistent estimators of conditional average treatment effects when large groups of movers are observed.[3]

This approach does not yield estimated average treatment effects for stayers because these effects are not identified in a model where treatment effects are heterogenous and selection into treatment can be determined in an unrestricted way by the unobserved heterogeneity that determines outcome.[4] In some empirical applications, average treatment effects for stayers might be of first-order interest. This is the case for instance when a policy maker is evaluating interventions aimed at increasing enrollment in a program, i.e. aimed at treating cross-sectional units who have so far never been treated.

Here we show that imposing particular restrictions on selection into treatment leads to the identification of ATE for stayers by a linear extrapolation using quantities identified by the CRC model.

Lemieux (1998) and Suri (2011) have used panel data models where a single term of unobserved heterogeneity determines both treated and untreated outcomes to identify average treatment effects among stayers. We establish a link between these methods and the

[2]See Chamberlain (1982), Wooldridge (2005), Wooldridge (2010), and de Chaisemartin and D'Haultfoeuille (2018b).

[3]This is the well-known difference-in-differences estimator. This method can also be applied to repeated cross-section data aggregated at the level of control and treatment groups rather than individual panel data, see e.g. de Chaisemartin and D'Haultfoeuille (2018a) for a review. Here we consider the case where individual panel data is available.

[4]This is discussed for instance in Chamberlain (1982) and Chernozhukov et al. (2013).

methods corresponding to CRC models by showing that the methods of Lemieux (1998) and Suri (2011) are equivalent to the linear extrapolation mentioned above.

This extrapolation to ATE for stayers generally requires that factors determining selection into treatment other than treatment effect be independent of baseline heterogeneity and treatment effect. In some empirical applications, this assumption might be implausible. We propose a test for the validity of the extrapolation to ATE for stayers which can be implemented when three or more time periods are observed. We also propose a generalized extrapolation which takes into account factors determining selection into treatment and originating from observable sources which may be correlated with baseline heterogeneity or treatment effect, so that this generalized extrapolation is valid under more general patterns of selection into treatment. In practice, we show that estimation can be based on a two-step procedure composed of a high-dimensional linear regression and an instrumental variable regression, so that implementation is straightforward.

We apply our results to revisit the empirical study of Suri (2011) on estimating the returns to using an agricultural technology, hybrid seeds, among maize farmers in Kenya.

2 Correlated Random Coefficient Model and Extrapolation to Stayers

With panel data where cross-sectional observations are indexed by $i = 1, ..., n$ and time periods are indexed by $t = 1, ..., T$, the correlated random coefficient (CRC) model for the effect of treatment $x_{it} \in \{0, 1\}$ on outcome y_{it} can be written:

$$y_{it} = a_i + b_i x_{it} + f_t + u_{it}, \qquad E(u_{it}|X) = 0 \tag{2.1}$$

where X collects all values of treatment status across time and cross-sectional observations, i.e. $X = \{x_{it}\}_{i=1,...,n,t=1,...,T}$. The unobserved term a_i captures baseline heterogeneity which affects a cross-sectional observation's outcome both when untreated or treated, while the unobserved term b_i is the treatment effect for cross-sectional observation i. Unobserved time

varying shocks that are common to all cross-sectional observations are captured by f_t, and u_{it} captures all other unobserved factors.

In the empirical example which we will use as an illustration throughout the paper, outcome y_{it} is the productivity of farmer i in year t and x_{it} indicates whether this farmer uses hybrid seeds. The farmer's baseline productivity is captured by a_i, the farmer's returns to using hybrid seeds is given by b_i, f_t are aggregate shocks to productivity common to all farmers, and u_{it} are idiosyncratic transitory shocks to productivity.

Selection on unobservables is captured in this model by not imposing any restriction on the relationship between treatment status (x_{it}) and the heterogeneity in outcomes across cross-sectional observations and time (a_i, b_i, f_t).

The assumption $E(u_{it}|X) = 0$ is referred to as an assumption of strict exogeneity, which in particular rules out feedback mechanisms from past values of the outcome to current treatment status.

For simplicity we do not consider including additional control covariates in this section, but including additional covariates has consequences for modeling and estimation which we discuss below and in the online appendix. We consider unbalanced panels with data missing at random in the empirical section below and in the online appendix. For simplicity we consider the case where there is only one treatment here, but our results could be extended to multiple treatments, which we discuss briefly in the conclusion section of the paper.

The variable x_{it} is taken to be binary here, but all of our results extend directly to the case where x_{it} is a discrete variable. If x_{it} had a continuous distribution function and were not perfectly persistent over time, the probability that a cross-sectional observation is a stayer would be zero. This is the case considered by Chamberlain (1992) which leads to requiring that three or more time periods are available, so that time effects f_t can be identified by comparisons among movers. Graham and Powell (2012) show that a trimming estimator can be used to estimate time effects f_t from two time periods only, although the resulting convergence rate is slower than the parametric rate. Arellano and Bonhomme (2012) study

how to recover information about the distribution of heterogenous effects among movers when additional restrictions can be imposed on the second moments of the transitory shocks u_{it}. Fernández-Val and Lee (2013) study a similar model in a long panel setting (i.e. where the number of time periods T is relatively large and serial dependence is weak), while here we concentrate on short panels where T is relatively small compared to the number of cross-sectional observations n.[5]

We begin this section by rewriting the CRC model (2.1) to establish an exhaustive list of the quantities that are identified under this model. For simplicity we will consider the case where only two time periods are observed, i.e. $T = 2$, through most of the paper. The online appendix considers the case where a general number of time periods is observed.

In order to shorten notation, we will write throughout the paper $E(w_i|x_1, ..., x_T)$ to denote $E(w_i|x_{i1} = x_1, ..., x_{iT} = x_T)$ for any random variable w_i and values $x_t \in \{0, 1\}\ \forall\, t$.

With two time periods, i.e. $T = 2$, the CRC model (2.1) yields identification of time effects f_t, up to a normalization such as $f_1 = 0$, from changes in outcomes of stayers:

$$\Delta f_2 = E(\Delta y_{i2}|0, 0) = E(\Delta y_{i2}|1, 1) \tag{2.2}$$

where Δ is the first-differencing operator.

ATE for movers are then identified by a difference-in-differences comparison:

$$E(b_i|0, 1) = E(\Delta y_{i2}|0, 1) - \Delta f_2, \qquad E(b_i|1, 0) = -(E(\Delta y_{i2}|1, 0) - \Delta f_2) \tag{2.3}$$

Average baseline heterogeneity is identified by average outcomes for untreated stayers and movers, average total heterogeneity is identified by average outcomes for treated stayers:

$$E(a_i|0, 0) = E(y_{it}|0, 0) - f_t, \qquad E(a_i + b_i|1, 1) = E(y_{it}|1, 1) - f_t \tag{2.4}$$

$$E(a_i|0, 1) = E(y_{i1}|0, 1) - f_1, \qquad E(a_i|1, 0) = E(y_{i2}|1, 0) - f_2 \tag{2.5}$$

When $T = 2$ and under cross-sectional independence, the CRC model (2.1) is equivalent to equations (2.2)-(2.5), from which we see that the CRC model imposes no restriction on

[5]We study the CRC model throughout this paper. Reviews of alternative non-linear models of panel data can be found in Arellano and Bonhomme (2011) and Ghanem (2017).

ATE for stayers, i.e. ATE for stayers are not identified with the CRC model.[6] The online appendix discusses the general case where $T \geq 2$. When $T > 2$, time effects f_t are identified by observations on both stayers and movers, but ATE remain non-identified for stayers.

2.1 Extrapolation to Stayers

Lemieux (1998) and Suri (2011) both consider datasets with two time periods ($T = 2$) and estimate a model for outcomes given by:[7]

$$y_{it} = b_i(x_{it} + \alpha_1) + \tilde{f}_t + \tilde{u}_{it}, \qquad E(\tilde{u}_{it}|x_{i1}, x_{i2}) = 0 \tag{2.6}$$

where $\tilde{f}_t$ are time effects and b_i is a single term of unobserved heterogeneity that affects both treated outcome (with a coefficient of $1 + \alpha_1$) and untreated outcome (with a coefficient of α_1).

Under the correlated random coefficient model (2.1) discussed above, the model (2.6) holds if an additional restriction holds:

$$a_i = \alpha_0 + \alpha_1 b_i + \epsilon_i, \qquad E(\epsilon_i|x_{i1}, ..., x_{iT}) = 0 \tag{2.7}$$

which we will call an extrapolation identifying assumption. Combining (2.1) and (2.7) yields the model (2.6) by defining $\tilde{f}_t = f_t + \alpha_0$ and $\tilde{u}_{it} = u_{it} + \epsilon_i$.

Before showing how assumption (2.7) leads to the identification of average treatment effects among stayers under the CRC model (2.1), we discuss its interpretation.

2.1.1 Interpretation of the Extrapolation Identifying Assumption

In order to interpret the extrapolation identifying assumption (2.7), we consider two sufficient conditions for the assumption (2.7) to hold. Using the law of iterated expectations,

[6] See also Chernozhukov et al. (2013), where partial identification with similar models when the outcome variable has bounded support is discussed.

[7] This model with additional covariates, which we are ignoring in this section for simplicity, is given by equations (7) and (8) in Lemieux (1998) and equations (20) and (26) in Suri (2011).

we can show that the extrapolation identifying assumption (2.7) holds if:

$$a_i = \alpha_0 + \alpha_1 b_i + \epsilon_i, \qquad E(\epsilon_i | b_i) = 0 \tag{2.8}$$

$$x_{it} = g(b_i, c_{it}) \; \forall \, t, \qquad \{c_{it}\}_{t=1,\ldots,T} \perp \{a_i, b_i\} \tag{2.9}$$

where g is a deterministic function and $\{c_{it}\}_{\forall t}$ are random variables.

The condition (2.8) captures the statistical dependence between baseline heterogeneity a_i and treatment effect b_i, but imposes linearity in the conditional mean of baseline heterogeneity a_i conditional on treatment effect b_i. While restrictive, this assumption is similar to assumptions of linear common factor models found in the literature on generalized Roy models with cross-sectional data, see e.g. Heckman and Vytlacil (2007). In a short panel setting, i.e. treating T as fixed, and for point identification to be preserved, it is likely that this assumption could only be relaxed at the expense of more stringent restrictions imposed on the conditional distribution of the error term u_{it} in the CRC model (2.1) than mean independence.

The condition (2.9) captures the possible dependence of treatment status x_{it} on treatment effect b_i (endogenous selection), but imposes that factors of selection into treatment other than treatment effect b_i, denoted as c_{it}, be independent of baseline heterogeneity and treatment effect. For illustration purposes, the function g can be taken to be a threshold crossing function, so that $x_{it} = 1[b_i \geq c_{it}]$ where $1[.]$ is the indicator function. In this case, observation i receives treatment in time period t if treatment effect b_i exceeds the cost of treatment given by c_{it}. With this representation, the condition (2.9) assumes that the cost of treatment c_{it} is independent of both baseline heterogeneity a_i and treatment effect b_i.[8]

In the context of our empirical example on maize production in Kenya, condition (2.9) has an intuitive interpretation as requiring that the cost of using hybrid seeds faced by farmers be statistically independent of their baseline productivity a_i and of their returns to

[8]Condition (2.9) is also similar to the restrictions on selection into treatment imposed in generalized Roy models with cross-sectional data, but it is weaker than what is required when only cross-sectional data is available. Indeed the extrapolation identifying assumption (2.7) obtained from conditions (2.8) and (2.9), together with the CRC model (2.1), contain no identifying power for ATE when $T = 1$ in settings where only y_{it} and x_{it} are observed, i.e. where additional exogenous instrumental variables are not available.

using hybrid seeds b_i. We discuss the plausibility of this restriction and the consequences of its violation in Section 2.2.

2.1.2 Identification of Average Treatment Effects for Stayers

When only two time periods are observed ($T = 2$), the extrapolation identifying assumption (2.7) can be rewritten as:

$$\alpha_1 = \frac{E(a_i|0,1) - E(a_i|1,0)}{E(b_i|0,1) - E(b_i|1,0)}, \qquad \alpha_0 = E(a_i|0,1) - \alpha_1 E(b_i|0,1) \tag{2.10}$$

$$E(b_i|0,0) = \frac{E(a_i|0,0) - \alpha_0}{\alpha_1}, \qquad E(b_i|1,1) = \frac{E(a_i|1,1) - \alpha_0}{1+\alpha_1} \tag{2.11}$$

if all quantities above are well-defined, i.e. if $E(b_i|0,1) \neq E(b_i|1,0)$ and $\alpha_1 \notin \{0,-1\}$.

Since all of the right-hand side quantities in equations (2.10) and (2.11) are identified under the CRC model, we see that assumption (2.7) yields identification of ATE for stayers by an extrapolation from ATE and average baseline heterogeneity for movers to ATE for stayers. This extrapolation takes a simple form: an extrapolation line is drawn through points $(E(a_i|1,0), E(b_i|1,0))$ and $(E(a_i|0,1), E(b_i|0,1))$. Average treatment effects for untreated stayers are identified by interpolating the vertical line given by $a = E(a_i|0,0)$ and the extrapolation line. Average treatment effects for treated stayers are identified by interpolating the -45° line given by $a + b = E(a_i|1,1) + E(b_i|1,1)$ and the extrapolation line. This is represented graphically in Figure 1. If the condition $E(b_i|1,0) \neq E(b_i|0,1)$ fails, the extrapolation line is not identified. If the conditions $\alpha_1 \neq 0$ and $\alpha_1 \neq -1$ fail, ATE are not identified for untreated and treated stayers, respectively.

Lemieux (1998) and Suri (2011) both consider estimation of the model given by (2.6). Lemieux (1998) uses generalized method of moments estimation to estimate the parameters of the model and ATE for all subpopulations of interest, Suri (2011) relies on minimum distance estimation.

The next proposition shows that both estimation methods can also be represented as the linear extrapolation from ATE among movers to ATE among stayers discussed above and

depicted in Figure 1.[9]

Proposition 1. *Assume that the CRC model given by (2.1) holds and that observations are cross-sectionally independently and identically distributed. Define the pseudo-true values:*

$$\alpha_1^\star = \frac{E(a_i|1,0) - E(a_i|0,1)}{E(b_i|1,0) - E(b_i|0,1)}, \qquad \alpha_0^\star = E(a_i|0,1) - \alpha_1^\star E(a_i|0,1) \tag{2.12}$$

and define the pseudo-true values for ATE of untreated stayers ($ATE_{00} = E(b_i|0,0)$), treated stayers ($ATE_{11} = E(b_i|1,1)$), and ATE of the entire population ($ATE = E(b_i)$):

$$ATE_{00}^\star = \frac{E(a_i|0,0) - \alpha_0^\star}{\alpha_1^\star}, \;\; ATE_{11}^\star = \frac{E(a_i + b_i|1,1) - \alpha_0^\star}{1 + \alpha_1^\star} \tag{2.13}$$

$$ATE^\star = \pi_{00} ATE_{00}^\star + \pi_{11} ATE_{11}^\star + \pi_{01} E(b_i|0,1) + \pi_{10} E(b_i|1,0) \tag{2.14}$$

where $\pi_{x_1 x_2} = P(x_{i1} = x_1, x_{i2} = x_2) \; \forall \, x_1, x_2 \in \{0,1\}$.

As long as they are well-defined, these pseudo-true values are solutions to the moment conditions used in Lemieux (1998) and to the link between reduced form and structural parameters used in Suri (2011).

The appendix provides details on the estimation methods of Lemieux (1998) and Suri (2011) and the proof of Proposition 1.

Proposition 1 provides a mechanical representation of the methods of Lemieux (1998) and Suri (2011) in terms of quantities obtained by well-known difference-in-differences comparisons. Additionally it shows that, compared to the CRC model (2.1), the only additional identifying power of models with a single term of unobserved heterogeneity is found in assuming that ATE of stayers can be obtained by a linear extrapolation from quantities identified by the CRC model.

The discussion at the beginning of this section also shows that the validity of this extrap-

[9] In cross-sectional data settings, there are results on extrapolating from subpopulations for which ATE are robustly identified to wider subpopulations that have a similar flavor to the extrapolation discussed here. Angrist and Fernández-Val (2013), Brinch et al. (2017), Kline and Walters (2018), and Mogstad et al. (2018) discuss extrapolation from local ATE obtained by instrumental variable regression. Angrist and Rokkanen (2015), Bertanha (2017), Bertanha and Imbens (2014), Cattaneo et al. (2016), Dong and Lewbel (2015), Rokkanen (2015) discuss extrapolation from local ATE obtained by regression discontinuity design. Wüthrich (2018) establishes a relationship between instrumental variable quantile regression estimation and local quantile treatment effect estimation.

olation is equivalent to the extrapolation identifying assumption (2.7) holding, and that this extrapolation identifying assumption holds under a linearity restriction and if the factors of selection into treatment other than treatment effect are independent of baseline heterogeneity and treatment effect.

With this interpretation of the extrapolation identifying assumption (2.7), researchers can evaluate the plausibility of the extrapolation to ATE for stayers in particular empirical applications. In applications where the assumption (2.7) is implausible, researchers may wish to test the validity of the extrapolation to ATE discussed above, and to use an extrapolation to ATE for stayers which relies on more flexible restrictions on selection into treatment. Before discussing these two extensions, we introduce a simple two-step estimation procedure for the model composed of the CRC assumption (2.1) and the extrapolation identifying assumption (2.7).

2.1.3 Two-Step Estimation

As an alternative to the estimation methods of Lemieux (1998) and Suri (2011), we discuss here a two-step estimation method for the model composed of the CRC assumption (2.1) and the extrapolation identifying assumption (2.7). The first step of the procedure consists of a high-dimensional regression while the second step consists of an instrumental variable regression, so that estimation can be performed easily with standard software. The advantage of this two-step estimation procedure, in addition to ease of implementation, is that it can be directly extended to accommodate additional control covariates in the CRC model, that it leads to natural testing procedures for the validity of the extrapolation discussed above, and that it can easily be extended to estimate ATE among stayers using a generalized extrapolation which allows for more flexible dependence between the factors that determine selection into treatment and the unobserved heterogeneity that determines outcomes. All three of these extensions are discussed below.

The first step of our estimation procedure, similarly as in Chamberlain (1992), consists of a regression of outcomes y_{it} on a set of indicator variables for each cross-sectional observation,

the interaction of these indicator variables with treatment status, and a set of indicator variables for each time period.

This procedure yields estimates of time effects f_t, $\hat{f}_t$. The procedure also yields noisy estimates of baseline heterogeneity a_i and treatment effect b_i, $\hat{a}_i$ and $\hat{b}_i$, for each cross-sectional observation which is a mover. For untreated stayers, only estimates of baseline heterogeneity a_i, $\hat{a}_i$, are obtained, while only estimates of total heterogeneity $a_i + b_i$, $\hat{a}_i + \hat{b}_i$, are obtained for treated stayers.

As before, we consider for simplicity the special case where two time periods are observed, i.e. $T = 2$. The case with a general number of time periods is considered in the online appendix.

With only two time periods, this first step estimation procedure takes a particularly simple form. The estimated time effects obtained with the normalization $\hat{f}_1 = 0$ are given by:

$$\hat{f}_2 = \frac{\sum_{i \notin M_n} \Delta y_{i2}}{n - |M_n|}$$

where M_n is the set of all cross-sectional observations that are movers, i.e. $M_n = \{i = 1, ..., n : x_{i1} \neq x_{i2}\}$ and $|.|$ denotes the cardinality of a set.

We also have:

$$\hat{a}_i = \frac{\sum_{t=1,2}(1 - x_{it})(y_{it} - \hat{f}_t)}{\sum_{t=1,2}(1 - x_{it})}, \qquad \hat{a}_i + \hat{b}_i = \frac{\sum_{t=1,2} x_{it}(y_{it} - \hat{f}_t)}{\sum_{t=1,2} x_{it}}, \qquad \hat{b}_i = \hat{a}_i + \hat{b}_i - \hat{a}_i$$

where, as discussed above, all three of these quantities are well-defined for movers only. For untreated stayers, only $\hat{a}_i$ is well-defined, while only $\hat{a}_i + \hat{b}_i$ is well-defined for treated stayers.

The first result in this section shows conditions under which time effects are estimated precisely and shows that the noise in the estimates of heterogeneity can be decomposed into two parts, one which vanishes asymptotically and the other which does not depend on sample size.

For simplicity we assume that observations are identically and independently distributed (i.i.d.) at the level of cross-sectional units. This assumption is relaxed to independence

across cross-sectional observations in the appendix and can easily be relaxed to accommodate limited forms of cross-sectional dependence such as cluster dependence as in Section 2.3 below.

Assumption 1. *Observations* $\{x_{i1}, y_{i1}, x_{i2}, y_{i2}, a_i, b_i\}_{i=1,...,n}$ *are i.i.d. across* i.

The second assumption imposes that all variables in the model have finitely bounded higher moments, that there is a positive probability of being a stayer, and that the error term u_{it} is not degenerate.

Assumption 2. *Define* $\pi_S = P(x_{i1} = x_{i2})$ *and* $\sigma^2_{\Delta u,S} = Var(\Delta u_{i2}|x_{i1} = x_{i2})$.

a) The support of a_i, b_i, u_{it} *is compact.*

b) $\pi_S > 0$.

c) $\sigma^2_{\Delta u,S} > 0$.

Assumption 2.a is imposed in this form for simplicity and could easily be relaxed to impose bounded higher moments only. Assumption 2.b is natural here since we are interested in cases where stayers are observed in the data. Assumption 2.c is a regularity condition which imposes that the error term in the CRC model (2.1) has variability, so that the model would not fit the data perfectly without this error term.

Under these assumptions, Proposition 2 establishes the asymptotic properties of our first-step estimates.

Proposition 2. *Under the CRC model (2.1) and Assumptions 1 and 2, as* $n \to \infty$*:*

$$\sqrt{n}(\hat{f}_2 - f_2) \xrightarrow{d} N(0, \frac{\sigma^2_{\Delta u,S}}{\pi_S}) \tag{2.15}$$

and wherever $\hat{a}_i$ *and* $\hat{a}_i + \hat{b}_i$ *are well-defined we can write:*

$$\hat{a}_i = a_i + \frac{\sum_{t=1,2}(1 - x_{it})u_{it}}{\sum_{t=1,2}(1 - x_{it})} + \zeta_{a,i,n}, \quad \hat{a}_i + \hat{b}_i = a_i + b_i + \frac{\sum_{t=1,2} x_{it}u_{it}}{\sum_{t=1,2} x_{it}} + \zeta_{a+b,i,n} \tag{2.16}$$

where $max_{i=1,...,n:x_{i1}=0 \, or \, x_{i2}=0}|\zeta_{a,i,n}| = O_p(\frac{1}{\sqrt{n}})$ *and* $max_{i=1,...,n:x_{i1}=1 \, or \, x_{i2}=1}|\zeta_{a+b,i,n}| = O_p(\frac{1}{\sqrt{n}})$.

Proposition 2 shows that the noise in the estimates of a_i and $a+b_i$ obtained from the first step of our estimation procedure is approximated by a noise term of mean zero conditional

on treatment status history, since under the CRC model (2.1) we have:

$$E(\frac{\sum_{t=1,2}(1-x_{it})u_{it}}{\sum_{t=1,2}(1-x_{it})}|x_{i1},x_{i2})=0, \qquad E(\frac{\sum_{t=1,2}x_{it}u_{it}}{\sum_{t=1,2}x_{it}}|x_{i1},x_{i2})=0 \tag{2.17}$$

for any combination of values of x_{i1} and x_{i2} such that these quantities are well-defined.

Therefore consistent estimators of ATE for movers, $ATE_{01} = E(b_i|0,1)$ and $ATE_{10} = E(b_i|1,0)$, are obtained by simply averaging these noisy estimates across all observations corresponding to movers. To shorten notation, define $n_{x_1x_2} = |\{i = 1,...,n : x_{i1} = x_1, x_{i2} = x_2\}|$, then the estimators for ATE of movers are given by:

$$\hat{ATE}_{01} = \bar{b}_{01} = \frac{1}{n_{01}} \sum_{i=1,...,n:x_{i1}=0,x_{i2}=1} \hat{b}_i, \qquad \hat{ATE}_{10} = \bar{b}_{10} = \frac{1}{n_{10}} \sum_{i=1,...,n:x_{i1}=1,x_{i2}=0} \hat{b}_i$$

In addition, the extrapolation identifying assumption (2.7) together with the result of Proposition 2 implies that for cross-sectional observations that are movers:

$$\hat{a}_i = \alpha_0 + \alpha_1\hat{b}_i + r_i + \zeta_{i,n}, \;\; E(r_i|x_{i1},x_{i2}) = 0, \;\; max_{i=1,...,n}\zeta_{i,n} = O_p(\frac{1}{\sqrt{n}}) \tag{2.18}$$

where $r_i = \epsilon_i + \sum_{t=1,2} u_{it}((1+\alpha_1)(1-x_{it}) - \alpha_1 x_{it})$ is a composite error term composed of the error term ϵ_i in the extrapolation identifying assumption (2.7) and the non-vanishing part of the estimation noise in the estimates of unobserved heterogeneity shown in Proposition 2, and $\zeta_{i,n}$ is a vanishing error term composed of the vanishing part of the estimation noise in the estimates of unobserved heterogeneity shown in Proposition 2.

Up to a vanishing error term, α_0 and α_1 are therefore parameters in an instrumental variable regression model where the observed variable $\hat{a}_i$ is the dependent variable, the observed variable $\hat{b}_i$ is the endogenous covariate, and instrumental variables are given by treatment status history $\{x_{i1}, x_{i2}\}$.

The second step of our estimation procedure estimates α_0 and α_1 by an instrumental variable regression of $\hat{a}_i$ on $\hat{b}_i$ using $\{x_{i1},...,x_{iT}\}$ as instrumental variables. Because the dependent variable $\hat{a}_i$ and the endogenous covariate $\hat{b}_i$ in this instrumental variable regression are only observed simultaneously for movers, this regression is performed using observations on movers only.

With only two time periods, this second step estimator takes the simple form of a Wald

estimator:

$$\hat{\alpha}_1 = \frac{\bar{a}_{01} - \bar{a}_{10}}{\bar{b}_{01} - \bar{b}_{10}}, \qquad \hat{\alpha}_0 = \bar{a}_{01} - \hat{\alpha}_1 \bar{b}_{10}$$
$$\bar{a}_{01} = \frac{1}{n_{01}} \sum_{i=1,\ldots,n: x_{i1}=0, x_{i2}=1} \hat{a}_i, \qquad \bar{a}_{10} = \frac{1}{n_{10}} \sum_{i=1,\ldots,n: x_{i1}=1, x_{i2}=0} \hat{a}_i$$

Given estimates of baseline heterogeneity a_i for untreated stayers, of total heterogeneity $a_i + b_i$ for treated stayers, and of the parameters α_0 and α_1 in the extrapolation identifying assumption, estimates of ATE among untreated and treated stayers are obtained by a simple plug-in estimator using the results from the previous section:

$$\hat{ATE}_{00} = \frac{\bar{a}_{00} - \hat{\alpha}_0}{\hat{\alpha}_1}, \qquad \hat{ATE}_{11} = \frac{\bar{a+b}_{11} - \hat{\alpha}_0}{1 + \hat{\alpha}_1}$$
$$\bar{a}_{00} = \frac{1}{n_{00}} \sum_{i=1,\ldots,n: x_{i1}=0, x_{i2}=0} \hat{a}_i \qquad \bar{a+b}_{11} = \frac{1}{n_{11}} \sum_{i=1,\ldots,n: x_{i1}=1, x_{i2}=1} (\hat{a}_i + \hat{b}_i)$$

In the rest of this section we show regularity conditions under which this second step yields asymptotically normal estimators for α_0 and α_1 and for ATE among stayers.[10]

Assumption 3. *Define* $\pi_{01} = P(x_{i1} = 0, x_{i2} = 1)$ *and* $\pi_{10} = P(x_{i1} = 1, x_{i2} = 0)$.

a) $\pi_{01} > 0$ *and* $\pi_{10} > 0$.

b) $E(b_i|0,1) \neq E(b_i|1,0)$.

c) $Var(r_i|x_{i1}, x_{i2}) > 0$ *if* $x_{i1} \neq x_{i2}$. $Var(a_i + \frac{1}{2}\sum_{t=1,2} u_{it} | \Delta u_{i2}, x_{i1} = 0, x_{i2} = 0) \geq c$ *and* $Var(a_i + b_i + \frac{1}{2}\sum_{t=1,2} u_{it} | \Delta u_{i2}, x_{i1} = 1, x_{i2} = 1) \geq c$ *a.s. for a constant* $c > 0$.

Assumption 3.a requires that there be two types of movers with positive probability. Assumption 3.b is an assumption of relevance of the instrumental variables which requires that the two groups of movers have different ATE. Assumption 3.c is a regularity condition which guarantees that the second step estimators defined above have non-degenerate asymptotic distributions. It imposes that there be variability in the composite error term of the approx-

[10]Only the first step of our estimation procedure is needed to compute estimates of ATE for movers, which are given by $\hat{ATE}_{01}$ and $\hat{ATE}_{10}$ above. Although this is not shown here for concision, one can use the results of Proposition 2 to show that these estimators have linear influence function representations and are asymptotically normal, so that one can estimate unconditional ATE for the entire population by weighing each conditional ATE by the observed frequency of the corresponding subpopulation and the resulting estimator will be asymptotically normal and have a linear influence function representation under the same assumptions as Proposition 3 below.

imate instrumental variable regression model (2.18) and that there be variability in baseline heterogeneity a_i and total heterogeneity $a_i + b_i$ conditional on the error term u_{it} of the CRC model or that there be variability in the error term u_{it} of the CRC model over time.

The following proposition shows that under the CRC model, the extrapolation identifying assumption, and the assumptions above, the second step estimators of α_0, α_1, and of ATE for stayers discussed above have a linear influence function representation and are asymptotically normal.

Proposition 3. *Under the CRC model (2.1), the extrapolation identifying assumption (2.7), and Assumptions 1-3, as $n \to \infty$ we have:*

$$\sqrt{n}(\begin{bmatrix}\hat{\alpha}_0\\ \hat{\alpha}_1\end{bmatrix} - \begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix}) = \sqrt{n}\sum_{i=1}^{n}\xi_{\alpha,i} + \zeta_{\alpha,n} \xrightarrow{d} N(0, V_\alpha) \tag{2.19}$$

where $\xi_{\alpha,i}$ is an i.i.d. sequence of random variables with $E(\xi_{\alpha,i}) = 0$ and $V_\alpha = Var(\xi_{\alpha,i})$, and where $\zeta_{\alpha,n} = o_p(1)$.

If in addition $\alpha_1 \notin \{0, -1\}$, $\pi_{00} > 0$, and $\pi_{11} > 0$, then:

$$\sqrt{n}(\begin{bmatrix}\hat{ATE}_{00}\\ \hat{ATE}_{11}\end{bmatrix} - \begin{bmatrix}ATE_{00}\\ ATE_{11}\end{bmatrix}) = \sqrt{n}\sum_{i=1}^{n}\xi_{ATE,i} + \zeta_{ATE,n} \xrightarrow{d} N(0, V_{ATE}) \tag{2.20}$$

where $\xi_{ATE,i}$ is an i.i.d. sequence of random variables with $E(\xi_{ATE,i}) = 0$ and $V_{ATE} = Var(\xi_{ATE,i})$, and where $\zeta_{ATE,n} = o_p(1)$.

Since these second-step estimators have a linear influence function representation, asymptotically valid inference can be based on Wald tests with variances estimated by bootstrap resampling, as shown in e.g. Mammen (1992). Note that this resampling should be clustered at the level of cross-sectional observations and that both steps of the estimation procedure outlined above need to be applied to each bootstrap sample for the estimated variance to be valid. The online appendix also provides analytical formulae for the asymptotic variance of the second step estimators, so that their variance can also be estimated as sample analogues of their asymptotic variance. The resulting variance estimator uses quantities that

are computed by default in commonly used statistical software, so that implementation is straightforward.

In addition, the online appendix shows that the approach above is easily extended to data with a general number of time periods and models where the extrapolation identifying assumption (2.7) holds but the CRC model (2.1) is replaced by the more general specification:

$$y_{it} = a_i + b_i x_{it} + z_{it}\gamma_0 + u_{it}, \qquad E(u_{it}|X_i, Z_i) = 0 \tag{2.21}$$

where z_{it} are control covariates, $X_i = [x_{it}]_{t=1,...,T}$, and $Z_i = [z_{it}]_{t=1,...,T}$.[11] With this more general specification, the first step of our procedure is replaced by a regression of outcomes y_{it} on a set of indicator variables for each cross-sectional observations, the interaction between these indicator variables and treatment status, and all of the covariates in z_{it}. With the resulting noisy estimators $\hat{a}_i$ of baseline heterogeneity, $\hat{b}_i$ of treatment effect, and $\hat{a}_i + \hat{b}_i$ of total heterogeneity, the second step of our procedure remains unchanged and consists of an instrumental variable regression using observations on movers and plug-in estimators of ATE for untreated and treated stayers.

2.2 Testing the Validity of the Extrapolation

A violation of the condition (2.9), which imposes that factors of selection into treatment other than treatment effect be independent of baseline heterogeneity and treatment effect, would generally lead to a violation of the extrapolation identifying assumption (2.7) and to the extrapolation discussed above being invalid, leading to a bias in estimated ATE among stayers. For concision we will refer below to factors of selection into treatment other than

[11] Setting $z_{it} = [1[t = s]]^{s=1,...,T}$ yields the CRC model considered in (2.1). Note that z_{it} can contain interactions between treatment status x_{it} and control covariates z^1_{it}, so that this model specification encompasses time varying heterogenous effects through observed variables. With additional covariates, Lemieux (1998) and Suri (2011) consider estimation of the model given by:

$$y_{it} = b_i(x_{it} + \alpha_1) + z_{it}\gamma_0 + \tilde{f}_t + \tilde{u}_{it}, \qquad E(\tilde{u}_{it}|X_i, Z_i) = 0$$

which in general requires that $E(u_{it}|X_i, Z_i) = 0$ and $E(\epsilon_i|X_i, Z_i) = 0$ hold, u_{it} being the error term in the CRC model (2.21) and ϵ_i the error term in the extrapolation identifying assumption (2.7). We only impose that $E(u_{it}|X_i, Z_i) = 0$ and $E(\epsilon_i|X_i) = 0$ hold, which allows for dependence between the unobserved heterogeneity terms a_i and b_i and the control covariates $\{z_{i1}, ..., z_{iT}\}$.

treatment effect as the cost of treatment.

To express the form of the bias in estimated ATE for stayers resulting from the cost of treatment being dependent with baseline heterogeneity or treatment effect, we consider for simplicity the case where a single time-constant cost shifter captures the correlation between the cost of treatment and the unobserved heterogeneity determining outcomes. We start by writing the decomposition of the cost of treatment into a time-constant cost shifter c_i and transitory variation in cost ν_{it}:

$$c_{it} = c_i + \nu_{it} \ \forall \, t \tag{2.22}$$

and, instead of the extrapolation identifying assumption (2.7), we assume that:

$$a_i = \alpha_0 + \alpha_1 b_i + \alpha_2 c_i + \epsilon_i, \qquad\qquad E(\epsilon_i | X_i) = 0 \tag{2.23}$$

where α_2 is the partial predictive effect of the cost shifter c_i on productivity.[12]

With two time periods, the pseudo-true values of the slope coefficient and intercept of the extrapolation line discussed in the previous section are then given by:

$$\alpha_1^\star = \alpha_1 + \alpha_2 \frac{E(c_i|1,0) - E(c_i|0,1)}{E(b_i|1,0) - E(b_i|0,1)}, \qquad \alpha_0^\star = E(a_i|1,0) - \alpha_1^\star E(b_i|1,0) \tag{2.24}$$

For simplicity assume that both groups of movers have the same average cost, $E(c_i|1,0) = E(c_i|0,1)$, so that the pseudo-true extrapolation line has the correct slope coefficient α_1, and an intercept given by $\alpha_0^\star = \alpha_0 + \alpha_2 E(c_i|0,1)$.

The pseudo-true values obtained for average treatment effects among stayers would then be given by:

$$ATE_{0,0}^\star = E(b_i|0,0) + \alpha_2 \frac{E(c_i|0,0) - E(c_i|0,1)}{\alpha_1} \tag{2.25}$$

$$ATE_{1,1}^\star = E(b_i|1,1) + \alpha_2 \frac{E(c_i|1,1) - E(c_i|0,1)}{1 + \alpha_1} \tag{2.26}$$

so that the bias in the estimated average treatment effects will depend on the differences in cost across subgroups ($E(c_i|x_1, x_2) \ \forall \, x_1, x_2 \in \{0,1\}$) and the predictive partial effect of cost

[12] Equation (2.23) is obtained from the selection equation (2.9) and the decomposition (2.22) by assuming that: i) the conditional mean of baseline heterogeneity is linear, i.e. $E(a_i|c_i, b_i) = \alpha_0 + \alpha_1 b_i + \alpha_2 c_i$, and that ii) the determinants of cost other than the cost shifter c_i, $\nu_{it} \ \forall \, t$, are independent of the cost shifter c_i and of baseline heterogeneity and treatment effects, i.e. $\{\nu_{it}\}_{t=1,\ldots,T} \perp \{a_i, b_i, c_i\}$.

on baseline heterogeneity (α_2).

In the context of maize production in Kenya, the main determinant of the cost of using hybrid seeds during the period considered in Suri (2011) (1997 to 2004) was a farmer's distance to the nearest seed seller, because the price of hybrid seeds was regulated. Suri (2011) reports evidence that less productive farmers tend to live further away from seed sellers. One can therefore expect that farmers facing higher costs of using hybrid seeds are predicted to have lower levels of baseline productivity a_i, i.e. $\alpha_2 < 0$.

Suri (2011) also reports evidence that farmers who live further away from seed sellers tend to adopt the use of hybrid seeds at lower rates, so that one can expect farmers who never used hybrid seeds to have higher average costs of using hybrid seeds than farmers who adopted the use of hybrid seeds in one or all time periods, and farmers who used hybrid seeds for only one time period to have higher average costs of using hybrid seeds than farmers who used hybrid seeds in all time periods, i.e. $E(c_i|0,0) - E(c_i|0,1) > 0$ and $E(c_i|1,1) - E(c_i|0,1) < 0$.

With α_1 estimated to be between -1 and 0 in Suri (2011), we see that in this particular empirical application one might expect the estimated ATE for both treated and untreated stayers to exhibit a positive bias.

The possibility of bias in the estimated ATE for stayers may lead researchers to wish to test the validity of the extrapolation identifying assumption (2.7). With two time periods, equations (2.10) and (2.11) above show that the extrapolation identifying assumption (2.7) is equivalent to an identity which defines four previously unrestricted parameters in terms of quantities identified by the CRC model (2.1), so that the extrapolation identifying assumption does not contain any testable implications under the CRC model. The online appendix articulates this result in a more direct way.[13]

When three or more time periods are available, the extrapolation identifying assumption can be tested because there are more than two groups of movers for which average baseline

[13]If the cost shifter c_i above were observed, one could estimate α_2 and test whether it is equal to zero by including c_i as a covariate in the instrumental variable regression of the noisy estimates of a_i on the noisy estimates of b_i discussed in the previous section. Here we consider testing the extrapolation identifying assumption (2.7) directly without relying on observed cost shifters.

heterogeneity and ATE are identified by the CRC model. One can test the extrapolation identifying assumption (2.7) by testing whether a linear relationship exists between the points $(E(a_i|x_1, ..., x_T), E(b_i|x_1, ..., x_T))$ identified by the CRC model for all combinations of values $\{x_1, ..., x_T\}$ corresponding to movers. The appendix provides the details of the test.

2.3 Extrapolation in the Presence of Confounding Factors

The extrapolation discussed in Section 2.1 relies on a linear relationship existing between $E(b_i|x_{i1}, ..., x_{iT})$ and $E(a_i|x_{i1}, ..., x_{iT})$. Movers are used to identify the parameter of this linear function, and stayers' observed outcomes can be used to pin down ATE for stayers. This extrapolation is valid if treatment status is irrelevant for predicting baseline heterogeneity conditional on treatment effect ($E(\epsilon_i|X_i) = 0$ in (2.7)). As discussed above, this condition will generally be violated if the cost of treatment is correlated with baseline heterogeneity or treatment effect, which could be a likely scenario in many empirical applications.

In this section we show that confounding factors can be accounted for in a generalized extrapolation as long as they originate from observable sources. We define v_i to be a variable observed by the researcher which indexes cost shifters $d_{v,t}$ common to all observations sharing the same value v of the variable v_i, while the rest of the cost of treatment is captured by a variable ξ_{it}, so that we can write:

$$c_{it} = d_{v_i,t} + \xi_{it} \tag{2.27}$$

In our empirical example, we will take the variable v_i to index a farmer's village, so that $d_{v,t}$ captures all village-specific cost shifters, while ξ_{it} captures the rest of the variation in the cost of using hybrid seeds across farmers.[14]

We assume that the dependence between baseline heterogeneity a_i, treatment effects b_j for j such that $v_j = v_i$, and cost shifters $d_{v_i,t}$ takes the form of a partially linear relationship:

$$a_i = e_{v_i} + \alpha_1 b_i + \epsilon_i, \qquad E(\epsilon_i|\{b_j\}_{j:v_j=v_i}, \{d_{v_i,t}\}_{t=1,...,T}) = 0 \tag{2.28}$$

where e_{v_i} is an unobserved variable which captures the predictive effect on baseline het-

[14] In our empirical example, farmers do not move across villages over time.

erogeneity of group-level information on treatment effects and cost shifters, while α_1 is the partial predictive effect of an observation's own treatment effect on their baseline heterogeneity.

This condition captures the possible statistical dependence between baseline heterogeneity or treatment effect and group-level cost shifters, but imposes that this dependence take the form of a partially linear relationship, i.e. such that the partial predictive effect of b_i on a_i given the group-level variables $\{b_j\}_{j:v_j=v_i}, \{d_{v_i,t}\}_{t=1,...,T}$ is equal to a constant α_1 added to the partial predictive effect of b_i on a_j for $j \neq i$ but $v_j = v_i$.[15]

We also assume that the idiosyncratic cost components are independent of baseline heterogeneity, treatment effect, and of the cost shifters d_{v_i}:

$$\{\xi_{i1}, ..., \xi_{iT}\}_{i:v_i=v} \perp \{a_i, b_i\}_{i:v_i=v}, d_v \tag{2.29}$$

In our empirical example, the structure imposed above allows for dependence between the cost of using hybrid seeds and productivity, provided that this dependence originates from village specific cost shifters only. This accounts for instance for a farmer's distance to the nearest seed seller being correlated with a farmer's productivity since there is little variation in distance to the nearest seed seller across farmers who live in the same village, so that this cost shifter can be taken as common to all farmers who live in the same village. It will also account for the dependence between productivity and any other village-level cost shifter such as transportation amenities, information on hybrid seeds shared across farmers within villages, or the price of hybrid seeds if it varies geographically.[16]

[15]Condition (2.28) is obtained for instance if the group-level dependence in a_i and b_i is captured by:

$$\begin{bmatrix} a_i \\ b_i \end{bmatrix} = \begin{bmatrix} e_{a,v_i} \\ e_{b,v_i} \end{bmatrix} + \begin{bmatrix} \xi_{a,i} \\ \xi_{b,i} \end{bmatrix}, \qquad \{\xi_{a,i}, \xi_{b,i}\}_{i:v_i=v} \perp \{e_{a,v}, e_{b,v}, \{d_{v,t}\}_{t=1,...,T}\}$$

where $\xi_{a,i}$ and $\xi_{b,i}$ are cross-sectionally independent, and if the mean of the idiosyncratic part of baseline heterogeneity conditional on the idiosyncratic part of treatment effect is linear:

$$E(\xi_{a,i}|\xi_{b,i}) = \alpha_0 + \alpha_1 \xi_{b,i}$$

[16]Note that the cost shifters $d_{v,t}$ themselves need not be observed, only the indexing variable v_i is assumed to be observed.

Under (2.9), (2.27), (2.28), and (2.29) we obtain:

$$a_i = e_{v_i} + \alpha_1 b_i + \epsilon_i, \qquad E(\epsilon_i | \{x_{j1}, ..., x_{jT}\}_{j:v_j=v_i}) = 0 \tag{2.30}$$

which we call a generalized extrapolation identifying assumption.

To estimate ATE for stayers with this generalized extrapolation identifying assumption, the first step of our estimation procedure remains unchanged from Section 2.1.3, so that we rely on the same noisy estimates $\hat{a}_i$ and $\hat{b}_i$ of baseline heterogeneity and treatment effect for cross-sectional observations corresponding to movers, $\hat{a}_i$ of baseline heterogeneity for untreated stayers, and $\hat{a}_i + \hat{b}_i$ of total heterogeneity for treated stayers.

Similarly as before, we can show under the CRC model (2.1), the generalized extrapolation assumption (2.30), and new regularity conditions listed below, that an approximate fixed-effects instrumental variable regression model links our noisy estimates $\hat{a}_i$ and $\hat{b}_i$ among cross-sectional observations that are movers:

$$\hat{a}_i = e_{v_i} + \alpha_1 \hat{b}_i + r_i + \zeta_{i,n}, \quad E(r_i | \{x_{j1}, ..., x_{jT}\}_{j:v_j=v_i}) = 0, \quad max_{i=1,...,n} \zeta_{i,n} = O_p(\frac{1}{\sqrt{n}}) \tag{2.31}$$

where as before $r_i = \epsilon_i + \sum_{t=1,2} u_{it}((1+\alpha_1)(1-x_{it}) - \alpha_1 x_{it})$ but with ϵ_i denoting the error term in the generalized extrapolation identifying assumption (2.30).

The second step of our estimation procedure under the generalized extrapolation identifying assumption is therefore given by a fixed-effects instrumental variable regression of $\hat{a}_i$ on $\hat{b}_i$ using $\{x_{i1}, ..., x_{iT}\}$ as instrumental variables, with fixed effects indexed by the variable v_i. As before, only observations on movers are used for this regression.

This new estimation procedure will yield an estimate of α_1, which we redefine to be $\hat{\alpha}_1$, and will also yield estimated fixed effects which we denote by $\hat{e}_v$ for value v of the indexing variable v_i.

Given these estimates, estimated ATE for stayers are redefined to be:

$$\hat{ATE}_{00} = \frac{1}{n_{00}} \frac{\sum_{i:x_{i1}=x_{i2}=0} (\hat{a}_i - \hat{e}_{v_i})}{\hat{\alpha}_1}, \qquad \hat{ATE}_{11} = \frac{1}{n_{11}} \frac{\sum_{i:x_{i1}=x_{i2}=1} (\hat{a}_i + \hat{b}_i - \hat{e}_{v_i})}{1+\hat{\alpha}_1}$$

In the rest of this section we list conditions that guarantee that these estimators are

consistent and asymptotically normal. As before we only consider the case where two time periods are observed here, while the appendix considers the case where a general number of time periods is observed.

For simplicity we assume that observations are obtained by a random sample of clusters, with clusters indexed by the indexing variable v_i. In the online appendix we show that results can be obtained by imposing only independence across observations with different values of the indexing variable v_i. In addition we assume for simplicity that cross-sectional observations belonging to the same cluster are exchangeable. Finally we assume that there are few cross-sectional observations per value of the indexing variable v_i since this corresponds to our empirical example where few farmers are observed in each village.[17]

To state the next assumption, define $N_v = |\{i = 1, ..., n : v_i = v\}|$ to be the number of cross-sectional observations with value v of the indexing variable v_i, $N = |\{v : \exists\, i = 1, ..., n \; s.t. \; v_i = v\}|$ to be the number of values of the indexing variable v_i, and index cross-sectional observations with the same value v of v_i by i_v, so that $\{i_v : i = 1, ..., n_v\} = \{i = 1, ..., n : v_i = v\}$.

Assumption 4. *Observations* $\{\{x_{i_v 1}, y_{i_v 1}, x_{i_v 2}, y_{i_v 2}, a_{i_v}, b_{i_v}\}_{i=1,...,n_v}\}_{v=1,...,N}$ *are i.i.d. across* v. *The number of observations sharing the same value of the indexing variable* v_i *is bounded, i.e.* $N_v \leq C \;\forall\, v = 1, ..., N, \;\forall\, N$ *for a constant* C. *Cross-sectional observations with the same value of* v_i *are exchangeable.*

With cluster dependence rather than cross-sectional independence as in section 2.1.3, the assumption that the error term u_{it} in the CRC model (2.1) is not degenerate needs to be slightly reformulated compared to Assumption 2 above.

Assumption 5. *Assumptions 2.a and 2.b hold. Redefine* $\sigma^2_{\Delta u,S} = Var(\sum_{i:v_i=v, x_{i1}=x_{i2}} \Delta u_{i2})$, *then* $\sigma^2_{\Delta u,S} > 0$.

[17] We observe an average of twelve farmers per village and a total of 1,130 farmers in our empirical application. Similar results as the ones we derive in this section are obtained in a straightforward way when v_i takes only few values as long as the strength of cross-sectional dependence is limited. One could also extend these results to continuously valued indexing variables v_i by considering local differencing estimators.

The generalized extrapolation (2.30) discussed in this section accounts for more flexible patterns of dependence between the factors that determine selection into treatment (c_{it}) and the terms unobserved heterogeneity that determine outcome (a_i and b_i), but consistent estimation will require overlap conditions that were not needed with the simple extrapolation discussed in Section 2.1. The overlap conditions stated in the next assumption guarantee that instrumental variables in the second step of our procedure are relevant even after partialling out variation across observations sharing the same value of the indexing variable v_i and that stayers can be compared to movers with the same value of the indexing variable v_i.

Assumption 6. *Define the events* $M_v = \{\exists\, i \; s.t. \; v_i = v \text{ and } x_{i1} \neq x_{i2}\}$ *and* $T_v = \{\exists\, i, j \; s.t. \; v_i = v_j = v, \text{ and } x_{i1} \neq x_{i2},\, x_{j1} \neq x_{j2},\, x_{i1} \neq x_{j1}\}$. *Define* $\pi_M = P(M_v)$, $\pi_T = P(T_v)$.

a) $\pi_T > 0$.

b) $\pi_M = 1$.

Define $b_{x_1x_2,v} = E(b_i | v_i = v,\, (x_{i1}, x_{i2}) = (x_1, x_2), \{x_{j1}, x_{j2}\}_{j \neq i : v_j = v})$ *for* $x_1, x_2 \in \{0, 1\}$.

c) $b_{01,v} > b_{10,v}$ *in* T_v *or* $b_{01,v} < b_{10,v}$ *in* T_v.

Define $n_v = |\{i \in M_n : v_i = v\}|$, $\dot{r}_i = r_i - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} r_j$ *and* $\tilde{u}_i = \frac{1}{2} \sum_{t=1,2} u_{it} - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} r_j$.

d) $Var\left(\begin{bmatrix} \sum_{i : v_i = v, x_{i1} = x_{i2}} \Delta u_{i2} \\ \sum_{i : v_i = v, x_{i1} = 0, x_{i2} = 1} \dot{r}_i \\ \sum_{i : v_i = v, x_{i1} = x_{i2} = 0} (a_i - e_{v_i} + \tilde{u}_i) \\ \sum_{i : v_i = v, x_{i1} = x_{i2} = 1} (a_i + b_i - e_{v_i} + \tilde{u}_i) \end{bmatrix} \right)$ *is positive definite.*

In the context of our empirical application, Assumption 6.a requires that many villages have movers of two profiles, i.e. adopters and disadopters. Assumption 6.a is imposed since otherwise there would be no variation in the instrumental variables $\{x_{i1}, x_{i2}\}$ across movers in the same village. Assumption 6.b requires that all villages have at least one mover. This condition is required for the village level fixed effect $\hat{e}_{v_i}$ to be defined for all cross-sectional observations, i.e. for a comparison to be possible between every stayer and at least one mover living in the same village. In practice if some villages do not have movers, one can

report ATE for stayers conditional on belonging to a village with at least one mover. In our empirical application, this leads us to report ATE for 91% of stayers, as 9% of stayers lived in a village without any mover. Assumption 6.c imposes that movers of different profiles living in the same village have different ATE.

Similarly as for Assumption 3.c, Assumption 6.d is a regularity condition which guarantees that the second step estimators discussed here have a non-degenerate asymptotic distribution. It imposes that there be no perfect linear dependence between cross-sectional observations belonging to the same group, that there be within-group variation in the composite error term of the new approximate instrumental variable regression model (2.31), and that there be variation in the terms of heterogeneity a_i and b_i that is not perfectly linearly dependent of the error term u_{it} of the CRC model or variation in the error term u_{it} over time.

Under these assumptions, our new estimators of α_1 and of ATE among stayers are asymptotically normal and have a linear influence function representation. Note that under Assumption 4, N and n are of the same order.

Proposition 4. *Under the CRC model (2.1), the generalized extrapolation identifying assumption (2.30), and Assumptions 4-6, as $N \to \infty$ we have:*

$$\sqrt{N}(\hat{\alpha}_1 - \alpha_1) = \sqrt{N} \sum_{v=1}^{N} \xi_{\alpha,v} + \zeta_{\alpha,N} \xrightarrow{d} N(0, V_\alpha) \tag{2.32}$$

where $\xi_{\alpha,v}$ is an i.i.d. sequence of random variables with $E(\xi_{\alpha,v}) = 0$ and $V_\alpha = Var(\xi_{\alpha,v})$, and where $\zeta_{\alpha,N} = o_p(1)$.

If in addition $\alpha_1 \notin \{0, -1\}$, then:

$$\sqrt{N}\left(\begin{bmatrix} \hat{ATE}_{00} \\ \hat{ATE}_{11} \end{bmatrix} - \begin{bmatrix} ATE_{00} \\ ATE_{11} \end{bmatrix}\right) = \sqrt{N} \sum_{v=1}^{N} \xi_{ATE,v} + \zeta_{ATE,N} \xrightarrow{d} N(0, V_{ATE}) \tag{2.33}$$

where $\xi_{ATE,v}$ is an i.i.d. sequence of random variables with $E(\xi_{ATE,v}) = 0$ and $V_{ATE} = Var(\xi_{ATE,v})$, and where $\zeta_{ATE,N} = o_p(1)$.

As before, Proposition 4 shows that asymptotically valid inference can be obtained by

using Wald tests with variance estimated by cluster bootstrap, clusters now being indexed by the indexing variable v_i. The appendix also provides formulae for analytical standard errors which are sample analogues of the asymptotic variance of these new second step estimators. In addition the generalized extrapolation assumption (2.30) can be tested when more than two time periods are observed by testing whether a linear relationship exists between the baseline heterogeneity and treatment effect of movers of different treatment status history profiles after partialling out variation at the level of the indexing variable v_i. Details of the test are provided in the appendix.

3 Empirical Application to Estimating Returns to Agricultural Technology Adoption

In this section we apply the extrapolations and tests defined in Section 2 to a longitudinal dataset of Kenyan maize farmers. The dataset was collected as part of the Tegemeo Agricultural Monitoring and Policy Analysis Project by the Tegemeo Institute at Egerton University and Michigan State University. This same dataset was also used in Suri (2011), and hence we refer the reader to Suri (2011) for a detailed discussion of the data and of the related empirical literature. The only notable difference here is that we use data on years 1997, 2004, 2007, and 2010 while Suri (2011) only used waves 1997 and 2004 for her study as these were the only two waves available at the time.

We observe a total of 1,130 farmers in our estimation sample but in an unbalanced panel with a total of 3,770 observations, so that farmers are observed for around 3.5 time periods on average. The sample is decomposed into 354 farmers observed both using and not using hybrid seeds over time (movers), 123 farmers never observed using hybrid seeds (untreated stayers), and 653 farmers always observed using hybrid seeds (treated stayers).

The adoption rate of hybrid seeds increased sharply over our period of observation: 72% of farmers in our sample used hybrid seeds in 1997, 71% in 2004, 77% in 2007, and 88% in 2010. This reflects an increase in the ease of access to hybrid seeds, which is also evidenced

by the decrease in average distance to the nearest seed seller over time, from 6.1 kilometers on average in 1997, to 2.6km in 2004, 2.9km in 2007, 3.3km in 2010.

We estimate returns to technology adoption, i.e. to using hybrid seeds, with the model:

$$y_{it} = a_i + b_i x_{it} + z_{it}\gamma_t + u_{it}, \qquad E(u_{it}|X, Z) = 0 \tag{3.1}$$

where y_{it} and x_{it} are maize yields (logarithm of kilograms harvested per acre) and hybrid seed use, $X = [x_{i1}, ..., x_{iT}]_{i=1,...,n}$, $Z = [z_{i1}, ..., z_{iT}]_{i=1,...,n}$, and z_{it} is the same vector of control covariates as in Suri (2011), namely: main season rainfall, variables measuring other inputs to production than hybrid seed use, acres planted, and demographics of the household such as size, gender distribution, and age. We also include province-by-year fixed effects in the controls to account for regional time varying shocks that might have occurred during the fairly long period of observation.

As discussed above, estimates of the coefficients γ_t and noisy estimates of baseline heterogeneity a_i, treatment effect b_i, total heterogeneity $a_i + b_i$ (depending on a cross-sectional observation's treatment status history) are obtained by an ordinary least squares regression of y_{it} on indicator variables for each cross-sectional observation, the interaction of these indicator variables with hybrid seed use, and the covariates z_{it} interacted with indicator variables for each time period.[18]

Given these estimates, estimates of average returns (ATE in the general discussion of Section 2) are obtained for different groups of movers which we present in Table 1. Returns for movers are estimated to be 23% on average, but with substantial variation across subgroups of movers. Movers who used hybrid seeds early in the period of observation, i.e. who abandoned the use of hybrid seeds in later years, are estimated to have relatively low average returns (movers who used hybrid seeds in 1997 and 2004 are estimated to have average returns to using hybrid seeds of 8% and 11%, respectively). Movers who used hybrid seeds later, i.e.

[18]Here we observe four time periods but an unbalanced panel. We treat data as missing at random. Every cross-sectional observation with three or more observed time periods participates in the estimation of the coefficients γ_t, regardless of whether they are stayers or movers. Among cross-sectional observations with only two observed time periods, only stayers participate in the estimation of the coefficients γ_t.

who adopted the use of hybrid seeds in later years, are estimated to have higher average returns (movers who used hybrid seeds in 2007 and 2010 are estimated to have average returns to using hybrid seeds of 23% and 28% respectively). Similarly movers who were not using hybrid seeds in early years are estimated to have higher average returns than movers who were not using hybrid seeds in later years (for instance average returns are estimated to be 37% for movers not using hybrid seeds in 1997 and around 0% for movers not using hybrid seeds in 2010).

Low average returns among early adopters who disadopted in later time periods are consistent with these farmers being marginal hybrid seed users. High average returns among late adopters are consistent with a technology diffusion process such that some farmers who would have benefited from high returns to adoption did not use hybrid seeds in early years - perhaps because of high costs of adoption - but did gain access to hybrid seeds in later years, providing further evidence that access to hybrid seeds has improved over time.

On the other hand, a significant number of farmers in our sample is still never observed using hybrid seeds. The methods developed above can be used to estimate whether these farmers would experience high returns from adopting hybrid seeds.

Using the noisy estimates of baseline heterogeneity a_i and returns b_i among movers obtained by the OLS regression described above, we can estimate the parameters α_0 and α_1 of the extrapolation identifying assumption given by (2.7). The appendix details our estimation procedure which is an extension of the procedure defined above to the case where an unbalanced panel with more than two time periods is observed. Table 1 reports these estimates, with α_1 in particular estimated to be -0.49, so that a negative statistical relationship is estimated to exist between baseline productivity and returns (i.e. on average low productivity farmers are estimated to benefit from higher returns from using hybrid seeds than high productivity farmers). We can also estimate the slope coefficient α_1 of the generalized extrapolation identifying assumption given by (2.30) where v_i indexes farmer i's village. With this generalized extrapolation, we estimate α_1 to be -0.95, so that the

predictive effect of returns on baseline productivity is still estimated to be negative when accounting for the predictive effect of village-level cost shifters, but the magnitude of this predictive effect is estimated to be larger after controlling for village-level cost shifters. Since this last estimate of α_1 is close to -1, we concentrate on non-hybrid (untreated) stayers in the remainder of this section, and do not obtain results for hybrid (treated) stayers. The generalized extrapolation requires that a stayer live in a village with at least one observed mover. In our data 91% of farmers who never used hybrid seeds live in a village with at least one mover, so the results below are reported for these farmers only.

Table 1 reports average returns for non-hybrid stayers observed in each year using either extrapolation (the simple extrapolation based on (2.7) is dubbed non-robust and denoted by NR, the generalized extrapolation based on (2.30) is dubbed robust and denoted by R). We find that there are large differences in the estimated ATE for non-hybrid stayers across extrapolations. The non-robust extrapolation estimates the average returns for non-hybrid stayers to be 66%, i.e. much larger than the average returns for movers, while the robust extrapolation estimates average returns for non-hybrid stayers to be 37%, i.e. larger than returns for movers but only marginally so.[19] Both sets of results point toward high returns farmers still being excluded from access, possibly because of high costs of adoption, but using the robust extrapolation leads to estimating more moderate gains from expanding access to hybrid seeds among non-hybrid stayers than when using the non-robust extrapolation.

Finally Table 1 presents results from testing the validity of the non-robust and robust extrapolations. In line with the suggestive evidence discussed in Section 2.2 that the cost of using hybrid seeds might be predictive of productivity and that untreated stayers might face higher costs of using hybrid seeds, we find strong evidence against the non-robust extrapolation using our proposed testing procedure, with a p-value close to zero. We find significantly weaker evidence against the robust extrapolation, with a p-value of 0.26.

[19]For comparison, Carter et al. (2017) find an average increase in productivity of 41% from using hybrid seeds in a randomized control trial. Suri (2011) finds average returns of 100% for non-hybrid stayers over the period 1997 and 2004 only.

4 Conclusion

In this paper we explored how to combine models of selection with correlated random coefficient models of panel data to identify ATE for stayers instead of restricting one's attention to movers. We propose simple estimation and testing procedures and find that when applied to estimating the returns to technology adoption, being able to test the extrapolation of ATE to non-hybrid stayers and being able to estimate a generalized extrapolation has first-order implications for empirical results. We hope that these results participate in widening the applicability of correlated random coefficient models when estimating treatment or partial effects with panel data.

The discussion above could be generalized to models of the form:

$$Y_i = h_1(Z_{1,i}, \gamma_{01})B_i + h_2(Z_{2,i}, \gamma_{02}) + U_i, \qquad E(U_i|Z_{1,i}, Z_{2,i}) = 0 \tag{4.1}$$

where $h_1(.,.)$ is a known function of dimension $T \times K$ and $h_2(.,.)$ is a known function of dimension $T \times 1$. This is the class of models considered in Section 4 of Chamberlain (1992).

With continuous covariates, estimation of these models requires $K + 1$ time periods to achieve a parametric convergence rate, or K time periods if a trimming estimator is used as in Graham and Powell (2012). Reducing the dimension of the random coefficient in the model would allow for estimation with fewer time periods or for more precise estimation where trimming is not required.

With discrete covariates, the subpopulation for which ATE or average partial effects are identified is more narrow when the the dimension of $h_1(Z_{1,i}, \gamma_0)$ is larger, while reducing the dimension of the random coefficients would yield identification of ATE or average partial effects on broader subpopulations.

One could model the statistical relationship between elements of B_i by using a common factor structure

$$B_i = \Gamma_0 E_i + \epsilon_i, \qquad E(\epsilon_i|E_i) = 0 \tag{4.2}$$

where the dimension of E_i is $K^{'} < K$. Selection into all covariates in $Z_{1,i}$ could be represented

as $Z_{1,i} = G(E_i, C_i)$ and assuming that $C_i \perp \{B_i, E_i\}$ would imply:

$$E(B_i|Z_{1,i}) = \Gamma_0 E(E_i|Z_{1,i}) \tag{4.3}$$

One could then study what variation in $Z_{1,i}$ would identify Γ_0 and average partial effects for stayers, how to test the validity of the extrapolation in this context, and how to accommodate for dependence between selection factors C_i and unobserved heterogeneity $\{B_i, E_i\}$.

References

Alderman, H. (2007): "Improving Nutrition through Community Growth Promotion: Longitudinal Study of the Nutrition and Early Child Development Program in Uganda," *World Development*, 35, 1376–1389.

Angrist, J. D. and I. Fernández-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," *Advances in Economics and Econometrics: Tenth World Congress*.

Angrist, J. D. and M. Rokkanen (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff," *Journal of the American Statistical Association*, 110, 1331–1344.

Arcidiacono, P., J. Cooley, and A. Hussey (2008): "The Economic Returns to an MBA," *International Economic Review*, 49, 873–899.

Arellano, M. and S. Bonhomme (2011): "Nonlinear Panel Data Analysis," *Annual Review of Economics*, 3, 395–424.

——— (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *The Review of Economic Studies*, 79, 987–1020.

Ashenfelter, O. and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics*, 67, 648–660.

Behrman, J. R. and J. Hoddinott (2005): "Programme Evaluation with Unobserved Heterogeneity and Selective Implementation: The Mexican PROGRESA Impact on Child Nutrition," *Oxford Bulletin of Economics and Statistics*, 67, 547–569.

Bertanha, M. (2017): "Regression Discontinuity Design with Many Thresholds," SSRN Scholarly Paper ID 2712957, Social Science Research Network, Rochester, NY.

Bertanha, M. and G. W. Imbens (2014): "External Validity in Fuzzy Regression Discontinuity Designs," Working Paper 20773, National Bureau of Economic Research.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039.

Browning, M. and J. Carro (2007): "Heterogeneity and Microeconometrics Modeling," *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*.

Card, D. (1996): "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957–979.

Card, D. and D. Sullivan (1988): "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica*, 56, 497–530.

Carter, M., M. Mathenge, S. Bird, T. Lybbert, T. Njagi, and E. Tjernström (2017): "Policy Brief: Local Seed Company Fills a Niche to Increase Maize Productivity in Kenya," *Innovation Lab for Assets and Market Access Policy Brief*.

Cattaneo, M. D., L. Keele, R. Titiunik, and G. Vazquez-Bare (2016): "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *The Journal of Politics*, 78, 1229–1248.

Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.

——— (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567–596.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81, 535–580.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor (2010): "Teacher Credentials and Student Achievement in High School A Cross-Subject Analysis with Student Fixed Effects," *Journal of Human Resources*, 45, 655–681.

de Chaisemartin, C. and X. D'Haultfoeuille (2018a): "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, 85, 999–1028.

——— (2018b): "Two-way fixed effects estimators with heterogeneous treatment effects," *Working Paper*.

Dong, Y. and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *The Review of Economics and Statistics*, 97, 1081–1092.

Fernández-Val, I. and J. Lee (2013): "Panel data models with nonadditive unobserved heterogeneity: Estimation and inference," *Quantitative Economics*, 4, 453–481.

Freeman, R. B. (1984): "Longitudinal Analyses of the Effects of Trade Unions," *Journal of Labor Economics*, 2, 1–26.

Ghanem, D. (2017): "Testing identifying assumptions in nonseparable panel data models," *Journal of Econometrics*, 197, 202–217.

Graham, B. S. and J. L. Powell (2012): "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models," *Econometrica*, 80, 2105–2152.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin (1998): "Does Special Education Raise Academic Achievement for Students with Disabilities?" Working Paper 6690, National Bureau of Economic Research.

Heckman, J. J. and E. J. Vytlacil (2007): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 4875–5143.

Jakubson, G. (1991): "Estimation and Testing of the Union Wage Effect Using Panel Data," *The Review of Economic Studies*, 58, 971–991.

Kline, P. and C. R. Walters (2018): "On Heckits, LATE, and Numerical Equivalence | The Econometric Society," *Econometrica*, Forthcoming.

Kowaleski-Jones, L. and G. J. Duncan (2002): "Effects of Participation in the WIC Program on Birthweight: Evidence From the National Longitudinal Survey of Youth," *American Journal of Public Health*, 92, 799–804.

Lemieux, T. (1998): "Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection," *Journal of Labor Economics*, 16, 261–291.

Mammen, E. (1992): *When Does Bootstrap Work?: Asymptotic Results and Simulations*, Lecture Notes in Statistics, New York: Springer-Verlag.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619.

Rokkanen, M. A. T. (2015): "Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design," .

Rouse, C. E. (1998): "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *The Quarterly Journal of Economics*, 113, 553–602.

Suri, T. (2011): "Selection and Comparative Advantage in Technology Adoption," *Econometrica*, 79, 159–209.

WOOLDRIDGE, J. M. (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *The Review of Economics and Statistics*, 87, 385–390.

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, second ed.

WÜTHRICH, K. (2018): "A comparison of two quantile models with endogeneity," *Journal of Business & Economic Statistics*, 0, 1–36.

Figure 1: Extrapolation from difference-in-differences estimates to stayers

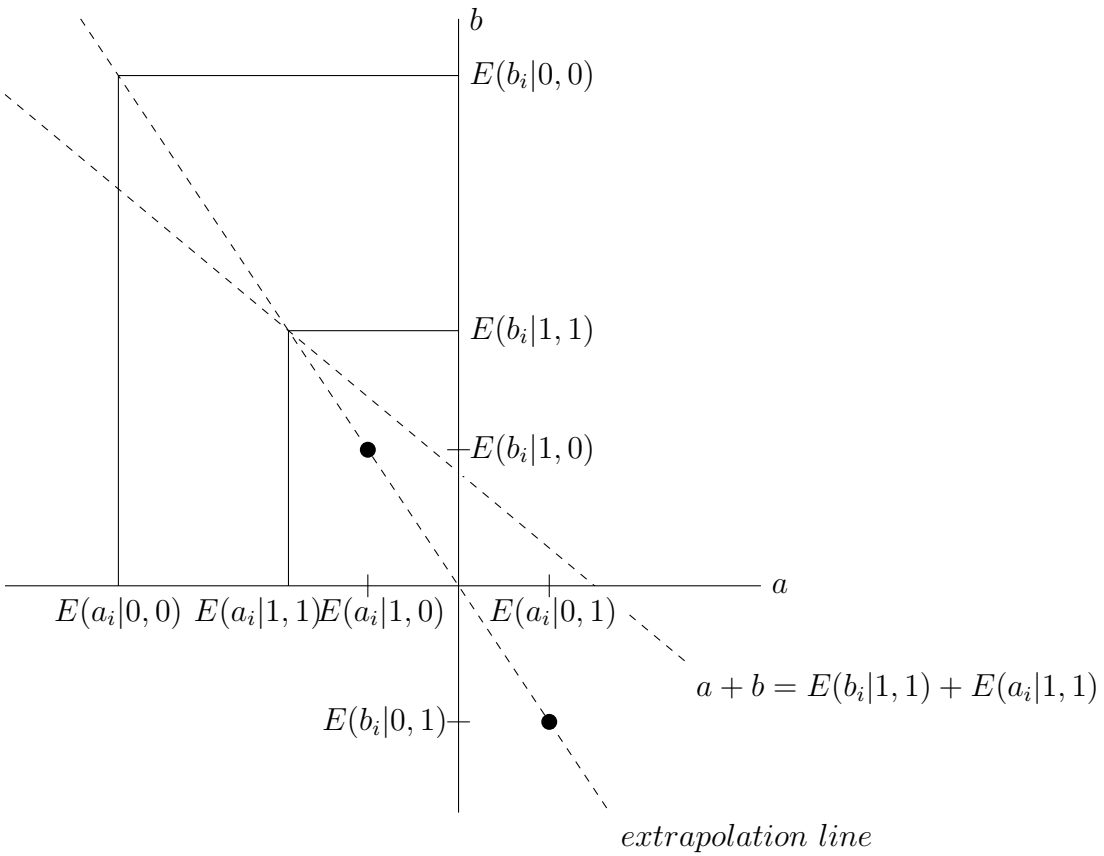

Table 1: Average baseline productivity and returns to using hybrid seeds for different sub-populations and estimation methods.

	Mover currently using hybrid					
	observations	baseline		return		
1997	137	4.896	(0.355)	0.080	(0.088)	
2004	115	4.937	(0.390)	0.112	(0.066)	
2007	165	4.928	(0.358)	0.232	(0.077)	
2010	251	4.882	(0.331)	0.283	(0.069)	
	Mover currently not using hybrid					
	observations	baseline		return		
1997	173	4.804	(0.338)	0.373	(0.072)	
2004	198	4.878	(0.322)	0.230	(0.087)	
2007	125	4.697	(0.323)	0.295	(0.102)	
2010	38	4.908	(0.330)	-0.001	(0.170)	
	non-hybrid stayer					
	observations	baseline		return (NR)		return (R)	
1997	95	4.692	(0.351)	0.717	(0.374)	0.352	(0.121)
2004	84	4.748	(0.354)	0.603	(0.320)	0.326	(0.103)
2007	71	4.638	(0.344)	0.826	(0.444)	0.447	(0.112)
2010	59	4.738	(0.327)	0.623	(0.299)	0.370	(0.148)
α_0				5.044	(0.407)		
α_1				-0.491	(0.348)	-0.949	(0.329)
p-value				0.004		0.257	

Standard errors, which are between parenthesis, account for the estimation noise originating from both steps of estimation and are robust to cluster dependence at the village level. There are 95 clusters (villages) in the first step of our estimation procedure, and 78 clusters in the second step. For non-hybrid stayers, NR denotes that the non-robust extrapolation was used, while R denotes that the robust extrapolation was used. p-value refers to the p-value obtained from testing the validity of the non-robust and robust extrapolations.