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

Online Appendix

Valentin Verdier*

May 26, 2020

*Department of Economics, University of North Carolina - Chapel Hill.

Contents

In this appendix we provide the technical details corresponding to our discussion in the main text. Section A briefly discusses identification of time effects and average treatment effects (ATE) with a correlated random coefficient (CRC) model and a general number of time periods, while the main text focused on the case where two time periods are observed for simplicity. Section B discusses the estimation methods of Lemieux (1998) and Suri (2011). Section C compares the sufficient set of conditions for the extrapolation identifying assumption to hold discussed in the main text with generalized Roy models. Section D discusses CRC models with time-varying treatment effects. Section E discusses the two-step estimation method defined in the main text and the test of validity of the extrapolation identifying assumption for the simple extrapolation. Section F discusses the extension of these methods to using the generalized extrapolation identifying assumption discussed in the main text. Section G discusses the implications of using unbalanced panels due to data missing at random. Section H contains proofs for the results of Sections E and F. Propositions 1-3 in the main text are obtained as special cases of the propositions stated in Sections E and F, but with more cumbersome notation. Therefore we also write in Section H proofs for Propositions 1-3 in the main text. In Section I we briefly discuss the consequences of learning about treatment effects (returns in our empirical application) upon being treated on the validity of the CRC model and extrapolation identification assumptions discussed in the main text.

Throughout the appendix, c will denote an arbitrary positive constant $c > 0$ and C will denote an arbitrary constant $C < \infty$. We use $|\mathcal{A}|$ to denote the cardinality of any set $\mathcal{A}$. We use the notation O_p to denote that a sequence is bounded in probability and o_p to denote that a sequence converges in probability. Throughout the appendix, referenced equations corresponding to numbered sections are found in the main text, while equations corresponding to sections indexed by letters are found in this appendix.

A Identification with a General Number of Time Periods

In this section we discuss the information contained in the CRC model (2.4) for a general number of time periods T, while the main text only considered $T = 2$ for simplicity.

Under cross-sectional independence, we can stack observations across time and rewrite the CRC model (2.4) as:

$$Y_i = W_i \begin{bmatrix} a_i \\ b_i \end{bmatrix} + f + U_i, \qquad E(U_i|X_i) = 0 \tag{A.1}$$

where $Y_i = [y_{it}]_{t=1,...,T}$, $W_i = \begin{bmatrix} j_T & X_i \end{bmatrix}$, $j_T = [1]_{t=1,...,T}$, $X_i = [x_{it}]_{t=1,...,T}$, $f = [f_t]_{t=1,...,T}$, $U_i = [u_{it}]_{t=1,...,T}$.

As in Chamberlain (1992), since the relationship between baseline heterogeneity (a_i), treatment effect (b_i), and treatment status history (X_i) is left unrestricted, the information for estimating f contained in (A.1) is equivalent to the information contained in:

$$E(M_{W_i}(Y_i - f)|X_i) = 0 \tag{A.2}$$

where $M_{W_i} = I_T - W_i(W_i'W_i)^- W_i'$ and $(.)^-$ is a generalized inverse operator.

As in the main text, we will apply the normalization $f_1 = 0$, so that time effects f_t are identified if M_{W_i} has rank greater than $T - 1$ for some values of X_i corresponding to a positive probability. For values of X_i corresponding to stayers, M_{W_i} is the projection matrix of a regression on an constant using T observations, so that it has rank $T - 1$, leading to identification of time effects from observations on stayers.

With two time periods, $M_{W_i} = 0$ for cross-sectional observations i that correspond to movers. However with three or more time periods, $rank(M_{W_i}) \geq T-2 > 0$, and observations on movers participate in the identification of time effects f_t. Depending on the profiles of

treatment status history observed in the data, it is possible for all time effects f_t, $t = 2, ..., T$ to be identified by observations on movers only when $T \geq 3$.

The CRC model (A.1) is equivalent to (A.2) and:

$$E(\begin{bmatrix} a_i \\ b_i \end{bmatrix} |X_i) = E(B_i(Y_i - f)|X_i) + (I - B_i W_i)\zeta_i \tag{A.3}$$

where $B_i = (W_i' W_i)^{-} W_i'$ and ζ_i is an unknown, unrestricted term of heterogeneity.

This shows that conditional average treatment effect $E(b_i|X_i)$ is only identified for cross-sectional observations such that $W_i' W_i$ is non-singular. With x_{it} being binary, $W_i' W_i$ is non-singular for movers and singular for stayers, so that average treatment effects are only identified for movers.

B Notation and Comparison with the Methods of Suri 2011 and Lemieux 1998

In this section we describe the estimation procedures used by Lemieux (1998) and Suri (2011) and show that they can be represented by the linear extrapolation discussed in the main text when there are no additional covariates in the model, i.e. when treatment status x_{it} is the only covariate.

First we map the notation used by Lemieux (1998) and Suri (2011) to the notation used in the main text. In the simple setting without additional covariates, our notation for the correlated random coefficient model with the extrapolation identifying assumption and two time periods is given by:

$$y_{it} = a_i + b_i x_{it} + f_t + u_{it} \qquad E(u_{it}|x_{i1}, x_{i2}) = 0 \tag{B.1}$$

$$a_i = \alpha_0 + \alpha_1 b_i + \epsilon_i \qquad E(\epsilon_i|x_{i1}, x_{i2}) = 0 \tag{B.2}$$

In the simple setting without additional covariates, the notation used by Lemieux (1998) writes potential outcomes without (N) or with (U) treatment as:

$$y_{it}^N = \delta_t^N + \theta_i^N + \epsilon_{it}^{'}$$
$$y_{it}^U = \delta_t^N + \bar{\delta} + \theta_i^U + \epsilon_{it}^{'}$$

where θ_i^U and θ_i^N have mean zero, so that the average treatment effect is given by $\bar{\delta}$ and $\epsilon_{it}^{'}$ are unobserved wage shocks.

Therefore our notation writes $f_t = \delta_t^N$, $a_i = \theta_i^N$, $b_i = \bar{\delta} + \theta_i^U - \theta_i^N$, $u_{it} = \epsilon_{it}^{'}$.

The notation for the extrapolation identifying assumption in Lemieux (1998) is given by the linear projections:

$$\theta_i^N = b_N(\theta_i^N - \theta_i^U) + \xi_i$$
$$\theta_i^U = b_U(\theta_i^N - \theta_i^U) + \xi_i$$

where $b_N = \frac{Cov(\theta_1^N, \theta_i^N - \theta_i^U)}{Var(\theta_i^N - \theta_i^U)}$ and $b_U = \frac{Cov(\theta_1^U, \theta_i^N - \theta_i^U)}{Var(\theta_i^N - \theta_i^U)}$.

Lemieux (1998) then defines $\epsilon_{it} = \xi_i + \epsilon_{it}^{'}$ and assumes:

$$E(\epsilon_{it}|x_{i1}, x_{i2}) = 0 \tag{B.3}$$

In our notation we have $\alpha_1 = b_N$, $\epsilon_i = \xi_i$, $\alpha_1 + 1 = b_U$.

Lemieux (1998) defines $\theta_i = b_N(\theta_i^N - \theta_i^U)$ and $\psi = \frac{b_U}{b_N}$ so that:

$$\theta_i^N = \theta_i + \xi_i$$
$$\theta_i^U = \psi\theta_i + \xi_i$$

In our notation we have $\frac{\alpha_1+1}{\alpha_1} = \psi$, $\alpha_1 b_i = \theta_i$.

The model for observed outcomes estimated by Lemieux (1998) is then given by:

$$y_{it} = \delta_t^N + \bar{\delta} x_{it} + \theta_i(1 + (\psi - 1)x_{it}) + \epsilon_{it} \qquad E(\epsilon_{it}|x_{i1}, x_{i2}) = 0 \tag{B.4}$$

In the next two subsections we show that under the CRC model assumption (B.1) only, estimation of all the parameters in the model (B.4) that combines the CRC model with the extrapolation identifying assumption (B.2) by generalized method of moments estimation (Lemieux (1998)) or minimum distance estimation (Suri (2011)) leads to the linear extrapolation from ATE among movers to ATE among stayers depicted in Figure 1 of the main text.

B.1 Lemieux 1998

The estimation procedure proposed by Lemieux (1998) is GMM estimation from the moment conditions:

$$E(\begin{bmatrix} x_{i1}x_{i2} \\ (1 - x_{i1})(1 - x_{i2}) \\ x_{i1}(1 - x_{i2}) \\ (1 - x_{i1})x_{i2} \end{bmatrix} e_i(\delta_2^N, \bar{\delta}, \psi, \delta_1^N) = 0$$

$$E(\theta_i) = 0$$

where the moment function $e_i(.)$ is defined to be:

$$e_i(\delta_2^N, \bar{\delta}, \psi, \delta_1^N) = y_{i2} - \delta_2^N - \bar{\delta} x_{i2} - \frac{1 + (\psi - 1)x_{i2}}{1 + (\psi - 1)x_{i1}}(y_{i1} - \delta_1^N - \bar{\delta} x_{i1})$$

These moment conditions can be rewritten as:

$$E(\Delta y_{i2} - \Delta\delta_2^N|0,0) = 0$$
$$E(\Delta y_{i2} - \Delta\delta_2^N|1,1) = 0$$
$$E(y_{i2} - \delta_2^N - \bar{\delta} - \psi(y_{i1} - \delta_1^N)|0,1) = 0$$
$$E(y_{i2} - \delta_2^N - \frac{1}{\psi}(y_{i1} - \delta_1^N - \bar{\delta})|1,0) = 0$$
$$\sum_{t=1}^{2} E(\frac{y_{it} - \delta_t^N - \bar{\delta}x_{it}}{1 + (\psi - 1)x_{it}}) = 0$$

Under the CRC model (B.1), with the normalization $f_1 = 0$, we can further re-write these moment conditions:

$$\Delta f_2 - \Delta\delta_2^N = 0 \tag{B.5}$$
$$\Delta f_2 - \Delta\delta_2^N = 0 \tag{B.6}$$
$$E(a_i + b_i - \delta_1^N - \bar{\delta} - \psi(a_i - \delta_1^N)|0,1) = 0 \tag{B.7}$$
$$E(a_i - \delta_1^N - \frac{1}{\psi}(a_i + b_i - \delta_1^N - \bar{\delta})|1,0) = 0 \tag{B.8}$$
$$\sum_{t=1}^{2} E(\frac{a_i + b_i x_{it} + f_t - \delta_t^N - \bar{\delta}x_{it}}{1 + (\psi - 1)x_{it}}) = 0 \tag{B.9}$$

We see that the first two moment conditions (B.5) and (B.6) contain the same information under the CRC model (B.1), setting $\Delta\delta_2^N$ equal to Δf_2. Therefore the remaining three moment conditions (B.7)-(B.9) are exactly identifying for ψ, δ_1^N, $\bar{\delta}$.

Equations (B.7) and (B.8) imply:

$$\psi = \frac{E(a_i + b_i|0,1) - E(a_i + b_i|1,0)}{E(a_i|0,1) - E(a_i|1,0)} \tag{B.10}$$

Defining $\alpha_1^\star = \frac{E(a_i|0,1) - E(a_i|1,0)}{E(b_i|0,1) - E(b_i|1,0)}$, we can therefore use $\psi = \frac{\alpha_1^\star + 1}{\alpha_1^\star}$ below to shorten notation.[1]

[1]Here we use $\alpha_1^\star$ to denote a pseudo-true value of a parameter since the extrapolation identifying assumption is not assumed to hold here, only the CRC model is assumed to hold throughout this section.

We can re-write equation (B.8) as:

$$\delta_1^N = E(a_i|1,0) - \alpha_1^\star(E(b_i|1,0) - \bar{\delta}) \tag{B.11}$$

Therefore defining $\alpha_0^\star = E(a_i|1,0) - \alpha_1^\star E(b_i|1,0)$ we can write that $\delta_1^N = \alpha_0^\star + \alpha_1^\star\bar{\delta}$.

Equation (B.9) can then be written:

$$\pi_{00}(E(a_i|0,0) - \alpha_0^\star - \alpha_1^\star\bar{\delta}) + \pi_{11}\frac{\alpha_1^\star}{\alpha_1^\star+1}(E(a_i+b_i|1,1) - \alpha_0^\star - (\alpha_1^\star+1)\bar{\delta})$$
$$+\pi_{01}\alpha_1^\star(E(b_i|0,1) - \bar{\delta}) + \pi_{10}\alpha_1^\star(E(b_i|1,0) - \bar{\delta}) = 0$$

where we define $\pi_{x_1x_2} = P(x_{i1} = x_1, x_{i2} = x_2)$ to shorten notation.

This yields:

$$\bar{\delta} = \pi_{00}\frac{E(a_i|0,0) - \alpha_0^\star}{\alpha_1^\star} + \pi_{11}\frac{E(a_i+b_i|1,1) - \alpha_0^\star}{1+\alpha_1^\star}$$
$$+ \pi_{01}E(b_i|0,1) + \pi_{10}E(b_i|1,0)$$

so that the ATE for the entire population is indeed obtained by the linear extrapolation represented in Figure 1 of the main text.[2]

B.2 Suri 2011

The notation used in Suri (2011) is almost identical to the notation used in Lemieux (1998) but Suri (2011) uses minimum distance estimation instead of GMM estimation. The only differences between the notation in Suri (2011) and the notation in Lemieux (1998) are that the parameter ϕ is used, which is mapped to the notation of Lemieux (1998) by $\phi = \psi - 1$, so that this new parameter is mapped to our notation by $\phi = \frac{1}{\alpha_1}$. The expected value of returns is also defined to be β in Suri (2011) rather than $\bar{\delta}$ in Lemieux (1998), so that this

[2]Similarly ATE for untreated stayers would be taken to be $ATE_{00}^\star = \frac{E(a_i|0,0)-\alpha_0^\star}{\alpha_1^\star}$ and ATE for treated stayers would be taken to be $\frac{E(a_i+b_i|1,1)-\alpha_0^\star}{1+\alpha_1^\star}$.

new parameter is mapped to our notation by $\beta = E(b_i)$.

The reduced form parameters used in Suri (2011) are obtained from the conditional expectations:

$$E(y_{i1}|x_{i1}, x_{i2}) = \gamma_{01} + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i1} x_{i2}$$
$$E(y_{i2}|x_{i1}, x_{i2}) = \gamma_{02} + \gamma_4 x_{i1} + \gamma_5 x_{i2} + \gamma_6 x_{i1} x_{i2}$$

which are obtained in this form without loss of generality since x_{it} is a binary random variable.

In addition to the parameter ϕ and ATE β, the structural parameters to be identified in Suri (2011) also comprise the parameters in the conditional expected value of θ_i conditional on treatment history:

$$E(\theta_i|x_{i1}, x_{i2}) = \lambda_0 + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \lambda_3 x_{i1} x_{i2} \tag{B.12}$$

The structural parameters to be identified in Suri (2011) are the parameters in the conditional expectation $E(\theta_i|x_{i1}, x_{i2})$, ϕ, and the ATE β (a total of six structural parameters). These structural parameters are estimated by minimum distance estimation from the link:

$$\gamma_1 = (1 + \phi)\lambda_1 + \beta + \phi\lambda_0 \tag{B.13}$$
$$\gamma_2 = \lambda_2 \tag{B.14}$$
$$\gamma_3 = (1 + \phi)\lambda_3 + \phi\lambda_2 \tag{B.15}$$
$$\gamma_4 = \lambda_1 \tag{B.16}$$
$$\gamma_5 = (1 + \phi)\lambda_2 + \beta + \phi\lambda_0 \tag{B.17}$$
$$\gamma_6 = (1 + \phi)\lambda_3 + \phi\lambda_1 \tag{B.18}$$
$$\lambda_0 = -\lambda_1\pi_1 - \lambda_2\pi_2 - \lambda_3\pi_{11} \tag{B.19}$$

where as before $\pi_{x_1x_2} = P(x_{i1} = x_1, x_{i2} = x_2)$ and where $\pi_1 = P(x_{i1} = 1)$ and $\pi_2 = P(x_{i2} = 1)$.

The first six equalities (B.13)-(B.18) follow from:

$$
\begin{aligned}
&E(y_{it}|x_{i1}, x_{i2}) \\
&= f_t + (1 + \phi x_{it})E(\theta_i|x_{i1}, x_{i2}) + \beta x_{it} \\
&= f_t + (1 + \phi x_{it})(\lambda_0 + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \lambda_3 x_{i1}x_{i2}) + \beta x_{it} \\
&= \begin{cases} f_1 + \lambda_0 + ((1+\phi)\lambda_1 + \beta + \phi\lambda_0)x_{i1} + \lambda_2 x_{i2} + ((1+\phi)\lambda_3 + \phi\lambda_2)x_{i1}x_{i2} & if\, t = 1 \\ f_2 + \lambda_0 + \lambda_1 x_{i1} + ((1+\phi)\lambda_2 + \beta + \phi\lambda_0)x_{i2} + ((1+\phi)\lambda_3 + \phi\lambda_1)x_{i1}x_{i2} & if\, t = 2 \end{cases}
\end{aligned}
$$

where the first equality follows from the model given by (B.4) with the notation used in Suri (2011).

The last equality (B.19) follows from $E(\theta_i) = 0$.

Under the CRC model we can rewrite the reduced form parameters as:

$$
\begin{aligned}
\gamma_1 &= E(b_i|1,0) + E(a_i|1,0) - E(a_i|0,0) \\
\gamma_2 &= E(a_i|0,1) - E(a_i|0,0) \\
\gamma_3 &= E(b_i|1,1) - E(b_i|1,0) + E(a_i|1,1) - E(a_i|1,0) + E(a_i|0,0) - E(a_i|0,1) \\
\gamma_4 &= E(a_i|1,0) - E(a_i|0,0) \\
\gamma_5 &= E(b_i|0,1) + E(a_i|0,1) - E(a_i|0,0) \\
\gamma_6 &= E(b_i|1,1) - E(b_i|0,1) + E(a_i|1,1) - E(a_i|1,0) + E(a_i|0,0) - E(a_i|0,1)
\end{aligned}
$$

Therefore under the CRC model, equations (B.14), (B.15), (B.16), and (B.18) or equations (B.13), (B.14), (B.16), and (B.17) both lead to $\frac{\gamma_6-\gamma_3}{\gamma_4-\gamma_2} = \frac{\gamma_1-\gamma_5}{\gamma_4-\gamma_2} - 1 = \frac{E(b_i|1,0)-E(b_i|0,1)}{E(a_i|1,0)-E(a_i|0,1)}$. Therefore under the CRC model the system of seven equations (B.13)-(B.19) is at most exactly identifying for the six structural parameters.

From (B.13), (B.14), (B.16), and (B.17) we find:

$$\begin{aligned}\phi &= \frac{\gamma_1 - \gamma_5}{\gamma_4 - \gamma_2} - 1 \\ &= \frac{E(b_i|1,0) - E(b_i|0,1)}{E(a_i|1,0) - E(a_i|0,1)}\end{aligned}$$

We also have:

$$\lambda_2 = \gamma_2 = E(a_i|0,1) - E(a_i|0,0) \tag{B.20}$$

and:

$$\lambda_1 = \gamma_4 = E(a_i|1,0) - E(a_i|0,0) \tag{B.21}$$

When $\phi = -1$, the system linking reduced form parameters to structural parameters does not identify λ_3, so that β, λ_0, λ_3 are not identified, so that the probability limits of the estimators for average returns for stayers do not exist. In Figure 1 in the main text, this corresponds to the case where the line $a + b = E(b_i|1,1) + E(a_i|1,1)$ and the extrapolation line going through $(E(a_i|1,0), E(b_i|1,0))$ and $(E(a_i|0,1), E(b_i|0,1))$ have the same slope, -1, so that they do not intersect at a unique point.

When $\phi \neq -1$, then:

$$\lambda_3 = \frac{\gamma_6 - \phi\lambda_1}{1 + \phi} \tag{B.22}$$

β is then given by:

$$\begin{aligned}\beta &= \gamma_1 - (1+\phi)\lambda_1 - \phi\lambda_0 \\ &= \gamma_1 - \gamma_4 - \phi(\lambda_1 + \lambda_0) \\ &= E(b_i|1,0) - \phi(\lambda_1 + \lambda_0)\end{aligned}$$

We can show from the above equalities that:

$$\begin{aligned}\lambda_1 + \lambda_0 = &(E(a_i|1,0) - E(a_i|0,0))(1 - \pi_1 - \pi_2 + \pi_{11}) \\ &+ \pi_2(E(a_i|1,0) - E(a_i|0,1)) \\ &+ \pi_{11}\frac{1}{1+\phi}(E(a_i + b_i|0,1) - E(a_i + b_i|1,1))\end{aligned}$$

The above implies:

$$\begin{aligned}\beta = &\pi_{10}E(b_i|1,0) + \pi_{01}E(b_i|0,1) \\ &+ \pi_{00}(E(a_i|0,0)\frac{E(b_i|1,0) - E(b_i|0,1)}{E(a_i|1,0) - E(a_i|0,1)} + \frac{E(b_i|0,1)E(a_i|1,0) - E(b_i|1,0)E(a_i|0,1)}{E(a_i|1,0) - E(a_i|0,1)} \\ &+ \pi_{11}(E(a_i + b_i|1,1)\frac{E(b_i|1,0) - E(b_i|0,1)}{E(a_i + b_i|1,0) - E(a_i + b_i|0,1)} \\ &+ \frac{E(b_i|0,1)E(a_i + b_i|1,0) - E(b_i|1,0)E(a_i + b_i|0,1)}{E(a_i + b_i|1,0) - E(a_i + b_i|0,1)})\end{aligned}$$

Using $\alpha_1^\star = \frac{E(a_i|1,0) - E(a_i|0,1)}{E(b_i|1,0) - E(b_i|0,1)}$ and $\alpha_0^\star = E(a_i|0,1) - \alpha_1^\star E(b_i|0,1)$ we can rewrite:

$$\begin{aligned}\beta = &\pi_{10}E(b_i|1,0) + \pi_{01}E(b_i|0,1) \\ &+ \pi_{00}\frac{E(a_i|0,0) - \alpha_0^\star}{\alpha_1^\star} \\ &+ \pi_{11}\frac{E(a_i + b_i|1,1) - \alpha_0^\star}{\alpha_1^\star + 1}\end{aligned}$$

Similarly for expected returns for the population of stayers, which in Suri (2011) are given by:

$$\begin{aligned}ATE_{00}^\star &= \beta + \phi E(\theta_i|0,0) = \beta + \phi\lambda_0 \\ ATE_{11}^\star &= \beta + \phi E(\theta_i|1,1) = \beta + \phi(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3)\end{aligned}$$

Under the CRC model, we have:

$$\begin{aligned} ATE^{\star}_{00} &= \gamma_1 - (1+\phi)\lambda_0 \\ &= E(b_i|1,0)\frac{E(a_i|0,0) - E(a_i|0,1)}{E(a_i|1,0) - E(a_i|0,1)} + E(b_i|0,1)\frac{E(a_i|1,0) - E(a_i|0,0)}{E(a_i|1,0) - E(a_i|0,1)} \end{aligned}$$

and:

$$\begin{aligned} ATE^{\star}_{11} &= \gamma_1 - (1+\phi)(\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3) \\ &= E(b_i|1,0)\frac{E(a_i+b_i|1,1) - E(a_i+b_i|0,1)}{E(a_i+b_i|1,0) - E(a_i+b_i|0,1)} \\ &+ E(b_i|0,1)\frac{E(a_i+b_i|1,0) - E(a_i+b_i|1,1)}{E(a_i+b_i|1,0) - E(a_i+b_i|0,1)} \end{aligned}$$

Using $\alpha^{\star}_1 = \frac{E(a_i|1,0)-E(a_i|0,1)}{E(b_i|1,0)-E(b_i|0,1)}$ and $\alpha^{\star}_0 = E(a_i|0,1) - \alpha_1 E(b_i|0,1)$ we can rewrite:

$$\begin{aligned} ATE^{\star}_{00} &= \frac{E(a_i|0,0) - \alpha^{\star}_0}{\alpha^{\star}_1} \\ ATE^{\star}_{11} &= \frac{E(a_i+b_i|1,1) - \alpha^{\star}_0}{\alpha^{\star}_1 + 1} \end{aligned}$$

Therefore we see that the method used in Suri (2011) corresponds to the extrapolation represented in Figure 1 of the main text.

C Comparison of the Extrapolation Identifying Assumption with Generalized Roy Models

In this section we briefly compare generalized Roy models with the set of conditions (2.12) and (2.13) considered in the main text as sufficient for the extrapolation identifying assumption (2.11) to hold. We take as an example the model outlined in p. 365-366 of Carneiro et al. (2003), abstracting from observed covariates or instrumental variables, which can be

written as:

$$\mathbf{y}(0) = \beta_{0,0} + \beta_{1,0}\boldsymbol{\theta} + \boldsymbol{\epsilon}_0, \qquad \mathbf{y}(1) = \beta_{0,1} + \beta_{1,1}\boldsymbol{\theta} + \boldsymbol{\epsilon}_1,$$
$$\mathbf{x} = 1[\beta_{0,s} + \beta_{1,s}\boldsymbol{\theta} + \boldsymbol{\epsilon}_s \geq 0], \qquad \{\boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_s\} \perp \boldsymbol{\theta}$$

where variables in bold denote random variables, $\mathbf{y}(1)$ and $\mathbf{y}(0)$ are potential outcomes with and without treatment, $\mathbf{x}$ is treatment status, $\boldsymbol{\theta}$ is an unobserved common factor, and $\boldsymbol{\epsilon}_0, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_s$ are unobserved shocks to outcomes and selection into treatment.

We can define $\mathbf{a} = \beta_{0,0}+\beta_{1,0}\boldsymbol{\theta}+\boldsymbol{\epsilon}_0$, $\mathbf{b} = \beta_{0,1}-\beta_{0,0}+(\beta_{1,1}-\beta_{1,0})\boldsymbol{\theta}$, so that if $\beta_{1,1}-\beta_{1,0} \neq 0$ we have $\boldsymbol{a} = \beta_{0,0} - \frac{\beta_{1,0}}{\beta_{1,1}-\beta_{1,0}}(\beta_{0,1} - \beta_{0,0}) + \frac{\beta_{1,0}}{\beta_{1,1}-\beta_{1,0}}\mathbf{b} + \boldsymbol{\epsilon}_0$ and (2.12) holds. With these definitions, (2.13) also holds by defining $\mathbf{c} = \boldsymbol{\epsilon}_s$ and from the assumption that $\boldsymbol{\epsilon}_s$ is independent of $\boldsymbol{\theta}$. With generalized Roy models, observing instrumental variables that satisfy exogeneity and relevance conditions, observing proxies for the unobserved common factor $\boldsymbol{\theta}$, or observing several independent measurements of $\boldsymbol{\theta}$, would yield identification of ATE (see also Cunha et al. (2005), Abbring and Heckman (2007)), while here the restrictions imposed by the CRC model (2.4) yield identification, as discussed in the main text.

D CRC Model with Time-Varying Treatment Effects

In this section we discuss CRC models with time varying treatment effects. For simplicity we consider the case where there are no additional control covariates here, so that the model is given by:

$$y_{it} = a_i + b_{it}x_{it} + f_t + u_{it}, \qquad E(u_{it}|X) = 0 \tag{D.1}$$

Without additional restrictions on treatment effects, b_{it}, identification of differences in time effects relies on average changes in outcomes for a cross-sectional observation across pairs of time periods when she was untreated only (whereas with the CRC model with time-

constant treatment effects considered in the main text, pairs of time periods with the same treatment status - both treated and untreated - can be used to identify changes in time effects):

$$f_t - f_s = E(y_{it} - y_{is} | x_{it} = 0, x_{is} = 0) \tag{D.2}$$

ATE for movers who are currently treated can then be identified by difference-in-differences comparisons:

$$E(b_{it} | x_{it} = 1, x_{is} = 0) = E(y_{it} - y_{is} - (f_t - f_s) | x_{it} = 1, x_{is} = 0) \tag{D.3}$$

As in the main text, we can apply the normalization $f_1 = 0$ to shorten notation. Then average baseline heterogeneity is identified for movers and untreated stayers, and average total heterogeneity at each time period is identified for treated movers and stayers:

$$\begin{aligned}
E(a_i | x_{it} = 1, x_{is} = 0) &= E(y_{is} - f_s | x_{it} = 1, x_{is} = 0) \\
E(a_i + b_{it} | x_{it} = 1, x_{is} = 0) &= E(y_{it} - f_t | x_{it} = 1, x_{is} = 0) \\
E(a_i | \max_{s=1,...,T} x_{is} = 0) &= E(y_{it} - f_t | \max_{s=1,...,T} x_{is} = 0) \\
E(a_i + b_{it} | \min_{s=1,...,T} x_{is} = 1) &= E(y_{it} - f_t | \min_{s=1,...,T} x_{is} = 1)
\end{aligned}$$

Under the CRC model with time varying random coefficients (D.1), ATE for movers who are currently untreated or for stayers are not identified. In order to extrapolate from the quantities identified by the CRC model (D.1) to ATE for stayers or for movers who are currently untreated, we can assume that a single term of unobserved heterogeneity determines baseline heterogeneity, treatment effects at each time period, and treatment status:

$$a_i = \beta_{0,a} + \beta_{1,a} e_i + \nu_{a,i}, \qquad E(\nu_{a,i} | e_i) = 0 \tag{D.4}$$

$$b_{it} = \beta_{0,t} + \beta_{1,t} e_i + \nu_{t,i}, \qquad E(\nu_{t,i} | e_i) = 0 \; \forall \, t \tag{D.5}$$

$$x_{it} = g(e_i, c_{it}) \; \forall \, t, \qquad \{c_{i1}, ..., c_{iT}\} \perp \{e_i, \nu_{a,i}, \nu_{t,i} \; \forall \, t\} \tag{D.6}$$

Conditions (D.4) and (D.5) impose that a particular characteristic e_i determines both baseline heterogeneity a_i and treatment effects b_{it}, but the effect of e_i on treatment effect may vary over time. Condition (D.6) imposes that treatment effects b_{it} themselves do not enter the determination of treatment status, rather that this determination is based on e_i only. Intuitively this last restriction most likely implies that the shocks $\nu_{t,i}$ to treatment effect b_{it} are unknown and unpredictable at the time of determination of treatment status, so that the main advantage of this extrapolation with time-varying treatment effects compared to the extrapolation discussed in the main text with time-constant treatment effects is that the effect of the one-dimensional term of unobserved heterogeneity e_i on treatment effect b_{it}, $\beta_{1,t}$, is allowed to vary over time.

Under (D.4)-(D.6), we obtain:

$$a_i = \alpha_{0,t} + \alpha_{1,t} b_{it} + \epsilon_{it}, \qquad E(\epsilon_{it}|x_{i1}, ..., x_{iT}) = 0 \tag{D.7}$$

where we define $\alpha_{0,t} = \beta_{0,a} - \frac{\beta_{0,t}}{\beta_{1,t}}$ and $\alpha_{1,t} = \frac{\beta_{1,a}}{\beta_{1,t}}$, which are well-defined if $\beta_{1,t} \neq 0$.

Under the CRC model with time-varying random coefficients (D.1) and the new extrapolation identifying assumption (D.7), the second step of our estimation procedure can still take the form of an instrumental variable regression of $\hat{a}_i$ on $\hat{b}_{it}$ using treatment status history $\{x_{i1}, ..., x_{iT}\}$ as instrumental variables, but the estimation sample is now restricted to include only movers who are treated at time t. Because of this restriction in the sample that can be used for the second-step estimation, one must observe at least three time periods to observe several groups of movers who are treated at time t and be able to implement this estimation procedure. For instance with $T = 2$, the only group of movers treated at time $t = 1$ has treatment status history profile $x = (1, 0)$. With $T = 3$, movers treated at time $t = 1$ correspond to treatment status history profiles $x \in \{(1, 0, 0), (1, 1, 0), (1, 0, 1)\}$.

ATE for untreated movers and stayers can then be identified from:

$$E(b_{it}|x) = \frac{E(a_i|x) - \alpha_{0,t}}{\alpha_{1,t}} = \frac{E(a_i + b_{it}|x) - \alpha_{0,t}}{1 + \alpha_{1,t}} \tag{D.8}$$

if $\alpha_{1,t} \notin \{-1, 0\}$.

In practice, since identification relies on smaller groups of movers, the implementation of this approach may lead to estimation results that are significantly more imprecise than the results obtained when treatment effects are assumed to be time constant. In addition, the conditions (D.4)-(D.6) that lead to the extrapolation identifying assumption (D.7) are restrictive, even though we can note that they encompass the case when treatment effects are constant over time since we could then define $e_i = b_i$ and $\nu_{t,i} = 0$.

In some applications, accounting for dynamic treatment effects, i.e. allowing for past treatment status to affect current treatment effect, is an important feature of the models used. This could be accommodated here by explicitly using an extrapolation identifying assumption that allows, for instance, the most recent treatment status to affect current treatment effect:

$$b_{it} = \beta_{0,t} + \beta_{1,t}e_i + \beta_2 x_{it-1} + \beta_3 x_{it-1}e_i + \nu_{t,i}, \qquad E(\nu_{t,i}|e_i) = 0 \; \forall \, t \tag{D.9}$$

so that the difference in treatment effect at the time of the first exposure to treatment and one time period later is given by $\beta_2 + \beta_3 e_i$.

Under (D.4), (D.6), and (D.9) we obtain the extrapolation identifying assumption:

$$a_i = \alpha_{0,t} + \alpha_{1,t}(x_{it-1})b_{i,t} + \alpha_{2,t}(x_{it-1}) \tag{D.10}$$

where $\alpha_{1,t}(x_{it-1}) = \frac{\beta_{1,a}}{\beta_{1,t}+\beta_3 x_{it-1}}$ and $\alpha_{2,t}(x_{it-1}) = -\frac{\beta_{1,a}\beta_2}{\beta_{1,t}+\beta_3 x_{it-1}}$.

The parameters in this extrapolation identifying assumption can be estimated by an instrumental variable regression by conditioning on both possible values of $x_{it-1} \in \{0, 1\}$,

which will generally require than one additional time period be observed. For instance, suppose one is interested in estimating ATE at time $t = 2$. With $T = 3$, the treatment history profiles corresponding to movers that are treated at time $t = 2$ are given by:

$$(x_1, x_2, x_3) = (0, 1, 1),(0, 1, 0), (1, 1, 0)$$

from which we see that the extrapolation identifying assumption (D.10) would not be identified for cross-sectional observations who were treated at time $t = 1$ (i.e. with $x_{i1} = 1$). With $T = 4$, the treatment history profiles corresponding to movers that are treated at time $t = 2$ and were also treated at time $t = 1$ are given by:

$$(x_1, x_2,x_3, x_4) = (1, 1, 0, 0), (1, 1, 0, 1), (1, 1, 1, 0) \tag{D.11}$$

so that the extrapolation identifying assumption (D.10) would be identified as long as these three groups of movers have different average treatment effects.

Given identification of the parameters entering (D.10), extrapolation to movers who are currently untreated or to stayers can be performed as before. For instance the ATE for untreated stayers at time $t$at the first exposure to treatment is given by:

$$ATE_{t,0} = \frac{E(a_1|\max_{s=1,...,T} x_{is} = 0) - \alpha_{0,t} - \alpha_{2,t}(0)}{\alpha_{1,t}(0)} \tag{D.12}$$

while the ATE for untreated stayers at time t after the first exposure to treatment is given by:

$$ATE_{t,1} = \frac{E(a_1|\max_{s=1,...,T} x_{is} = 0) - \alpha_{0,t} - \alpha_{2,t}(1)}{\alpha_{1,t}(1)} \tag{D.13}$$

There are alternative approaches that could be used for extrapolation with the CRC model with time-varying treatment effects that could be used in particular applications, but we leave this question for later work.

From the discussion above, we see that the data requirements are significantly greater

with the models discussed in this section than with the models discussed in the main text. Not only more time periods need to be observed for identification, but identification of time effects in the CRC model and of the parameters in the extrapolation identifying assumption is obtained from much narrower subgroups of cross-sectional observations than in the main text. In practice this will imply that estimation results will be significantly more noisy than using the methods developed in the main text, and will require researchers to have access to larger datasets than what is used in the empirical application of the main text. It would be interesting to consider the extensions sketched in this section in more details in future work in applications that are more amenable to these methods.

E Estimation and Inference with the Simple Extrapolation Identifying Assumption

In this section we discuss the details of estimation for models of the form:

$$y_{it} = a_i + b_i x_{it} + z_{it}\gamma + u_{it}, \qquad E(u_{it}|X_i, Z_i) = 0 \tag{E.1}$$

$$a_i = \alpha_0 + \alpha_1 b_i + \epsilon_i, \qquad E(\epsilon_i|X_i) = 0 \tag{E.2}$$

where $X_i = \begin{bmatrix} x_{i1} \\ ... \\ x_{iT} \end{bmatrix}$, $Z_i = \begin{bmatrix} z_{i1} \\ ... \\ z_{iT} \end{bmatrix}$, and z_{it} is a vector of control covariates.

Setting z_{it} to be a set of indicator variables for each time period other than the first time period, i.e. $z_{it} = [1[t = s]]^{s=2,...,T}$, yields the special case:

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

which is the CRC model considered in the main text, where only time effects are included as control covariates and where the normalization $f_1 = 0$ has been applied.

In general z_{it} is a vector of controls that can include random variables. For notational simplicity we consider the case where z_{it} is a scalar variable below, as all results extend in a straightforward way to multiple control covariates.

We observe cross-sectional observations $i = 1, ..., n$ over time periods $t = 1, ..., T$. Throughout this appendix, we use an asymptotic framework where n is large and T is small. We also assume that observations are cross-sectionally independent for simplicity, all results extend in a straightforward way to many independent clusters as in the empirical application.

Assumption 7. *Observations are cross-sectionally independent.*

Note that Assumption 7 is implied by Assumption 1 in the main text.

Stacking observations over time, we obtain:

$$Y_i = W_i \begin{bmatrix} a_i \\ b_i \end{bmatrix} + Z_i\gamma + U_i, \qquad E(U_i|W_i, Z_i) = 0 \tag{E.3}$$

where $Y_i = \begin{bmatrix} y_{i1} \\ ... \\ y_{iT} \end{bmatrix}$, $W_i = \begin{bmatrix} w_{i1} \\ ... \\ w_{iT} \end{bmatrix}$, $U_i = \begin{bmatrix} u_{i1} \\ ... \\ u_{iT} \end{bmatrix}$, and $w_{it} = [1, x_{it}]$.

The estimation method we discuss in this section is decomposed into two steps. The first step yields consistent estimates of the homogenous coefficients γ and noisy estimates of treatment effect b_i and baseline heterogeneity a_i for cross-sectional units that are movers, noisy estimates of baseline heterogeneity a_i for untreated stayers, noisy estimates of total heterogeneity $a_i + b_i$ for treated stayers. The second step yields consistent estimates of α_0, α_1, and noisy estimates of the values of $\{a_i, b_i\}_{i=1,...,n}$ that were missing from the first step, i.e. corresponding to untreated and treated stayers. We show that averaging the resulting noisy estimates of treatment effect b_i across the entire population or large groups of stayers or movers yields consistent estimators of ATE.

E.1 Step 1: High-dimensional regression

The first step of our estimation procedure regresses y_{it} on z_{it}, indicator variables for each cross-sectional observation, and indicator variables for each cross-sectional observation interacted with x_{it}. By the Frisch-Waugh theorem, the resulting estimates are given by:

$$\hat{\gamma} = \gamma + (\sum_{i=1}^{n} Z_i' M_{W_i} Z_i)^{-1} \sum_{i=1}^{n} Z_i' M_{W_i} U_i$$

$$\begin{bmatrix} \hat{a}_i \\ \hat{b}_i \end{bmatrix} = \begin{bmatrix} a_i \\ b_i \end{bmatrix} + (W_i' W_i)^{-} W_i' (U_i - Z_i(\hat{\gamma} - \gamma))$$

where $M_{W_i} = I_T - W_i(W_i'W_i)^{-}W_i'$ and $(W_i'W_i)^{-}$ is the generalized inverse obtained by omitting the interaction of the indicator variable corresponding to a particular cross-sectional observation with treatment status x_{it} when this cross-sectional observation is a stayer, i.e. when there is no variation in x_{it} over time across observations corresponding to this cross-sectional observation.

The next assumption imposes restrictions on moments of the data.

Assumption 8.

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

For constants C and $c > 0$:

b) $\frac{1}{n}\sum_{i=1}^{n} P(x_{it} = x_{is} \, \forall \, t, s) \geq c \; \forall \, n \geq C.$

c) $\frac{1}{n}\sum_{i=1}^{n} E(Z_i' M_{W_i} Z_i) \geq c \; \forall \, n \geq C.$

d) $\frac{1}{n} Var(\sum_{i=1,\ldots,n: x_{it} = x_{is} \, \forall \, t,s} Z_i' M_{W_i} U_i) \geq c \; \forall \, n \geq C.$

Assumption 8.a is standard and imposed in this form for simplicity. It could easily be relaxed to higher moments of the random variables a_i, b_i, z_{it}, u_{it} being uniformly bounded. Assumption 8.b imposes that we observe a non-vanishing share of stayers in the data. Assumptions 8.c and 8.d impose that there is variation in covariates z_{it} and transitory shocks u_{it} over time. For instance z_{it} and u_{it} may not be time constant.

Assumption 8.d requires that there be variation in covariates z_{it} and transitory shocks u_{it} over time among observations that are stayers instead of among all cross-sectional observations. This additional regularity condition is imposed for simplicity as it guarantees that the first-step estimator discussed in this section is not approximately linearly dependent of the part of the influence function of the second-step estimator which does not depend on the first-step estimator, which guarantees that the second-step estimator is at most $\sqrt{n}$-consistent, i.e. is not super consistent.

Note that Assumption 8 is implied by Assumptions 1 and 2 in the main text for the special case where $T = 2$ and $z_{it} = 1[t = 2]$.

Under the CRC model and Assumptions 7 and 8, the first step of our estimation procedure yields consistent estimates of the homogenous coefficients γ and noisy estimates of the two terms of unobserved heterogeneity a_i and b_i with estimation noise that can be decomposed into a part which vanishes as sample size increases and a part which is unrelated to sample size.

Proposition 4. *Under (E.1) and Assumptions 7 and 8, as $n \to \infty$ while T remains fixed:*

$$V_{n,\gamma,0}^{-\frac{1}{2}} A_{n,\gamma,0} \sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, I_K) \tag{E.4}$$

where $A_{n,\gamma,0} = \frac{1}{n}\sum_{i=1}^{n} E(Z_i' M_{W_i} Z_i)$, $V_{n,\gamma,0} = \frac{1}{n}\sum_{i=1}^{n} Var(Z_i' M_{W_i} U_i)$.

For $i = 1, ..., n$ such that $\exists\, t, s$ s.t. $x_{it} \neq x_{is}$, we have:

$$\begin{bmatrix} \hat{a}_i \\ \hat{b}_i \end{bmatrix} - \begin{bmatrix} a_i \\ b_i \end{bmatrix} = (W_i' W_i)^{-1} W_i' U_i - (W_i' W_i)^{-1} W_i' Z_i A_{n,\gamma,0}^{-1} \frac{1}{\sqrt{n}} \zeta_{ab,n} + e_{i,n}$$

where $\zeta_{ab,n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i' M_{W_i} U_i = O_p(1)$ and $max_{i=1,...,n}|e_{i,n}| = o_p(\frac{1}{\sqrt{n}})$.

For $i = 1, ..., n$ such that $x_{it} = 0\ \forall\, t$, we have:

$$\hat{a}_i - a_i = \frac{1}{T}\sum_{t=1}^{T} u_{it} - \frac{1}{T}\sum_{t=1}^{T} z_{it} A_{n,\gamma,0}^{-1} \frac{1}{\sqrt{n}} \zeta_{ab,n} + e_{i,n} \tag{E.5}$$

For $i = 1, ..., n$ such that $x_{it} = 1 \; \forall \, t$, we have:

$$\hat{a}_i + \hat{b}_i - a_i - b_i = \frac{1}{T} \sum_{t=1}^{T} u_{it} - \frac{1}{T} \sum_{t=1}^{T} z_{it} A_{n,\gamma,0}^{-1} \frac{1}{\sqrt{n}} \zeta_{ab,n} + e_{i,n} \tag{E.6}$$

Proposition 4 shows that the estimator for γ is $\sqrt{n}$-consistent and asymptotically normal. It also shows that the estimation noise of the heterogeneity terms a_i and b_i is decomposed into idiosyncratic noise that would arise even if γ were known and vanishing noise originating from the estimation of γ which is dominated by a term of order $\frac{1}{\sqrt{n}}$.

Note that Proposition 1 in the main text is obtained as a special case of Proposition 4.

E.2 Step 2: Instrumental Variable Regression and Extrapolation

In a second step, we consider estimation of the parameters α_0 and α_1 in the linear extrapolation to ATE for stayers by generalized method of moments (GMM) estimation of the coefficients in an instrumental variable regression using noisy estimates of a_i ($\hat{a}_i$) as the dependent variable, noisy estimates of b_i ($\hat{b}_i$) as the explanatory variable, and treatment history X_i as instrumental variables.

Let M_n be the subset of cross-sectional observations that are movers, i.e.

$$M_n = \{i = 1, ..., n : \; \exists \, t, s \; s.t. \; x_{it} \neq x_{is}\} \tag{E.7}$$

Define $\mathcal{S} \subseteq \{1, ..., T\}$ to be one of the largest subsets of time periods such that the variables $\{x_{it}\}_{t \in \mathcal{S}}$ are linearly independent among observations corresponding to movers in the data. When only two time periods are observed, i.e. $T = 2$, we have $\mathcal{S} = \{1\}$ or $\mathcal{S} = \{2\}$ since $x_{i1} = 1 - x_{i2}$ among movers. In general when $T > 2$ and if there are treated and untreated observations in all time periods, the entire vector of treatment status history X_i will be included in the list of instrumental variables, i.e. $\mathcal{S} = \{1, ..., T\}$. Define

$\tilde{X}_i = \begin{bmatrix} 1 \\ [x_{it}]_{t\in\mathcal{S}} \end{bmatrix}$ to be the elements of the vector of treatment status history that are linearly independent among movers augmented with a constant.

The estimator for α_0 and α_1 in the extrapolation identifying assumption (E.2) that we consider in this section is:

$$\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} = (B_n' \Sigma_n^{-1} B_n)^{-1} B_n' \Sigma_n^{-1} \frac{1}{n} \sum_{i \in M_n} \tilde{X}_i \hat{a}_i$$

$$B_n = \frac{1}{n} \sum_{i \in M_n} \tilde{X}_i [1, \hat{b}_i]$$

$$\Sigma_n = \frac{1}{n} \sum_{i \in M_n} \tilde{\epsilon}_i^2 \tilde{X}_i \tilde{X}_i'$$

where $\tilde{\epsilon}_i$ are first-stage residuals obtained by a two-stage least squares regression of $\hat{a}_i$ on $\hat{b}_i$ using $\tilde{X}_i$ as instrumental variables.[3]

As in the main text, given estimates of the parameters α_0 and α_1, we obtain estimates of ATE for stayers, $ATE_{S,0} = \frac{\sum_{i=1,\ldots,n} E(1[x_{it}=0 \,\forall\, t] b_i)}{\sum_{i=1,\ldots,n} P(x_{it}=0 \,\forall\, t)}$ and $ATE_{S,1} = \frac{\sum_{i=1,\ldots,n} E(1[x_{it}=1 \,\forall\, t] b_i)}{\sum_{i=1,\ldots,n} P(x_{it}=1 \,\forall\, t)}$, by a

[3]Note that in this section we could have considered a two-stage least squares regression only instead of GMM estimation, i.e. we could have chosen $\Sigma_n = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i'$. This would also lead to a consistent and asymptotically normal estimator. We consider GMM estimation for a potential efficiency gain because of the heteroscedasticity in $\hat{a}_i - \alpha_0 - \alpha_1 \hat{b}_i$ conditional on X_i that is likely to exist because of measurement error. Indeed heteroscedasticity is likely to appear in the non-vanishing part of the estimation noise in the estimates $\hat{a}_i$ and $\hat{b}_i$: Even if $Var(\epsilon_i | X_i) = \sigma_\epsilon^2$, $Var(U_i | X_i) = \sigma_u^2 I_T$, and $Cov(\epsilon_i, U_i | X_i) = 0$, we have:

$$\begin{aligned} Var(\epsilon_i + [1, -\alpha_1](W_i' W_i)^{-1} W_i' U_i | X_i) &= \sigma_\epsilon^2 + \sigma_u^2 [1, -\alpha_1](W_i' W_i)^{-1} \begin{bmatrix} 1 \\ -\alpha_1 \end{bmatrix} \\ &\neq Var(\epsilon_i + [1, -\alpha_1](W_i' W_i)^{-1} W_i' U_i | i \in M_n) \end{aligned}$$

in general.

In addition GMM estimation extends easily to estimating and performing statistical inference on additional parameters such as ATE among different subpopulations, to accounting for cross-sectional dependence when computing the weighting matrix, and to accommodating unbalanced panel data originating from missing data. We encounter these three issues in our empirical application.

Note also that one could use interactions between elements of X_i to obtain additional valid moment functions. We only consider moment functions obtained by using linear terms for simplicity here and to avoid the proliferation of moment conditions. See for instance Newey and Windmeijer (2009) for a discussion of issues that arise with GMM estimation and many moment conditions.

simple plug-in using the derivations in the main text:

$$\hat{ATE}_{S,0} = \frac{\bar{a}_{S,0} - \hat{\alpha}_0}{\hat{\alpha}_1}, \qquad \hat{ATE}_{S,1} = \frac{\bar{a}_{S,1} - \hat{\alpha}_0}{1 + \hat{\alpha}_1}$$
$$\bar{a}_{S,0} = \frac{\sum_{i:x_{it}=0\,\forall\, t} \hat{a}_i}{|\{i : x_{it} = 0\,\forall\, t\}|} \qquad \bar{a}_{S,1} = \frac{\sum_{i:x_{it}=1\,\forall\, t} \hat{a}_i}{|\{i : x_{it} = 1\,\forall\, t\}|}$$

Define

$$B_{n,0} = \frac{1}{n} \sum_{i \in M_n} E(\tilde{X}_i[1, b_i])$$
$$r_i = \epsilon_i + [1, -\alpha_1](W_i^{'} W_i)^{-1} W_i^{'} U_i$$
$$\Sigma_{n,0} = \frac{1}{n} \sum_{i \in M_n} E(r_i^2 \tilde{X}_i \tilde{X}_i^{'})$$

Define $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ to be the two-stage least squares regression estimators of α_0 and α_1 obtained by regressing $\hat{a}_i$ on a constant and $\hat{b}_i$ using $\tilde{X}_i$ as instrumental variables.

Define λ_{min} to be the minimum eigenvalue of a matrix.

Assumption 9. *For constants C and $c > 0$:*

a) $\lambda_{min}(\Sigma_{n,0}) \geq c$, $\forall\, n \geq C$.

b) $\lambda_{min}(B_{n,0}^{'} \Sigma_{n,0}^{-1} B_{n,0}) \geq c$, $\forall\, n \geq C$.

c) $\tilde{\alpha}_0 \xrightarrow{p} \alpha_0$ *and* $\tilde{\alpha}_1 \xrightarrow{p} \alpha_1$ *as* $n \to \infty$ *while* T *remains fixed.*

d) $\frac{1}{n} Var(\sum_{i=1,...,n:x_{it}=0\,\forall\, t} (a_i + \frac{1}{T} \sum_{t=1}^{T} u_{it}) | \sum_{i=1}^{n} Z_i^{'} M_{W_i} U_i) \geq c$ *and* $\frac{1}{n} Var(\sum_{i=1,...,n:x_{it}=1\,\forall\, t} (a_i + b_i + \frac{1}{T} \sum_{t=1}^{T} u_{it}) | \sum_{i=1}^{n} Z_i^{'} M_{W_i} U_i) \geq c$ *a.s.,* $\forall\, n \geq C$.

Assumption 9.a requires that there be variation in X_i among observations that correspond to movers. For instance with two time periods, i.e. $T = 2$, Assumption 9.a would be obtained by imposing i) $Var(v_i | X_i) \geq c > 0\ \forall\, i = 1, ..., n,\ \forall\, n$ and ii) $0 < c \leq P(x_{i1} = 0, x_{i2} = 1) \leq 1 - c$ and $0 < c \leq P(x_{i1} = 1, x_{i2} = 0) \leq 1 - c\ \forall\, i = 1, ..., n,\ \forall\, n$. The first condition is standard and requires that the error terms of the CRC model u_{it} and of the extrapolation identifying assumption have positive variance, so that the resulting model is not degenerate.

The second condition guarantees that two different profiles of movers are represented by non-vanishing fractions of the data (in large samples), so that there is variation in x_{i1} (or x_{i2}) among observations that are movers.

Assumption 9.b requires that the variation in X_i among observations that correspond to movers be predictive of treatment effect b_i. For instance with two time periods Assumption 9.b would be obtained by $|\frac{\sum_{i=1,\ldots,n} E(1[(x_{i1},x_{i2})=(0,1)]b_i)}{\sum_{i=1,\ldots,n} P((x_{i1},x_{i2})=(0,1))} - \frac{\sum_{i=1,\ldots,n} E(1[(x_{i1},x_{i2})=(1,0)]b_i)}{\sum_{i=1,\ldots,n} P((x_{i1},x_{i2})=(1,0))}| \geq c > 0$ $\forall\, n \geq C$.

Assumption 9.c requires that the two-stage least squares estimators of α_0 and α_1 be consistent. This is imposed for convenience only since convergence in probability of the two-stage least squares estimators of α_0 and α_1 can be derived from primitive conditions as in the proof of Proposition 5 below.

Assumption 9.d is a regularity condition which guarantees that the estimator for ATE is at most $\sqrt{n}$-consistent, i.e. is not super consistent. It requires that there be no approximately exact dependence between the terms of unobserved heterogeneity a_i and $a_i + b_i$ and the error term of the CRC model u_{it}. This condition for the case $x_{it} = 0\ \forall\, t$ would be obtained if we assume that i) a_i is independent of $\{u_{it}\}_{t=1,\ldots,T}$ conditional on Z_i and W_i and that ii) $Var(\frac{1}{T}\sum_{t=1}^{T} u_{it} | \{u_{is} - \frac{1}{T}\sum_{t=1}^{T} u_{it}\}_{s=1,\ldots,T}, Z_i, x_{it} = 0\, \forall\, t) \geq c$. These are both natural conditions requiring that the idiosyncratic shocks to outcomes u_{it} be independent of baseline heterogeneity a_i and that the error term in the CRC model u_{it} not be degenerate. Similarly this condition for the case $x_{it} = 1\ \forall\, t$ would be obtained if we assume that i) $a_i + b_i$ is independent of $\{u_{it}\}_{t=1,\ldots,T}$ conditional on Z_i and W_i and that ii) $Var(\frac{1}{T}\sum_{t=1}^{T} u_{it} | \{u_{is} - \frac{1}{T}\sum_{t=1}^{T} u_{it}\}_{s=1,\ldots,T}, Z_i, x_{it} = 1\, \forall\, t) \geq c$.

Note that Assumption 9.a, 9.b, and 9.d are implied by Assumptions 1-3 in the main text and Assumption 9.c is irrelevant with two time periods since the moment conditions $E(\tilde{X}_i(\hat{a}_i - \alpha_0 - \alpha_1 \hat{b}_i)) \simeq 0$ are exactly identifying for the parameters α_0 and α_1 in this case.

To state Proposition , we define deterministic matrices which will determine the asymp-

totic distribution of our second-step estimators $\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix}$:

$$V_{n,\alpha\gamma,0} = \frac{1}{n}\sum_{i\in M_n} E(r_i \tilde{X}_i U_i^{'} M_{W_i} Z_i) \qquad C_{n,0} = \frac{1}{n}\sum_{i\in M_n} E(\tilde{X}_i[1,-\alpha_1](W_i^{'}W_i)^{-1}W_i^{'}Z_i)$$

$$A_{n,1,0} = (B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1} \qquad A_{n,2,0} = (B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}C_{n,0}A_{n,\gamma,0}^{-1}$$

$$\Omega_{n,0} = \begin{bmatrix} \Sigma_{n,0} & V_{n,\alpha\gamma,0} \\ V_{n,\alpha\gamma,0}^{'} & V_{n,\gamma,0} \end{bmatrix} \qquad V_{n,0} = [A_{n,1,0}, A_{n,2,0}]\Omega_{n,0}\begin{bmatrix} A_{n,1,0}^{'} \\ A_{n,2,0}^{'} \end{bmatrix}$$

We define deterministic matrices which will determine the asymptotic distribution of our second-step estimators $A\hat{T}E_{S,0}$:

$$\tilde{a}_i = a_i + \frac{1}{T}\sum_{t=1}^{T} u_{it} - E(a_i),$$

$$A_{n,ATE_{S,0},\gamma,0} = -\frac{1}{\alpha_1}\frac{\frac{1}{n}\sum_{i=1,\ldots,n} E(1[x_{it}=0\,\forall\, t]\frac{1}{T}\sum_{t=1}^{T} z_{it})}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=0\,\forall\, t)}A_{n,\gamma,0}^{-1},$$

$$A_{n,ATE_{S,0},\alpha,\gamma,0} = [\frac{1}{\alpha_1}, \frac{ATE_{S,0}}{\alpha_1}](B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}C_{n,0}A_{n,\gamma,0}^{-1}$$

$$A_{n,ATE_{S,0},\alpha,r,0} = -[\frac{1}{\alpha_1}, \frac{ATE_{S,0}}{\alpha_1}](B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}$$

$$A_{n,ATE_{S,0},0} = [\frac{1}{\alpha_1}, A_{n,ATE_{S,0},\gamma,0} + A_{n,ATE_{S,0},\alpha,\gamma,0}, A_{n,ATE_{S,0},\alpha,r,0}]$$

$$\Omega_{ATE_{S,0},0} = Var(\sqrt{n}\begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=0\,\forall t]\tilde{a}_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=0\,\forall\, t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i^{'}M_{W_i}U_i \\ \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i \end{bmatrix})$$

$$V_{n,ATE_{s,0},0} = A_{n,ATE_{S,0},0}\Omega_{ATE_{S,0},0}A_{n,ATE_{S,0},0}^{'}$$

We define deterministic matrices which will determine the asymptotic distribution of our

second-step estimators $\hat{ATE}_{S,1}$:

$$
\begin{aligned}
a \tilde{+} b_i &= a_i + b_i + \frac{1}{T}\sum_{t=1}^{T} u_{it} - E(a_i + b_i) \\
A_{n,ATE_{S,1},\gamma,0} &= -\frac{1}{1+\alpha_1}\frac{\frac{1}{n}\sum_{i=1,\ldots,n} E(1[x_{it}=1\,\forall\, t]\frac{1}{T}\sum_{t=1}^{T} z_{it})}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=1\,\forall\, t)} A_{n,\gamma,0}^{-1}, \\
A_{n,ATE_{S,1},\alpha,\gamma,0} &= [\frac{1}{1+\alpha_1}, \frac{ATE_{S,1}}{1+\alpha_1}](B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}C_{n,0}A_{n,\gamma,0}^{-1} \\
A_{n,ATE_{S,1},\alpha,r,0} &= -[\frac{1}{1+\alpha_1}, \frac{ATE_{S,1}}{1+\alpha_1}](B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1} \\
A_{n,ATE_{S,1},0} &= [\frac{1}{1+\alpha_1}, A_{n,ATE_{S,1},\gamma,0} + A_{n,ATE_{S,1},\alpha,\gamma,0}, A_{n,ATE_{S,1},\alpha,r,0}] \\
\Omega_{ATE_{S,1},0} &= Var(\sqrt{n}\begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=1\,\forall t]a\tilde{+}b_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=1\,\forall\, t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i^{'}M_{W_i}U_i \\ \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i \end{bmatrix}) \\
V_{n,ATE_{s,1},0} &= A_{n,ATE_{S,1},0}\Omega_{ATE_{S,1},0}A_{n,ATE_{S,1},0}^{'}
\end{aligned}
$$

Proposition 5. *Under (E.1), (E.2), and Assumptions 7-9, as $n \to \infty$ while T remains fixed:*

$$
\begin{aligned}
\sqrt{n}(\begin{bmatrix}\hat{\alpha}_0 \\ \hat{\alpha}_1\end{bmatrix} - \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix}) &= (B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}\frac{1}{\sqrt{n}}\sum_{i\in M_n}\tilde{X}_i r_i \\
&\quad - (B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}C_{n,0}A_{n,\gamma,0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i^{'}M_{W_i}U_i + o_p(1)
\end{aligned}
$$

and:

$$
V_{n,0}^{-\frac{1}{2}}\sqrt{n}(\begin{bmatrix}\hat{\alpha}_0 \\ \hat{\alpha}_1\end{bmatrix} - \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix}) \xrightarrow{d} N(0, I_2) \tag{E.8}
$$

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

$$\sqrt{n}(A\hat{T}E_{S,0} - ATE_{S,0}) = A_{n,ATE_{s,0},0}\sqrt{n}\begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=0\,\forall t]\tilde{a}_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=0\,\forall\, t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i r_i \end{bmatrix} + o_p(1)$$

$$\sqrt{n}(A\hat{T}E_{S,1} - ATE_{S,1}) = A_{n,ATE_{s,1},0}\sqrt{n}\begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=1\,\forall t]\tilde{a}_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=1\,\forall\, t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i r_i \end{bmatrix} + o_p(1)$$

and:

$$V_{n,ATE_{s,0},0}^{-\frac{1}{2}}\sqrt{n}(A\hat{T}E_{S,0} - ATE_{S,0}) \xrightarrow{d} N(0,1)$$
$$V_{n,ATE_{s,1},0}^{-\frac{1}{2}}\sqrt{n}(A\hat{T}E_{S,1} - ATE_{S,1}) \xrightarrow{d} N(0,1)$$

Note that Proposition 2 in the main text is obtained as a special case of Proposition 5. Proposition 5 shows that the estimators $\begin{bmatrix}\hat{\alpha}_0 \\ \hat{\alpha}_1\end{bmatrix}$, $A\hat{T}E_{S,1}$, and $A\hat{T}E_{S,0}$ adopt a linear influence function asymptotic representation, so that inference by cluster bootstrap, with clusters given by cross-sectional units, would be asymptotically valid (see e.g. Mammen (1992)). Alternatively, one can use the analytical formula for asymptotic variance to obtain consistent estimated variance-covariance matrix using cluster robust standard errors for two-step estimation, with clusters given by cross-sectional units.

For simplicity the next proposition shows that these standard errors are consistent and lead to asymptotically valid inference for the estimator $\begin{bmatrix}\hat{\alpha}_0 \\ \hat{\alpha}_1\end{bmatrix}$ only, since the same result can be obtained for the estimators $A\hat{T}E_{S,1}$ and $A\hat{T}E_{S,0}$.

Proposition 6. *Define* $C_n = \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i[1, -\hat{\alpha}_1](W_i'W_i)^{-1}W_i'Z_i$*,* $\hat{r}_i = \hat{a}_i - \hat{\alpha}_0 - \hat{\alpha}_1\hat{b}_i$*, and* $\hat{U}_i = Y_i - Z_i\hat{\gamma}$*. Define* $\Sigma_n = \frac{1}{n}\sum_{i\in M_n} \hat{r}_i^2\tilde{X}_i\tilde{X}_i'$*,* $\Sigma_{n,\gamma} = \frac{1}{n}\sum_{i=1,\ldots,n} Z_i' M_{W_i}\hat{U}_i\hat{U}_i' M_{W_i} Z_i$*,* $\Sigma_{n,\alpha\gamma} =$

$\frac{1}{n}\sum_{i\in M_n}\hat{r}_i\tilde{X}_i\hat{U}_i^{'}M_{W_i}Z_i$. *Define* $A_{n,1}=(B_n^{'}\Sigma_n^{-1}B_n)^{-1}B_n^{'}\Sigma_n^{-1}$, $A_{n,2}=(B_n^{'}\Sigma_n^{-1}B_n)^{-1}B_n^{'}\Sigma_n^{-1}C_nA_n^{-1}$, $\Omega_n=\begin{bmatrix}\Sigma_n & \Sigma_{n,\alpha\gamma}\\ \Sigma_{n,\alpha\gamma}^{'} & \Sigma_{n,\gamma}\end{bmatrix}$, $V_n=[A_{n,1},A_{n,2}]\Omega_n\begin{bmatrix}A_{n,1}^{'}\\ A_{n,2}^{'}\end{bmatrix}$.

Under (E.1), (E.2), and Assumptions 7-9, as $n\to\infty$ *while* T *remains fixed:*

$$V_n^{-\frac{1}{2}}\sqrt{n}(\begin{bmatrix}\hat{\alpha}_0\\ \hat{\alpha}_1\end{bmatrix}-\begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix})\xrightarrow{d}N(0,I_2) \tag{E.9}$$

Since the matrices used in Proposition 6 are stored by standard statistical software, variance-covariance matrices using the formula given by Proposition 6 are straightforward to compute.

Additionally, note that all estimators above can be computed as the solution to exactly identifying moment conditions:

$$\begin{aligned}\sum_{i=1}^{n}Z_i^{'}M_{W_i}(Y_i-Z_i\hat{\gamma})&=0\\ B_n^{'}\Sigma_n^{-1}\sum_{i\in M_n}\tilde{X}_i([1,-\hat{\alpha}_1](W_i^{'}W_i)^{-1}W_i^{'}(Y_i-Z_i\hat{\gamma})-\hat{\alpha}_0)&=0\\ \sum_{i:x_{it}=0\,\forall\,t}(\frac{\frac{1}{T}\sum_{t=1}^{T}(y_{it}-z_{it}\hat{\gamma})-\hat{\alpha}_0}{\hat{\alpha}_1}-A\hat{T}E_{S,0})&=0\\ \sum_{i:x_{it}=1\,\forall\,t}(\frac{\frac{1}{T}\sum_{t=1}^{T}(y_{it}-z_{it}\hat{\gamma})-\hat{\alpha}_0}{1+\hat{\alpha}_1}-A\hat{T}E_{S,1})&=0\end{aligned}$$

so that, instead of using the formulae from Proposition 6 to compute standard errors, analytical standard errors can also be obtained directly using any command capable of numerical differentiation in standard statistical software.

E.3 Testing the Validity of the Extrapolation

E.3.1 Lack of testable implications with two time periods

As discussed in the main text, under the CRC model (**??**) and with two time periods, the extrapolation identifying assumption (**??**) is equivalent to introducing identities for four parameters that were left unrestricted by the CRC model, so that the extrapolation identifying assumption does not contain testable implications under the CRC model.

We can see this directly by considering the CRC model:

$$y_{it} = a_i + b_i x_{it} + f_t + u_{it}, \qquad E(u_{it}|x_{i1}, x_{i2}) = 0$$

and defining $\alpha_1^\star = \frac{E(a_i|0,1)-E(a_i|1,0)}{E(b_i|0,1)-E(b_i|1,0)}$ and $\alpha_0^\star = E(a_i|0,1) - \alpha_1^\star E(b_i|0,1)$ if $E(b_i|0,1) \neq E(b_i|1,0)$.

For i such that $x_{i1} \neq x_{i2}$, define $\tilde{a}_i = a_i$ and $\tilde{b}_i = b_i$. For i such that $x_{i1} = 0$ and $x_{i2} = 0$, define $\tilde{b}_i = \frac{a_i - \alpha_0^\star}{\alpha_1^\star}$ if $\alpha_1^\star \neq 0$. For i such that $x_{i1} = 1$ and $x_{i2} = 1$, define $\tilde{b}_i = \frac{a_i + b_i - \alpha_0^\star}{1+\alpha_1^\star}$ and $\tilde{a}_i = \alpha_0^\star + \alpha_1^\star \tilde{b}_i$ if $\alpha_1^\star \neq -1$.

Then we can write:

$$y_{it} = \tilde{a}_i + \tilde{b}_i x_{it} + u_{it}, \qquad E(u_{it}|x_{i1}, x_{i2}) = 0$$

and

$$\tilde{\alpha}_i = \alpha_0^\star + \alpha_1^\star \tilde{b}_i + \tilde{\xi}_i, \qquad E(\tilde{\xi}_i|x_{i1}, x_{i2}) = 0$$

since $\tilde{\xi}_i = 0$ when $x_{i1} = x_{i2}$ and $E(\tilde{\xi}_i|x_{i1}, x_{i2}) = 0$ when $x_{i1} \neq x_{i2}$.

E.3.2 Testing with three or more time periods

The extrapolation identifying assumption (E.2) implies that

$$E(a_i - \alpha_0 - \alpha_1 b_i | X_i) = 0 \tag{E.10}$$

Here we propose testing an implication of this assumption, namely $LP(a_i - \alpha_0 - \alpha_1 b_i | X_i, i \in M_n) = 0$, where LP is the linear projection operator, and recall that M_n is the subset of cross-sectional observations that are movers. This condition is equivalently written:

$$E(a_i - \alpha_0 - \alpha_1 b_i | i \in M_n) = 0 \tag{E.11}$$

$$E(x_{it}(a_i - \alpha_0 - \alpha_1 b_i) | i \in M_n) = 0 \; \forall \, t \in \mathcal{S} \tag{E.12}$$

where recall that $\mathcal{S}$ is one of the largest subsets of time periods such that the variables $\{x_{it}\}_{t \in \mathcal{S}}$ are linearly independent among observations corresponding to movers in the data. When only two time periods are observed, i.e. $T = 2$, we have $\mathcal{S} = \{1\}$ or $\mathcal{S} = \{2\}$, while in general $\mathcal{S} = \{1, ..., T\}$ when $T \geq 3$.

With a very large number of cross-sectional observations, testing (E.10) directly would generally yield a test with larger power, similarly as in the previous section where using interactions of the elements of X_i would lead to more moment conditions and to a more efficient estimator in general when a large number of cross-sectional observations is available. However with more modest sample sizes, it is possible that some values of X_i correspond to relatively few cross-sectional observations, leading to a "small cell" problem, i.e. it is possible that $E(a_i - \alpha_0 - \alpha_1 b_i | X_i)$ can only be estimated imprecisely for some values of X_i. This could lead to size distortions in small samples. We propose a more parsimonious test based on (E.11) and (E.12) instead.

This test is straightforward given the discussion in the previous section as long as three or more time periods are observed. We simply add $|\mathcal{S}| + 1$ exactly identified moment conditions:

$$E(1[i \in M_n]([1, -\alpha_1](W_i' W_i)^{-1} W_i'(Y_i - Z_i \gamma) - \alpha_0) - \eta_0) = 0$$

$$E(1[i \in M_n] x_{it}([1, -\alpha_1](W_i' W_i)^{-1} W_i'(Y_i - Z_i \gamma) - \alpha_0) - \eta_t) = 0 \, \forall \, t \in \mathcal{S}$$

and test the null hypothesis $H_0 : \eta_0 = 0, \eta_t = 0 \, \forall \, t \in \mathcal{S}$ using a Wald test with critical values from a chi-squared distribution with $|\mathcal{S}| - 1$ degrees of freedom. Note that when $T = 2$,

$|\mathcal{S}| = 1$ so that this test cannot be performed, as discussed in the main text and above.

Estimated variances for this Wald test can be obtained by cluster bootstrap, or using analytical formulae as in Proposition 6, or by solving for exactly identifying moment conditions using any command capable of numerical differentiation in standard statistical software.

This is an over-identification test similar to the Sargan-Hansen J-test discussed in Hansen (1982) except that it accounts for the first-step estimation of the coefficients γ.

F Estimation and Inference with the Generalized Extrapolation Identifying Assumption

In this section we discuss how to extend the estimation and testing methods discussed in the previous section to the use of a generalized extrapolation where cost shifters shared by all observations with the same value of indexing variable v_i can be correlated with productivity (a_i, b_i). The model considered in this section is given by:

$$y_{it} = a_i + b_i x_{it} + z_{it}\gamma + u_{it}, \qquad E(u_{it}|\{X_j, Z_j\}_{j:v_j=v_i}) = 0 \tag{F.1}$$

$$a_i = e_{v_i} + \alpha_1 b_i + \epsilon_i, \qquad E(\epsilon_i|\{X_j\}_{j:v_j=v_i}) = 0 \tag{F.2}$$

where all variables are defined as in the previous section, and v_i is a deterministic discrete indexing variable.[4] In our empirical example v_i indexes farmer i's village.

The same first-step estimator of γ is used as in the previous section. As before, it also leads to noisy estimates of baseline heterogeneity and treatment effects, $\hat{a}_i$ and $\hat{b}_i$, for movers, noisy estimates of baseline heterogeneity for untreated stayers, and noisy estimates of total heterogeneity for treated stayers.

[4]Note that here the exogeneity of covariates X_i and Z_i in the CRC model (F.1) has been strengthened to be strict across all observations with the same value of the indexing variable v_i. This is because the assumption of independence is relaxed below so that observations are only assumed to be independent across different values of the indexing variable v_i instead of being independent cross-sectionally. If the assumption of cross-sectional independence held, the assumptions of exogeneity in (F.1) could be relaxed to $E(u_{it}|X_i, Z_i) = 0$ as before.

The coefficient α_1 is then estimated by GMM estimation after demeaning the noisy estimates of a_i and b_i obtained for movers in the first stage within the groups defined by each value of the indexing variable v_i:

$$\ddot{a}_i = \hat{a}_i - \bar{a}_i, \qquad \bar{a}_i = \frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} \hat{a}_j$$

$$\ddot{b}_i = \hat{b}_i - \bar{b}_i, \qquad \bar{b}_i = \frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} \hat{b}_j$$

where $n_v = |\{i \in M_n : v_i = v\}|$. In our empirical example, this amounts to demeaning among movers within each village.

As before, define $\mathcal{S}$ to be one of the largest sets of time periods such that $[x_{it}]_{t \in \mathcal{S}}$ is a set of linearly independent variables in the data among observations that correspond to movers, in order to accommodate the case where $T = 2$. Redefine $\tilde{X}_i = [x_{it}]_{t \in \mathcal{S}}$ to be as in the previous section but without the constant. Then the estimator of α_1 in this section is given by:

$$\hat{\alpha}_1 = (\ddot{B}_n' \ddot{\Sigma}_n^{-1} \ddot{B}_n)^{-1} \ddot{B}_n' \ddot{\Sigma}_n^{-1} \frac{1}{n} \sum_{i \in M_n} \tilde{X}_i \ddot{a}_i$$

$$\ddot{B}_n = \frac{1}{n} \sum_{i \in M_n} \tilde{X}_i \ddot{b}_i$$

$$\ddot{\Sigma}_n = \frac{1}{n} \sum_{i \in M_n, j \in M_n : v_j = v_i} \ddot{\epsilon}_i \ddot{\epsilon}_j \tilde{X}_i \tilde{X}_j'$$

$$\ddot{\epsilon}_i = \ddot{a}_i - \tilde{\alpha}_1 \ddot{b}_i$$

and $\tilde{\alpha}_1$ is the first-step estimator of α_1 obtained by a fixed effects two-stage least squares regression of $\hat{a}_i$ on $\hat{b}_i$ using $\tilde{X}_i$ as instrumental variable and with fixed effects indexed by v_i.

Given this estimator of α_1, we obtain a noisy estimator of the "fixed effect" term e_v in

the generalized extrapolation assumption (F.2):

$$\hat{e}_v = \frac{1}{n_v} \sum_{i \in M_n : v_i = v} (\hat{a}_i - \hat{\alpha}_1 \hat{b}_i) \tag{F.3}$$

Estimators of ATE for stayers are then defined in this section by:

$$\hat{ATE}_{S,0} = \frac{1}{\hat{\alpha}_1} \frac{\sum_{i:x_{it}=0 \, \forall \, t} (\hat{a}_i - \hat{e}_{v_i})}{|\{i : x_{it} = 0 \, \forall \, t\}|}, \qquad \hat{ATE}_{S,1} = \frac{1}{1+\hat{\alpha}_1} \frac{\sum_{i:x_{it}=1 \, \forall \, t} (\hat{a}_i + \hat{b}_i - \hat{e}_{v_i})}{|\{i : x_{it} = 1 \, \forall \, t\}|}$$

Here we consider the case where there are few cross-sectional observations per value of the indexing variable v_i, so that the indexing variable v_i takes many values. As discussed in the main text, this corresponds to the data structure of our application where few farmers live in each village. Considering the case where v_i takes few values and where there are many cross-sectional observations per value of v_i is straightforward with cross-sectional independence or limited forms of cross-sectional dependence.

With v_i taking many values, Assumption 1 of cross-sectional independence can be relaxed to independence across values of v_i. 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.

Assumption 10. *Observations are independent if they do not share the same value of the indexing variable v_i and the number of observations per group is uniformly bounded, i.e. $max_{i=1,...,n} N_{v_i} \leq C \; \forall \, n$ for a constant C.*

Note that Assumption 10 is implied by Assumption 4 in the main text.

Similarly as before, the new estimator of α_1 defined above and the estimators of ATE for stayers will be consistent and asymptotically normal if conditions hold on second moments

of the data. Define:

$$\dot{b}_i = b_i - \frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} b_j$$

$$\ddot{B}_{n,0} = \frac{1}{n} \sum_{i \in M_n} E(\tilde{X}_i \dot{b}_i)$$

$$\dot{\epsilon}_i = \epsilon_i - \frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} \epsilon_j$$

$$\dot{r}_i = \dot{\epsilon}_i + [-1, \alpha_1]((W_i' W_i)^- W_i' U_i - \frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} (W_j' W_j)^- W_j' U_j)$$

$$\ddot{\Sigma}_{n,0} = \frac{1}{n} \sum_{i \in M_n, j \in M_n : v_j = v_i} E(\dot{r}_i \dot{r}_j \tilde{X}_i \tilde{X}_j')$$

$$\tilde{e}_{v_i} = \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} (a_j - \alpha_1 b_j)$$

$$\tilde{u}_i = \frac{1}{T} \sum_{t=1}^{T} u_{it} - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} [1, -\alpha_1](W_j' W_j)^- W_j' U_j$$

Assumption 11. *For constants C and $c > 0$:*

a) $\lambda_{min}(\ddot{\Sigma}_{n,0}) \geq c \; \forall \, n \geq C$.

b) $\ddot{B}_{n,0}' \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0} \geq c \; \forall \, n \geq C$.

c) $\tilde{\alpha}_1 \xrightarrow{p} \alpha_1$ *as* $n \to \infty$ *while* T *remains fixed.*

d) $\lambda_{min}(\frac{1}{n} Var(\begin{bmatrix} \sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \sum_{i \in M_n} \dot{r}_i \\ \sum_{i=1,...,n : x_{it} = 0 \, \forall \, t} (a_i - \tilde{e}_{v_i} + \tilde{u}_i) \\ \sum_{i=1,...,n : x_{it} = 1 \, \forall \, t} (a_i + b_i - \tilde{e}_{v_i} + \tilde{u}_i) \end{bmatrix})) \geq c \; \forall \, n \geq C$.

Assumption 11.a requires that there be within-group variation in X_i among movers since we can rewrite $\ddot{\Sigma}_{n,0} = \frac{1}{n} \sum_{i \in M_n, j \in M_n : v_j = v_i} E(r_i r_j \dot{X}_i \dot{X}_j')$ where $\dot{X}_i = \tilde{X}_i - \frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} \tilde{X}_j$. Define N to be the number of values taken by v_i. Under Assumption 10, note that N is of the same order as n. With two time periods, Assumption 11.a is implied by i) $Var(r_i | \{X_j\}_{j : v_j = v_i}) \geq c > 0 \; \forall \, i = 1, ..., n, \; \forall \, n$, ii) $Cov(r_i, r_j | \{X_{i'}\}_{i' : v_{i'} = v_i}) = 0 \; \forall \, i = 1, ..., n,$

$j \neq i$ s.t. $v_i = v_j$, $\forall\, n$ and iii)

$$\frac{1}{N}\sum_{v=1}^{N} P(\exists\, i, j \,\text{s.t.}\, v_i = v_j = v,\, x_{i1} = 0,\, x_{i2} = 1,\, x_{j1} = 1,\, x_{j2} = 0) \geq c > 0 \,\forall\, N \qquad \text{(F.4)}$$

As in the previous section, the first condition is fairly innocuous and simply requires that the model composed of the CRC model (F.1) and the generalized extrapolation identifying assumption (F.2) is not degenerate.

The second condition would be obtained by conditional cross-sectional independence of the error terms of the CRC model (F.1) and of the generalized extrapolation identifying assumption (F.2). This condition is imposed for simplicity but could be relaxed as long as the within-group dependence is not strong enough to lead to linear dependence across several observations belonging to the same group.

The third condition requires that a non-vanishing fraction of groups have both possible profiles of movers (adopters and disadopters), so that $\frac{1}{n}\sum_{i=1}^{n} E(\dot{X}_i^2) \geq c > 0 \;\forall\, i = 1, ..., n$, $\forall\, n \geq C$.

Assumption 11.b requires that the within-group variation in X_i be predictive of treatment effect b_i.

As before, Assumption 11.c is imposed for convenience since the preliminary estimator of α_1, $\tilde{\alpha}_1$, can be shown to be consistent under a similar argument as in the proof of the following proposition.

As before, Assumption 11.d is a regularity condition which guarantees that the estimators for α_1 and ATE are $\sqrt{n}$-consistent rather than being super consistent and requires that there be no approximately exact dependence between cross-sectional observations belonging to the same group and between the terms of unobserved heterogeneity a_i and $a_i + b_i$ and the average error term of the CRC model u_{it}.

Note that Assumptions 11.a, 11.b and 11.d are implied by Assumptions 4 and 6 in the main text and that Assumptions 11.c is irrelevant when $T = 2$ since the moment condition

used to estimate α_1 in this case is exactly identifying.

Note also that Assumption 8 is implied by Assumptions 4 and 5 in the main text.

To state the next proposition, define the deterministic matrices that determine the asymptotic distribution of our second-step estimator $\hat{\alpha}_1$:

$$\ddot{C}_{n,0} = \frac{1}{n}\sum_{i\in M_n} E(\tilde{X}_i[1,-\alpha_1]((W_i^{'}W_i)^{-1}W_i^{'}Z_i - \frac{1}{n_{v_i}}\sum_{j\in M_n:v_j=v_i}(W_j^{'}W_j)^{-1}W_j^{'}Z_j)) \tag{F.5}$$

and

$$\ddot{V}_{n,\alpha\gamma,0} = \frac{1}{n}\sum_{i\in M_n,j:v_j=v_i} E(\dot{r}_i X_i U_j^{'} M_{W_j} Z_j) \quad \ddot{V}_{n,\gamma,0} = \frac{1}{n}\sum_{i=1,\ldots,n,j:v_j=v_i} E(Z_i^{'} M_{W_i} U_i U_j^{'} M_{W_j} Z_j)$$

$$\ddot{A}_{n,1,0} = (\ddot{B}_{n,0}^{'}\ddot{\Sigma}_{n,0}^{-1}\ddot{B}_{n,0})^{-1}\ddot{B}_{n,0}^{'}\ddot{\Sigma}_{n,0}^{-1} \qquad \ddot{A}_{n,2,0} = (\ddot{B}_{n,0}^{'}\ddot{\Sigma}_{n,0}^{-1}\ddot{B}_{n,0})^{-1}\ddot{B}_{n,0}^{'}\ddot{\Sigma}_{n,0}^{-1}\ddot{C}_{n,0}A_{n,\gamma,0}^{-1}$$

$$\ddot{\Omega}_{n,0} = \begin{bmatrix} \ddot{\Sigma}_{n,0} & \ddot{V}_{n,\alpha\gamma,0} \\ \ddot{V}_{n,\alpha\gamma,0}^{'} & \ddot{V}_{n,\gamma,0} \end{bmatrix} \qquad \ddot{V}_{n,0} = [\ddot{A}_{n,1,0}, \ddot{A}_{n,2,0}]\ddot{\Omega}_{n,0}\begin{bmatrix} \ddot{A}_{n,1,0}^{'} \\ \ddot{A}_{n,2,0}^{'} \end{bmatrix}$$

Define the deterministic matrices which will determine the asymptotic distribution of our

second-step estimators $\hat{ATE}_{S,0}$:

$$\begin{aligned}
\tilde{a}_i &= a_i + \frac{1}{T}\sum_{t=1}^{T} u_{it} - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} [1, -\alpha_1](W_j^{'} W_j)^{-1} W_j^{'}(Y_j - Z_j\gamma) \\
&\quad - (E(a_i) - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} E(a_j - \alpha_1 b_j)), \\
\ddot{A}_{n,ATE_{S,0},\gamma,0} &= -\frac{1}{\alpha_1} \frac{1}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it} = 0 \,\forall\, t)} \frac{1}{n} \sum_{i=1,\ldots,n} E(1[x_{it} = 0 \,\forall\, t](\frac{1}{T}\sum_{t=1}^{T} z_{it} - \\
&\quad \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} [1, -\alpha_1](W_j^{'} W_j)^{-1} W_j^{'} Z_j)) A_{n,\gamma,0}^{-1}, \\
\ddot{A}_{n,ATE_{S,0},\alpha,\gamma,0} &= \frac{ATE_{S,0}}{\alpha_1} (\ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0})^{-1} \ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{C}_{n,0} A_{n,\gamma,0}^{-1} \\
\ddot{A}_{n,ATE_{S,0},\alpha,r,0} &= -\frac{ATE_{S,0}}{\alpha_1} (\ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0})^{-1} \ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \\
\ddot{A}_{n,ATE_{S,0},0} &= [\frac{1}{\alpha_1}, \ddot{A}_{n,ATE_{S,0},\gamma,0} + \ddot{A}_{n,ATE_{S,0},\alpha,\gamma,0}, \ddot{A}_{n,ATE_{S,0},\alpha,r,0}] \\
\ddot{\Omega}_{ATE_{S,0},0} &= Var(\sqrt{n} \begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=0 \,\forall t]\tilde{a}_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=0 \,\forall\, t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i^{'} M_{W_i} U_i \\ \frac{1}{n}\sum_{i \in M_n} \tilde{X}_i r_i \end{bmatrix}) \\
\ddot{V}_{n,ATE_{s,0},0} &= \ddot{A}_{n,ATE_{S,0},0} \ddot{\Omega}_{ATE_{S,0},0} \ddot{A}_{n,ATE_{S,0},0}^{'}
\end{aligned}$$

Define the deterministic matrices which will determine the asymptotic distribution of our

second-step estimators $\hat{ATE}_{S,1}$:

$$
\begin{aligned}
a \tilde{+} b_i &= a_i + b_i + \frac{1}{T}\sum_{t=1}^{T} u_{it} - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} [1, -\alpha_1](W_j^{'} W_j)^{-1} W_j^{'}(Y_j - Z_j \gamma) \\
&\quad - (E(a_i + b_i) - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} E(a_j - \alpha_1 b_j)), \\
\ddot{A}_{n, ATE_{S,1}, \gamma, 0} &= -\frac{1}{1+\alpha_1} \frac{1}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it} = 1 \,\forall\, t)} \frac{1}{n} \sum_{i=1,\ldots,n} E(1[x_{it} = 1 \,\forall\, t](\frac{1}{T}\sum_{t=1}^{T} z_{it} - \\
&\quad \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} [1, -\alpha_1](W_j^{'} W_j)^{-1} W_j^{'} Z_j)) A_{n,\gamma,0}^{-1}, \\
\ddot{A}_{n, ATE_{S,1}, \alpha, \gamma, 0} &= \frac{ATE_{S,1}}{1+\alpha_1} (\ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0})^{-1} \ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{C}_{n,0} A_{n,\gamma,0}^{-1} \\
\ddot{A}_{n, ATE_{S,1}, \alpha, r, 0} &= -\frac{ATE_{S,1}}{1+\alpha_1} (\ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0})^{-1} \ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \\
\ddot{A}_{n, ATE_{S,1}, 0} &= [\frac{1}{1+\alpha_1}, \ddot{A}_{n, ATE_{S,1}, \gamma, 0} + \ddot{A}_{n, ATE_{S,1}, \alpha, \gamma, 0}, \ddot{A}_{n, ATE_{S,1}, \alpha, r, 0}] \\
\ddot{\Omega}_{ATE_{S,1},0} &= Var(\sqrt{n} \begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it} = 1 \,\forall t] a \tilde{+} b_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it} = 1 \,\forall\, t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i^{'} M_{W_i} U_i \\ \frac{1}{n}\sum_{i \in M_n} \tilde{X}_i r_i \end{bmatrix}) \\
\ddot{V}_{n, ATE_{s,1}, 0} &= \ddot{A}_{n, ATE_{S,1}, 0} \ddot{\Omega}_{ATE_{S,1},0} \ddot{A}_{n, ATE_{S,1}, 0}^{'}
\end{aligned}
$$

Proposition 7. *Under (F.1), (F.2) and Assumptions 8, 10, and 11, as $n \to \infty$ while T remains fixed:*

$$
\begin{aligned}
\sqrt{n}(\hat{\alpha}_1 - \alpha_1) &= (\ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0})^{-1} \ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \frac{1}{\sqrt{n}} \sum_{i \in M_n} X_i \dot{r}_i \\
&\quad - (\ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0})^{-1} \ddot{B}_{n,0}^{'} \ddot{\Sigma}_{n,0}^{-1} \ddot{B}_{n,0} \ddot{C}_{n,0} A_{n,\gamma,0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i^{'} M_{W_i} U_i + o_p(1)
\end{aligned}
$$

and:

$$
\ddot{V}_{n,0}^{-\frac{1}{2}} \sqrt{n}(\hat{\alpha}_1 - \alpha_1) \xrightarrow{d} N(0,1) \tag{F.6}
$$

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

$$\sqrt{n}(\hat{ATE}_{S,0} - ATE_{S,0}) = \ddot{A}_{n,ATE_{s,0},0}\sqrt{n}\begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=0\,\forall t]\tilde{a}_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=0\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i r_i \end{bmatrix} + o_p(1)$$

$$\sqrt{n}(\hat{ATE}_{S,1} - ATE_{S,1}) = \ddot{A}_{n,ATE_{s,1},0}\sqrt{n}\begin{bmatrix} \frac{\frac{1}{n}\sum_{i=1,\ldots,n} 1[x_{it}=1\,\forall t]\tilde{a}_i}{\frac{1}{n}\sum_{i=1,\ldots,n} P(x_{it}=1\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i r_i \end{bmatrix} + o_p(1)$$

and:

$$\ddot{V}_{n,ATE_{s,0},0}^{-\frac{1}{2}}\sqrt{n}(\hat{ATE}_{S,0} - ATE_{S,0}) \xrightarrow{d} N(0,1)$$
$$\ddot{V}_{n,ATE_{s,1},0}^{-\frac{1}{2}}\sqrt{n}(\hat{ATE}_{S,1} - ATE_{S,1}) \xrightarrow{d} N(0,1)$$

Note that Proposition 3 in the main text is obtained as a special case of Proposition 7.

Proposition 7 shows that the second stage estimators of α_1 and of ATE for stayers have a linear influence function asymptotic representation, so that as before consistent variance estimation could be obtained by bootstrap resampling, although here resampling should be clustered at the level of the indexing variable v_i. Alternatively, one can also use analytical standard errors for inference, although these standard errors should now be clustered at the level of the indexing variable v_i. As before, we only show consistency of the analytical standard errors for $\hat{\alpha}_1$ here, as the same result for ATE of stayers is obtained in a similar way.

Proposition 8. *Define* $\ddot{C}_n = \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i[1, -\alpha_1]((W_i'W_i)^{-1}W_i'Z_i - \frac{1}{n_{v_i}}\sum_{j\in M_n: v_j=v_i}(W_j'W_j)^{-1}W_j'Z_j)$, $\hat{r}_i = \hat{a}_i - \hat{\alpha}_1\hat{b}_i$, $\ddot{r}_i = \hat{r}_i - \frac{1}{n_{v_i}}\sum_{j\in M_n, v_j=v_i}\hat{r}_j$, *and* $\hat{U}_i = Y_i - Z_i\hat{\gamma}$.

Define $\ddot{\Sigma}_{n,\gamma} = \frac{1}{n}\sum_{i,j=1,\ldots,n: v_j=v_i} Z_i' M_{W_i}\hat{U}_i\hat{U}_j' M_{W_j} Z_j$, $\ddot{\Sigma}_{n,\alpha\gamma} = \frac{1}{n}\sum_{i\in M_n, j\in M_n: v_j=v_i}\ddot{r}_i\tilde{X}_i\hat{U}_j' M_{W_j} Z_j$.

Define $A_{n,1} = (\ddot{B}_n'\ddot{\Sigma}_n^{-1}\ddot{B}_n)^{-1}\ddot{B}_n'\ddot{\Sigma}_n^{-1}$, $A_{n,2} = (\ddot{B}_n'\ddot{\Sigma}_n^{-1}\ddot{B}_n)^{-1}\ddot{B}_n'\ddot{\Sigma}_n^{-1}\ddot{C}_n A_n^{-1}$, $\ddot{\Omega}_n = \begin{bmatrix} \ddot{\Sigma}_n & \ddot{\Sigma}_{n,\alpha\gamma} \\ \ddot{\Sigma}_{n,\alpha\gamma}' & \ddot{\Sigma}_{n,\gamma} \end{bmatrix}$,

$$\ddot{V}_n = [\ddot{A}_{n,1}, \ddot{A}_{n,2}]\ddot{\Omega}_n \begin{bmatrix} \ddot{A}'_{n,1} \\ \ddot{A}'_{n,2} \end{bmatrix}.$$

Under (F.1), (F.2) and Assumptions 8, 10, and 11, as $n \to \infty$ while T remains fixed:

$$\ddot{V}_n^{-\frac{1}{2}}\sqrt{n}(\hat{\alpha}_1 - \alpha_1) \xrightarrow{d} N(0,1) \tag{F.7}$$

As in the previous section, this proposition shows that asymptotically valid inference for α_1 can be based on Wald tests with analytical standard errors clustered at the level of the indexing variable v_i which account for both steps of estimation.

As in the previous section, the estimators defined above and consistent standard errors can also be obtained by solving for exactly identifying moment conditions using any command capable of numerical differentiation in standard statistical software.

As in the previous section, we can test the extrapolation identifying assumption (F.2) by including additional exactly identifying moment conditions:

$$E(1[i \in M_n]x_{it}([1, -\alpha_1](W_i'W_i)^{-1}W_i'(Y_i - Z_i\gamma)$$
$$-\frac{1}{n_{v_i}} \sum_{j \in M_n, v_j = v_i} [1, -\alpha_1](W_j'W_j)^{-1}W_j'(Y_j - Z_j\gamma)) - \eta_t) = 0 \; \forall \, t \in \mathcal{S}$$

and testing the null hypothesis $H_0 : \eta_t = 0 \, \forall \, t \in \mathcal{S}$ using a Wald test with critical values from a chi-squared distribution with $|\mathcal{S}| - 1$ degrees of freedom.

G Estimation and Inference with Unbalanced Panels

Many panel datasets available in empirical work are unbalanced, i.e. some cross-sectional observations are only observed for a subset of the time periods $t = 1, ..., T$. In this section we briefly discuss the consequences of missing data if one assumes that observations are missing at random, i.e. that whether an observation (i, t) is observed are not is independent of all of the variables included in our model.

Let $o_{it} = 1[observation\,(i,t)\,is\,observed]$. Redefine $Y_i = [y_{it}]_{t:o_{it}=1}$, $W_i = [1, x_{it}]_{t:o_{it}=1}$, $Z_i = [z_{it}]_{t:o_{it}=1}$, $X_i = [x_{it}]_{t:o_{it}=1}$. Redefine M_n to be the set of cross-sectional observations that have a change in treatment status across the time periods for which observations are observed, i.e. $M_n = \{i = 1, ..., n : \exists\, t, s\, with\, x_{it} \neq x_{is},\, o_{it} = 1,\, o_{is} = 1\}$. With data missing at random, under the model used in Section B given by (E.1) and (E.2), we have:

$$E(M_{W_i}(Y_i - Z_i\gamma)) = 0$$

so that γ can be estimated by a linear regression of y_{it} on indicator variables for each cross-sectional observation, these indicator variables interacted with treatment status, and z_{it}, pooling over all observations that are observed.

We also have:

$$E(1[i \in M_n]([1, -\alpha_1](W_i'W_i)^{-1}W_i'(Y_i - Z_i\gamma) - \alpha_0)) = 0$$
$$E(1[i \in M_n]o_{it}x_{it}([1, -\alpha_1](W_i'W_i)^{-1}W_i'(Y_i - Z_i\gamma) - \alpha_0)) = 0\,\forall\, t$$

so that the parameters α_1 and α_0 can be estimated by GMM as in Section B, pooling over observations that are observed for each moment condition separately for each time period t.

Similarly testing the validity of the extrapolation identifying assumption and performing asymptotically valid inference for objects of interest such as ATE for untreated stayers can be obtained by relying on GMM estimation and pooling across all observations that are observed separately for each moment condition.

The same results apply to the use of the generalized extrapolation identifying assumption discussed in Section C.

H Proofs

H.1 Proof of Proposition 1

Recall the normalization $\hat{f}_1 = 0$ and the definition in the main text:

$$\hat{f}_2 = \frac{\sum_{i \notin M_n} \Delta y_{i2}}{n - |M_n|} \qquad = f_2 + \frac{\sum_{i \notin M_n} \Delta u_{i2}}{n - |M_n|}$$

where the second equality follows from the normalization $f_1 = 0$ and the CRC model (**??**).

By convergence in mean-square error and Assumption 1 we have:

$$\frac{1}{n}(n - |M_n|) = \frac{1}{n}\sum_{i=1}^{n} 1[x_{i1} = x_{i2}] \xrightarrow{p} \pi_S$$

where π_S was defined in the main text to be $\pi_S = P(x_{i1} = x_{i2})$.

Note that by the law of total variance and the CRC model (**??**) we have:

$$Var(1[x_{i1} = x_{i2}]\Delta u_{i2}) = \sigma^2_{\Delta u,S}\pi_S$$

where the main text defined $\sigma^2_{\Delta u,S} = Var(\Delta u_{i2}|x_{i1} = x_{i2})$. Therefore we have $Var(1[x_{i1} = x_{i2}]\Delta u_{i2}) > 0$ under Assumption 2.b and 2.c.

Assumption 2.a of bounded support implies that $1[x_{i1} = x_{i2}]\Delta u_{i2}$ has bounded support.

Therefore by the Lindeberg-Levy central limit theorem for i.i.d. observations, we have:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} 1[x_{i1} = x_{i2}]\Delta u_{i2} \xrightarrow{d} N(0, \sigma^2_{\Delta u,S}\pi_S) \tag{H.1}$$

Therefore by Slutsky's theorem, since $\pi_S > 0$ by Assumption 2.b, we have:

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

This establishes the first result of Proposition 1.

By definition, we have:

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

where as discussed above these estimators are both well-defined only for cross-sectional observations that are movers.

For all cross-sectional observations such that $\hat{a}_i$ is well-defined (movers and untreated stayers), we can write:

$$\begin{aligned}\hat{a}_i &= \frac{\sum_{i=1,2}(1-x_{it})(y_{it}-f_t)}{\sum_{i=1,2}(1-x_{it})} - (1-x_{i2})\frac{(\hat{f}_2-f_2)}{\sum_{i=1,2}(1-x_{it})} \\ &= a_i + \frac{\sum_{i=1,2}(1-x_{it})u_{it}}{\sum_{i=1,2}(1-x_{it})} - (1-x_{i2})\frac{(\hat{f}_2-f_2)}{\sum_{i=1,2}(1-x_{it})}\end{aligned}$$

where the first equality follows from the normalizations $f_1 = \hat{f}_1 = 0$ and the second equality follows from the CRC model (**??**).

The first result of this proposition shows that $\hat{f}_2 - f_2 = O_p(\frac{1}{\sqrt{n}})$, and we have $(1-x_{i2}) \in \{0,1\}$ and $\sum_{i=1,2}(1-x_{it}) \in \{1,2\}$ for movers and untreated stayers.

Therefore defining $\zeta_{a,i,n} = (1-x_{i2})\frac{(\hat{f}_2-f_2)}{\sum_{i=1,2}(1-x_{it})}$ we obtain $max_{i=1,\ldots,n:x_{i1}=0 \text{ or } x_{i2}=0}|\zeta_{a,i,n}| = O_p(\frac{1}{\sqrt{n}})$.

Similarly we have for movers and treated stayers:

$$\hat{a}_i + \hat{b}_i = a_i + b_i + \frac{\sum_{i=1,2} x_{it}u_{it}}{\sum_{i=1,2} x_{it}} - x_{i2}\frac{(\hat{f}_2-f_2)}{\sum_{i=1,2} x_{it}}$$

and defining $\zeta_{a+b,i,n} = x_{i2}\frac{(\hat{f}_2-f_2)}{\sum_{i=1,2} x_{it}}$ we obtain $max_{i=1,\ldots,n:x_{i1}=1 \text{ or } x_{i2}=1}|\zeta_{a+b,i,n}| = O_p(\frac{1}{\sqrt{n}})$, which completes the proof of this proposition.

H.2 Proof of Proposition 2

Linear influence function representation for $\hat{\alpha}_0$ and $\hat{\alpha}_1$.

Define $A_n = \begin{bmatrix} \frac{|M_n|}{n} & \frac{1}{n}\sum_{i\in M_n}\hat{b}_i \\ \frac{n_{01}}{n} & \frac{1}{n}\sum_{i\in M_n:x_{i1}=0}\hat{b}_i \end{bmatrix}$ and $w_n = \begin{bmatrix} \frac{1}{n}\sum_{i\in M_n}\hat{a}_i \\ \frac{1}{n}\sum_{i\in M_n:x_{i1}=0}\hat{a}_i \end{bmatrix}$ so that

$$\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} = A_n^{-1} w_n \tag{H.3}$$

The extrapolation identifying assumption (**??**) and Proposition 1 imply:

$$\hat{a}_i = \alpha_0 + \alpha_1 \hat{b}_i + r_i + \zeta_{i,n} \tag{H.4}$$

where $\zeta_{i,n} = \zeta_{a,i,n} - \alpha_1 \zeta_{b,i,n}$ and $\zeta_{b,i,n} = \zeta_{a+b,i,n} - \zeta_{a,i,n}$ and r_i is defined in the main text as $r_i = \epsilon_i + \sum_{t=1,2} u_{it}((1+\alpha_1)(1-x_{it}) - \alpha_1 x_{it})$.

Therefore we have:

$$\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = A_n^{-1} e_n \tag{H.5}$$

where $e_n = \begin{bmatrix} \frac{1}{n}\sum_{i\in M_n}(r_i + \zeta_{i,n}) \\ \frac{1}{n}\sum_{i\in M_n:x_{i1}=0}(r_i + \zeta_{i,n}) \end{bmatrix}$.

From Proposition 1, for cross-sectional observations that are movers:

$$\begin{aligned} \zeta_{i,n} &= (1+\alpha_1)\zeta_{a,i,n} - \alpha_1 \zeta_{ab,i,n} \\ &= (1+\alpha_1)(1-x_{i2})(\hat{f}_2 - f_2) - \alpha_1 x_{i2}(\hat{f}_2 - f_2) \end{aligned}$$

By convergence in mean-squared error:

$$\begin{bmatrix} \frac{1}{n}\sum_{i\in M_n}((1+\alpha_1)(1-x_{i2}) - \alpha_1 x_{i2}) \\ \frac{1}{n}\sum_{i\in M_n:x_{i1}=0}((1+\alpha_1)(1-x_{i2}) - \alpha_1 x_{i2}) \end{bmatrix} \xrightarrow{p} c_0$$

where $c_0 = \begin{bmatrix} E((1+\alpha_1)(1-x_{i2}) - \alpha_1 x_{i2} | x_{i1} \neq x_{i2})\pi_M \\ E((1+\alpha_1)(1-x_{i2}) - \alpha_1 x_{i2} | 0,1)\pi_{01} \end{bmatrix}$.

From the proof of Proposition 1, we also have:

$$\sqrt{n}(\hat{f}_2 - f_2) = \frac{1}{\pi_S}\frac{1}{\sqrt{n}} \sum_{i:x_{i1}=x_{i2}} \Delta u_{i2} + o_p(1) \tag{H.6}$$

From proposition 1, we also have $\hat{b}_i = b_i + \frac{\sum_t x_{it}u_{it}}{\sum_t x_{it}} - \frac{\sum_t (1-x_{it})u_{it}}{\sum_t (1-x_{it})} + \zeta_{b,i,n}$ where $max_i \zeta_{b,i,n} = o_p(1)$ and $E(\frac{\sum_t x_{it}u_{it}}{\sum_t x_{it}} - \frac{\sum_t (1-x_{it})u_{it}}{\sum_t (1-x_{it})} | x_{i1}, x_{i2}) = 0$ for $x_{i1} \neq x_{i2}$.

Therefore under Assumption 2.a of bounded support and Assumption 1, convergence in mean squared error implies:

$$A_n \xrightarrow{p} A_0 \tag{H.7}$$

where $A_0 = \begin{bmatrix} \pi_M & E(b_i | x_{i1} \neq x_{i2})\pi_M \\ \pi_{01} & E(b_i|0,1)\pi_{01} \end{bmatrix}$ where $\pi_M = \pi_{01} + \pi_{10}$.

In addition Assumptions 3.a and 3.b imply $\lambda_{min}(A_0) > 0$. Therefore by the continuous mapping theorem:

$$A_n^{-1} - A_0^{-1} = o_p(1) \tag{H.8}$$

Under Assumption 2.a of bounded support and Assumption 1 of cross-sectional independence, we have $\frac{1}{\sqrt{n}} \sum_{i \in M_n} r_i = O_p(1)$, $\frac{1}{\sqrt{n}} \sum_{i \in M_n : x_{i1}=0} r_i = O_p(1)$, and $\frac{1}{\sqrt{n}} \sum_{i:x_{i1}=x_{i2}} \Delta u_{i2} = O_p(1)$.

Therefore we have:

$$\sqrt{n}(\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}) = A_0^{-1}[I_2, c_0] \begin{bmatrix} \frac{1}{\sqrt{n}} \sum_{i \in M_n} r_i \\ \frac{1}{\sqrt{n}} \sum_{i \in M_n : x_{i1}=0} r_i \\ \frac{1}{\sqrt{n}} \sum_{i:x_{i1}=x_{i2}} \Delta u_{i2} \end{bmatrix} + o_p(1) \tag{H.9}$$

where I_2 is the 2×2 identity matrix.

Defining $\xi_{\alpha,i} = A_0^{-1}[I_2, c_0] \begin{bmatrix} 1[x_{i1} \neq x_{i2}]r_i \\ 1[x_{i1} = 0, x_{i2} = 1]r_i \\ 1[x_{i1} = x_{i2}]\Delta u_{i2} \end{bmatrix}$ completes the proof of the first result of Proposition 2, showing that the estimator $\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix}$ has an asymptotic influence function representation.

Asymptotic normality of $\hat{\alpha}_0$ and $\hat{\alpha}_1$.

Define

$$w_i = \begin{bmatrix} 1[x_{i1} \neq x_{i2}]r_i \\ 1[x_{i1} = 0, x_{i2} = 1]r_i \\ 1[x_{i1} = x_{i2}]\Delta u_{i2} \end{bmatrix} \tag{H.10}$$

From Assumption 2.a of bounded support, w_i has bounded support.

As in the proof of Proposition 1, Assumptions 2.b and 2.c guarantee that $Var(1[x_{i1} = x_{i2}]\Delta u_{i2}) > 0$. Assumptions 3.a and 3.c guarantee that $Var(1[x_{i1} \neq x_{i2}]r_i) > 0$ and $Var(1[x_{i1} = 0, x_{i2} = 1]r_i) > 0$.

Assumption 1 of cross-sectional independence guarantees that $1[x_{i1} = x_{i2}]\Delta u_{i2}$ is uncorrelated with $1[x_{i1} \neq x_{i2}]r_i$ and $1[x_{i1} = 0, x_{i2} = 1]r_i$.

Finally

$$\begin{aligned} Cov(1[x_{i1} \neq x_{i2}]r_i, 1[x_{i1} = 0, x_{i2} = 1]r_i) &= E(1[x_{i1} = 0, x_{i2} = 1]r_i^2) \\ &= Var(1[x_{i1} = 0, x_{i2} = 1]r_i) \end{aligned}$$

and

$$\begin{aligned} Var(1[x_{i1} \neq x_{i2}]r_i) &= Var(1[x_{i1} = 0, x_{i2} = 1]r_i) + Var(1[x_{i1} = 1, x_{i2} = 0]r_i) \\ &> Var(1[x_{i1} = 0, x_{i2} = 1]r_i) \end{aligned}$$

where the strict inequality follows from $Var(r_i|1, 0) > 0$ (Assumption 3.c) and $\pi_{10} > 0$

(Assumption 3.a).

Therefore $|Corr(1[x_{i1} \neq x_{i2}]r_i, 1[x_{i1} = 0, x_{i2} = 1]r_i)| < 1$, so that:

$$\lambda_{min} Var(w_i) > 0 \tag{H.11}$$

where λ_{min} denotes the smallest eigenvalue of a matrix.

Therefore all conditions (bounded support, positive variance, i.i.d.) are met for the Lindeberg-Levy central limit theorem to apply:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \xrightarrow{d} N(0, Var(w_i)) \tag{H.12}$$

so that we obtain by Slutsky's theorem:

$$\sqrt{n}(\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}) \xrightarrow{d} N(0, V_\alpha) \tag{H.13}$$

where $V_\alpha = A_0^{-1}[I_2, c_0] Var(w_i) \begin{bmatrix} I_2 \\ c_0^{'} \end{bmatrix} A_0^{-1'}$.

Linear influence function representation and asymptotic normality of $A\hat{T}E_{S,0}$ and $A\hat{T}E_{S,1}$

By definition:

$$\begin{aligned} A\hat{T}E_{S,0} &= \frac{\bar{a}_{00} - \hat{\alpha}_0}{\hat{\alpha}_1} \\ &= \frac{\frac{1}{n_{00}} \sum_{i: x_{i1} = x_{i2} = 0} (a_i + \frac{1}{2} \sum_{t=1,2} u_{it} - \frac{1}{2}(\hat{f}_2 - f_2)) - \hat{\alpha}_0}{\hat{\alpha}_1} \end{aligned}$$

Define $\tilde{a}_i = a_i + \frac{1}{2}\sum_{t=1,2} u_{it}$ and define:

$$w_{ATE,i} = \begin{bmatrix} 1[x_{i1} = x_{i2} = 0](\tilde{a}_i - E(a_i|0,0)) \\ w_i \end{bmatrix} \tag{H.14}$$

where w_i was defined above as $w_i = \begin{bmatrix} 1[x_{i1} \neq x_{i2}]r_i \\ 1[x_{i1} = 0, x_{i2} = 1]r_i \\ 1[x_{i1} = x_{i2}]\Delta u_{i2} \end{bmatrix}$.

As before, the variance of w_i is positive definite. In addition Assumption 1 implies that $1[x_{i1} = x_{i2} = 0](\tilde{a}_i - E(a_i|0,0))$ is uncorrelated with $1[x_{i1} \neq x_{i2}]r_i$ and $1[x_{i1} = 0, x_{i2} = 1]r_i$.

Assumption 3.c implies that $Var(\tilde{a}_i|\Delta u_{i2}, 0, 0) > 0$, so that $|Corr(1[x_{i1} = x_{i2} = 0](\tilde{a}_i - E(a_i|0,0)), 1[x_{i1} = x_{i2}]\Delta u_{i2})| < 1$.

Therefore $\lambda_{min} Var(w_{ATE,i}) > 0$. In addition Assumption 2.a guarantees that $w_{ATE,i}$ has bounded support.

Therefore by the Lindeberg-Levy central limit theorem we have:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{ATE,i} \xrightarrow{d} N(0, Var(w_{ATE,i})) \tag{H.15}$$

Therefore since $ATE_{S,0} = \frac{E(a_i|0,0)-\alpha_0}{\alpha_1}$, by the δ-method we obtain:

$$\sqrt{n}(\hat{ATE}_{S,0} - ATE_{S,0}) = A_{ATE,0}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{ATE,i} + o_p(1) \xrightarrow{d} N(0, A_{ATE,0} Var(w_{ATE,i}) A'_{ATE,0}) \tag{H.16}$$

where $A_{ATE,0} = [\frac{1}{\alpha_1}, -\frac{1}{\alpha_1}, -\frac{ATE_{S,0}}{\alpha_1}]B_{ATE,0}$ and $B_{ATE,0} = \begin{bmatrix} \frac{1}{\pi_{00}} & 0'_2 & -\frac{1}{\pi_{00}}\frac{1}{2}\frac{1}{\pi_S} \\ 0_2 & A_0^{-1} & A_0^{-1}c_0 \end{bmatrix}$ where 0_2 is a 2×1 vector of zeros and A_0, c_0 are defined above.

The same steps can be used for $\hat{ATE}_{S,1}$, which completes the proof of Proposition 2.

H.3 Definitions and Lemma for the Proof of Proposition 3

The first step of our estimation procedure is unchanged when estimating ATE under the generalized extrapolation identifying assumption (3.7). We first establish the same result for these first step estimators as Proposition 1 but under Assumptions 4 and 5.

Recall the redefinition in the main text of $\sigma^2_{\Delta u,S}$ to $\sigma^2_{\Delta u,S} = Var(\sum_{i:v_i=v,x_{i1}=x_{i2}} \Delta u_{i2})$. Also redefine $\pi_S = E(\sum_{i:v_i=v} 1[x_{i1} = x_{i2}])$.

Lemma 1. *Under the CRC model (??) and Assumptions 4 and 5, as $N \to \infty$:*

$$\sqrt{N}(\hat{f}_2 - f_2) \xrightarrow{d} N(0, \frac{\sigma^2_{\Delta u,S}}{\pi^2_S}) \tag{H.17}$$

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

$$\hat{a}_i = a_i + \frac{\sum_{t=1,2}(1 - x_{it})u_{it}}{\sum_{t=1,2}(1 - x_{it})} + \zeta_{a,i,N}, \quad \hat{a}_i + \hat{b}_i = a_i + b_i + \frac{\sum_{t=1,2} x_{it}u_{it}}{\sum_{t=1,2} x_{it}} + \zeta_{a+b,i,N} \tag{H.18}$$

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

Proof. As in the proof of Proposition 1, we have:

$$\sqrt{N}(\hat{f}_2 - f_2) = \frac{\sum_{v=1,...,N} \sum_{i:v_i=v,x_{i1}=x_{i2}} \Delta u_{i2}}{\sum_{v=1,...,N} \sum_{i:v_i=v} 1[x_{i1} = x_{i2}]} \tag{H.19}$$

Under Assumption 4, convergence in mean-squared error implies:

$$\frac{1}{N} \sum_{v=1,...,N} \sum_{i:v_i=v} 1[x_{i1} = x_{i2}] \xrightarrow{p} \pi_S \tag{H.20}$$

and Assumption 2.b implies $\pi_S > 0$.

Assumption 4 and Assumption 2.a imply that $\sum_{i:v_i=v,x_{i1}=x_{i2}} \Delta u_{i2}$ is i.i.d. across v with bounded support. Assumption 5 implies $Var(\sum_{i:v_i=v,x_{i1}=x_{i2}} \Delta u_{i2}) = \sigma^2_{\Delta u,S} > 0$. Therefore by the continuous mapping theorem, Slutsky's theorem, and the Lindeberg-Levy central limit

theorem, we have:

$$\sqrt{N}(\hat{f}_2 - f_2) = \frac{1}{\pi_S}\frac{1}{\sqrt{N}}\sum_{v=1,\ldots,N}\sum_{i:v_i=v,x_{i1}=x_{i2}}\Delta u_{i2} + o_p(1)$$
$$\xrightarrow{d} N(0, \frac{\sigma^2_{\Delta u,S}}{\pi^2_S})$$

Given this result, the rest of the proof of Lemma 1 is as in the proof of Proposition 1. □

Here we also define explicitly the second-step estimator for α_1 used with the generalized extrapolation identifying assumption (3.7). A fixed effects instrumental variable regression of $\hat{a}_i$ on $\hat{b}_i$ using x_{i2} as an instrumental variable, with fixed effects indexed by v_i, yields the estimator:

$$\hat{\alpha}_1 = \frac{\sum_{v=1,\ldots,N}\sum_{i\in M_n:v_i=v} x_{i2}\ddot{a}_i}{\sum_{v=1,\ldots,N}\sum_{i\in M_n:v_i=v} x_{i2}\ddot{b}_i} \tag{H.21}$$

where

$$\ddot{a}_i = \hat{a}_i - \bar{a}_i, \qquad \bar{a}_i = \frac{1}{n_{v_i}}\sum_{j\in M_n,v_j=v_i}\hat{a}_j$$
$$\ddot{b}_i = \hat{b}_i - \bar{b}_i, \qquad \bar{b}_i = \frac{1}{n_{v_i}}\sum_{j\in M_n,v_j=v_i}\hat{b}_j$$

The estimated fixed effects are given by:

$$\hat{e}_v = \frac{1}{n_v}\sum_{i\in M_n,v_i=v}(\hat{a}_i - \hat{\alpha}_1\hat{b}_i) \tag{H.22}$$

H.4 Proof of Proposition 3

From the CRC model (**??**), the generalized extrapolation identifying assumption (3.7), and Lemma 1:

$$\hat{a}_i = e_{v_i} + \alpha_1\hat{b}_i + r_i + \zeta_{i,n} \tag{H.23}$$

where as before $r_i = \epsilon_i + \sum_{t=1,2} u_{it}((1+\alpha_1)(1-x_{it}) - \alpha_1 x_{it})$ and $\zeta_{i,n} = (1+\alpha_1)(1-x_{i2})(\hat{f}_2 - f_2) - \alpha_1 x_{i2}(\hat{f}_2 - f_2)$.

Therefore we have:

$$\hat{\alpha}_1 - \alpha_1 = \frac{\sum_{v=1,...,N} \sum_{i \in M_n : v_i = v} x_{i2}(\dot{r}_i + \dot{\zeta}_{i,n})}{\sum_{v=1,...,N} \sum_{i \in M_n : v_i = v} x_{i2} \ddot{b}_i} \tag{H.24}$$

where $\dot{r}_i = r_i - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} r_j$ and $\dot{\zeta}_{i,n} = \zeta_{i,n} - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} \zeta_{j,n}$.

Considering the denominator, we have:

$$\begin{aligned}\sum_{v=1,...,N} \sum_{i \in M_n : v_i = v} x_{i2} \ddot{b}_i &= \sum_{v=1,...,N} \sum_{i \in M_n : v_i = v} x_{i2} \hat{b}_i - \sum_{v=1,...,N} \frac{n_{v,01}}{n_v} \sum_{i \in M_n : v_i = v} \hat{b}_i \\ &= \sum_{v=1,...,N} \frac{n_{v,01} n_{v,10}}{n_v} \left(\frac{1}{n_{v,01}} \sum_{i \in M_n : v_i = v} x_{i2} \hat{b}_i - \frac{1}{n_{v,10}} \sum_{i \in M_n : v_i = v} (1 - x_{i2}) \hat{b}_i \right)\end{aligned}$$

where $n_{v,x_1 x_2} = |\{i = 1, ..., n : v_i = v, x_{i1} = x_1, x_{i2} = x_2\}|$.

Under Assumption 4 we have $\frac{n_{v,01} n_{v,10}}{n_v} \leq C$. From Lemma 1 we therefore have:

$$\begin{aligned}&\frac{1}{N} \sum_{v=1,...,N} \frac{n_{v,01} n_{v,10}}{n_v} \left(\frac{1}{n_{v,01}} \sum_{i \in M_n : v_i = v} x_{i2} \hat{b}_i - \frac{1}{n_{v,10}} \sum_{i \in M_n : v_i = v} (1 - x_{i2}) \hat{b}_i \right) \\ = &\frac{1}{N} \sum_{v=1,...,N} \frac{n_{v,01} n_{v,10}}{n_v} \left(\frac{1}{n_{v,01}} \sum_{i \in M_n : v_i = v} x_{i2}(b_i + \Delta u_{i2}) - \frac{1}{n_{v,10}} \sum_{i \in M_n : v_i = v} (1 - x_{i2})(b_i - \Delta u_{i2}) \right) + o_p(1)\end{aligned}$$

and convergence in mean-squared error implies:

$$\begin{aligned}&\frac{1}{N} \sum_{v=1,...,N} \frac{n_{v,01} n_{v,10}}{n_v} \left(\frac{1}{n_{v,01}} \sum_{i \in M_n : v_i = v} x_{i2}(b_i + \Delta u_{i2}) - \frac{1}{n_{v,10}} \sum_{i \in M_n : v_i = v} (1 - x_{i2})(b_i - \Delta u_{i2}) \right) \\ &\qquad \xrightarrow{p} E\left(\frac{n_{v,01} n_{v,10}}{n_v} (b_{01,v} - b_{10,v}) \right)\end{aligned}$$

Define $\Delta_b = E(\frac{n_{v,01} n_{v,10}}{n_v}(b_{01,v} - b_{10,v}))$. We have $\Delta_b \neq 0$ since $\frac{n_{v,01} n_{v,10}}{n_v} \geq 0$, $\frac{n_{v,01} n_{v,10}}{n_v} \geq c$ with positive probability under Assumptions 6.a and 4, $\frac{n_{v,01} n_{v,10}}{n_v}$ has discrete support under Assumption 4, and $b_{01,v} - b_{10,v} > 0$ whenever $\frac{n_{v,01} n_{v,10}}{n_v} > 0$ or $b_{01,v} - b_{10,v} < 0$ whenever

$\frac{n_{v,01}n_{v,10}}{n_v} > 0$ under Assumption 6.c.

Convergence in mean-squared error and Lemma 1 imply:

$$\frac{1}{\sqrt{N}} \sum_{v=1,...,N} \sum_{i \in M_n : v_i = v} x_{i2}(\dot{r}_i + \dot{\zeta}_{i,n})$$
$$= \frac{1}{\sqrt{N}} \sum_{v=1,...,N} \sum_{i \in M_n : v_i = v} x_{i2}\dot{r}_i + c_0 \frac{1}{\pi_S} \frac{1}{\sqrt{N}} \sum_{v=1,...,N} \sum_{i : v_i = v, x_{i1} = x_{i2}} \Delta u_{i2} + o_p(1)$$

where $c_0 = E(\sum_{i \in M_n : v_i = v} x_{i2}((1+\alpha_1)(1-\dot{x}_{i2}) - \alpha_1 \dot{x}_{i2}))$, $\dot{x}_{i2} = x_{i2} - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} x_{j2}$.

Define $w_v = \begin{bmatrix} \sum_{i \in M_n : v_i = v} x_{i2}\dot{r}_i \\ \sum_{i : v_i = v, x_{i1} = x_{i2}} \Delta u_{i2} \end{bmatrix}$. $E(w_v) = 0$ under the CRC model and the generalized extrapolation identifying assumption. Assumption 6.d imposes that $\lambda_{min}(Var(w_v)) > 0$. w_v has bounded support under Assumption 2.a.

Therefore under Assumption 4, the continuous mapping theorem, Slutsky's theorem, and the Lindeberg-Levy central limit theorem imply:

$$\sqrt{N}(\hat{\alpha}_1 - \alpha_1) = A_{\alpha,0} \frac{1}{\sqrt{N}} \sum_{v=1,...,N} w_v \xrightarrow{d} N(0, V_\alpha) \tag{H.25}$$

where $V_\alpha = A_{\alpha,0} Var(w_v) A_{\alpha,0}^{'}$ and $A_{\alpha,0} = \frac{1}{\Delta_b}[1, c_0 \frac{1}{\pi_S}]$.

By definition:

$$\hat{ATE}_{S,0} = \frac{\frac{1}{n_{00}} \sum_{i : x_{i1} = x_{i2} = 0} (\hat{a}_i - \frac{1}{n_{v_i}} \sum_{j \in M_n : v_j = v_i} (\hat{a}_j - \hat{\alpha}_1 \hat{b}_j))}{\hat{\alpha}_1}$$

Redefine $\pi_{00} = E(\sum_{i : v_i = v} 1[x_{i1} = x_{i2} = 0])$, note that Assumption 6.d implies $\pi_{00} > 0$. By convergence in mean-squared error we obtain:

$$\frac{n_{00}}{N} \xrightarrow{p} \pi_{00} \tag{H.26}$$

By convergence in mean-squared error and the previous results, we can also write:

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{v=1}^{N}\sum_{i:v_i=v,x_{i1}=x_{i2}=0}(\hat{a}_i-\frac{1}{n_v}\sum_{j\in M_n:v_j=v_i}(\hat{a}_j-\hat{\alpha}_1\hat{b}_j))\\
=&\frac{1}{\sqrt{N}}\sum_{v=1}^{N}\sum_{i:v_i=v,x_{i1}=x_{i2}=0}(a_i+\frac{1}{2}\sum_{t=1,2}u_{it}-e_v-\frac{1}{n_v}\sum_{j\in M_n:v_j=v_i}r_j)\\
&+d_0\sqrt{n}(\hat{f}_2-f_2)+e_0\sqrt{n}(\hat{\alpha}_1-\alpha_1)+o_p(1)\\
=&\frac{1}{\sqrt{N}}\sum_{v=1}^{N}\sum_{i:v_i=v,x_{i1}=x_{i2}=0}(a_i+\frac{1}{2}\sum_{t=1,2}u_{it}-e_v-\frac{1}{n_v}\sum_{j\in M_n:v_j=v_i}r_j)\\
&+d_0\frac{1}{\pi_S}\frac{1}{\sqrt{N}}\sum_{v=1,\ldots,N}\sum_{i:v_i=v,x_{i1}=x_{i2}}\Delta u_{i2}\\
&+e_0\frac{1}{\Delta_b}[1,c_0\frac{1}{\pi_S}]\frac{1}{\sqrt{N}}\sum_{v=1,\ldots,N}w_v+o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
d_0&=\frac{1}{N}\sum_{v=1}^{N}E(\sum_{i:v_i=v}(-\frac{1[x_{i1}=x_{i2}]}{2}-\frac{1}{n_v}\sum_{j\in M_n:v_j=v_i}(-(1+\alpha_1)(1-x_{j2})+\alpha_1x_{j2})))\\
e_0&=E(\sum_{i:v_i=v,x_{i1}=x_{i2}=0}\frac{1}{n_v}\sum_{j\in M_n:v_j=v}b_j)
\end{aligned}$$

Therefore by the δ-method we have:

$$\sqrt{n}(\hat{ATE}_{S,0}-ATE_{S,0})=\frac{1}{\sqrt{N}}\sum_{v=1,\ldots,N}A_{ATE,0}w_{ATE_0,v}+o_p(1) \tag{H.27}$$

where $A_{ATE,0}=[\frac{1}{\alpha_1},-\frac{ATE_{S,0}}{\alpha_1}]B_{ATE,0}$, $B_{ATE,0}=\begin{bmatrix}\frac{1}{\pi_{00}} & e_0\frac{1}{\Delta_b} & (d_0+e_0\frac{1}{\Delta_b}c_0)\frac{1}{\pi_S}\\ 0 & \frac{1}{\Delta_b} & \frac{1}{\Delta_b}c_0\frac{1}{\pi_S}\end{bmatrix}$, and:

$$w_{ATE_0,v}=\begin{bmatrix}\sum_{i:v_i=v,x_{i1}=x_{i2}=0}(a_i-e_{v_i}-E(a_i-e_{v_i}|x_{i1}=x_{i2}=0)+\tilde{u}_i))\\ w_v\end{bmatrix} \tag{H.28}$$

where $\tilde{u}_i$ was defined in the main text to be $\tilde{u}_i = \frac{1}{2}\sum_{t=1}^{T} u_{it} - \frac{1}{n_{v_i}}\sum_{j \in M_n : v_j = v_i} r_j$ and w_v is defined above.

Bounded support, Assumption 6.d, and Assumption 4 lead to the applicability of the Lindeberg-Levy central limit theorem, so that:

$$\sqrt{n}(\hat{ATE}_{S,0} - ATE_{S,0}) \xrightarrow{d} N(0, V_{ATE}) \tag{H.29}$$

where $V_{ATE} = A_{ATE,0} Var(w_{ATE_0,v}) A'_{ATE,0}$.

This completes the proof of this proposition since the same steps can be used to derive the asymptotic normality of the estimator of ATE for treated stayers, $\hat{ATE}_{S,1}$.

H.5 Proof of Proposition 4

The choice of generalized inverse used here (excluding the interaction between the cross-sectional indicator variables and treatment status for stayers) implies:

$$(W_i' W_i)^{-} W_i' Z_i = \begin{cases} [\bar{z}_{it}^0, \bar{z}_{it}^1 - \bar{z}_{it}^0]' & if\ i\ is\ a\ mover \\ [\bar{z}_{it}^0, 0]' & if\ x_{it} = 0\ \forall\, t \\ [\bar{z}_{it}^1, 0]' & if\ x_{it} = 1\ \forall\, t \end{cases} \tag{H.30}$$

where $\bar{z}_{it}^0 = \frac{\sum_t (1 - x_{it}) z_{it}}{\sum_t (1 - x_{it})}$ and $\bar{z}_{it}^1 = \frac{\sum_t x_{it} z_{it}}{\sum_t x_{it}}$, and $(W_i' W_i)^{-} W_i' U_i$ has a similar representation.

Therefore under Assumption 8.a: $Z_i' M_{W_i} Z_i$, $Z_i' M_{W_i} U_i$, $(W_i' W_i)^{-} W_i' Z_i$ and $(W_i' W_i)^{-} W_i' U_i$ have bounded support.

The definition of the estimator and (E.3) implies:

$$\sqrt{n}(\hat{\gamma} - \gamma) = (\frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} Z_i)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i' M_{W_i} U_i$$

Under Assumption 7, $Z_i' M_{W_i} Z_i$ and $Z_i' M_{W_i} U_i$ are cross-sectionally independent.

Therefore convergence in mean squared error implies:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} Z_i - A_{n,\gamma,0} \xrightarrow{p} 0$$

Assumption 8.d implies that $V_{n,\gamma,0} = Var(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i' M_{W_i} U_i) \geq c, \forall\, n \geq C$, for constants $c > 0$ and C, and as discussed above $Z_i' M_{W_i} U_i$ has bounded support.

Therefore a central limit theorem for independent observations such as Theorem 5.11 in White (2001) implies:

$$V_{n,\gamma,0}^{-\frac{1}{2}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i' M_{W_i} U_i \xrightarrow{d} N(0, I_K) \tag{H.31}$$

These two results complete the proof of $V_{n,\gamma,0}^{-\frac{1}{2}} A_{n,\gamma,0} \sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, I_K)$.

From the definition of the estimators $\hat{a}_i$ and $\hat{b}_i$ we also have:

$$\begin{aligned}
\begin{bmatrix} \hat{a}_i \\ \hat{b}_i \end{bmatrix} &= (W_i' W_i)^{-} W_i' (Y_i - Z_i \hat{\gamma}) \\
&= \begin{bmatrix} a_i \\ b_i \end{bmatrix} + (W_i' W_i)^{-} W_i' (U_i - Z_i(\hat{\gamma} - \gamma))
\end{aligned}$$

From the above, we have:

$$\begin{aligned}
(W_i' W_i)^{-} W_i' Z_i (\hat{\gamma} - \gamma) = {} & (W_i' W_i)^{-} W_i' Z_i A_{n,\gamma,0}^{-1} \frac{1}{\sqrt{n}} \zeta_n \\
& + (W_i' W_i)^{-} W_i' Z_i ((\frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} Z_i)^{-1} - A_{n,\gamma,0}^{-1}) \frac{1}{\sqrt{n}} \zeta_n
\end{aligned}$$

By the continuous mapping theorem and Assumption 8.c:

$$(\frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} Z_i)^{-1} - A_{n,\gamma,0}^{-1} = o_p(1) \tag{H.32}$$

The first result in this proof shows that $(W_i' W_i)^{-} W_i' Z_i$ has bounded support and we have

shown that $\frac{1}{\sqrt{n}}\zeta_n \xrightarrow{d} N(0,1)$, so that $\frac{1}{\sqrt{n}}\zeta_n$ is bounded in probability.

Therefore:

$$\begin{aligned} max_{i=1,\ldots,n}((W_i'W_i)^{-}W_i'Z_i)((\frac{1}{n}\sum_{i=1}^{n} Z_i'M_{W_i}Z_i)^{-1} - A_{n,\gamma,0}^{-1})\frac{1}{\sqrt{n}}\zeta_n \\ = O_p(1) \times o_p(1) \times O_p(1) \\ = o_p(1) \end{aligned}$$

which completes the proof.

H.6 Proof of Proposition 5

From the definition of the estimator $\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix}$ we can rewrite:

$$\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = (B_n'\Sigma_n^{-1}B_n)^{-1}B_n'\Sigma_n^{-1}\frac{1}{n}\sum_{i\in M_n}\tilde{X}_i(\epsilon_i + [1,-\alpha_1](W_i'W_i)^{-1}W_i'(U_i + Z_i(\gamma - \hat{\gamma}))) \tag{H.33}$$

Define $C_n^\star = \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i[1,-\alpha_1](W_i'W_i)^{-1}W_i'Z_i$. From Assumption 7 and Assumption 8.a, by convergence in mean squared error:

$$C_n^\star - C_{n,0} = o_p(1) \tag{H.34}$$

By definition:

$$\begin{aligned} B_n &= \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i[1,\hat{b}_i] \\ &= \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i[1, b_i + [0,1](W_i'W_i)^{-1}W_i'(U_i + Z_i(\hat{\gamma} - \gamma))] \\ &= \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i[1, b_i + [0,1](W_i'W_i)^{-1}W_i'U_i] + o_p(1) \end{aligned}$$

where the last equality follows from Proposition 4.

As in the proof of Proposition 4, $\tilde{X}_i[1, b_i + [0,1](W_i'W_i)^{-1}W_i'U_i]$ has bounded support under Assumption 8.a, so that by convergence in mean-squared error:

$$B_n - B_{n,0} = o_p(1) \tag{H.35}$$

By definition of Σ_n:

$$\begin{aligned}
\Sigma_n &= \frac{1}{n}\sum_{i\in M_n} \tilde{\epsilon}_i^2 \tilde{X}_i\tilde{X}_i' \\
&= \frac{1}{n}\sum_{i\in M_n} (\hat{a}_i - \tilde{\alpha}_0 - \tilde{\alpha}_1\hat{b}_i)^2 \tilde{X}_i\tilde{X}_i' \\
&= \frac{1}{n}\sum_{i\in M_n} (\hat{a}_i - \alpha_0 - \alpha_1\hat{b}_i - [1,\hat{b}_i]\begin{bmatrix}\tilde{\alpha}_0 - \alpha_0 \\ \tilde{\alpha}_1 - \alpha_1\end{bmatrix})^2 \tilde{X}_i\tilde{X}_i' \\
&= \frac{1}{n}\sum_{i\in M_n} (r_i - [1,-\alpha_1](W_i'W_i)^{-1}W_i'Z_i(\hat{\gamma}-\gamma) - [1,\hat{b}_i]\begin{bmatrix}\tilde{\alpha}_0 - \alpha_0 \\ \tilde{\alpha}_1 - \alpha_1\end{bmatrix})^2 \tilde{X}_i\tilde{X}_i'
\end{aligned}$$

From Assumption 9.c, Proposition 4, and Assumption 8.a:

$$\begin{aligned}
max_{i=1,\ldots,n}[1,\hat{b}_i]\begin{bmatrix}\tilde{\alpha}_0 - \alpha_0 \\ \tilde{\alpha}_1 - \alpha_1\end{bmatrix} &= o_p(1) \\
max_{i=1,\ldots,n}[1,-\alpha_1](W_i'W_i)^{-1}W_i'Z_i(\hat{\gamma}-\gamma) &= o_p(1) \\
max_{i=1,\ldots,n} r_i &= O_p(1)
\end{aligned}$$

Therefore:

$$\Sigma_n = \frac{1}{n}\sum_{i\in M_n} r_i^2 \tilde{X}_i\tilde{X}_i' + o_p(1) \tag{H.36}$$

As previously, under Assumptions 7 and 8.a, convergence in mean squared error implies:

$$\Sigma_n - \Sigma_{n,0} = o_p(1) \tag{H.37}$$

Therefore under Assumptions 9.a and 9.b, by the continuous mapping theorem and Proposition 4, we have:

$$\begin{aligned}\sqrt{n}(\begin{bmatrix}\hat{\alpha}_0\\ \hat{\alpha}_1\end{bmatrix} - \begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix}) &= (B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}\frac{1}{\sqrt{n}}\sum_{i\in M_n}\tilde{X}_i r_i \\ &\quad - (B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}C_{n,0}A_{n,\gamma,0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i^{'}M_{W_i}U_i + o_p(1)\end{aligned}$$

which completes the proof of the first result of Proposition 5.

With the first result of Proposition 5 established, the second result can be obtained by using a central limit theorem for independent observations. To apply this central limit theorem, we show that higher moments of the linear influence function for $\begin{bmatrix}\hat{\alpha}_0\\ \hat{\alpha}_1\end{bmatrix}$ are bounded and that the variance of the linear influence function is uniformly positive-definite.

From Assumptions 8.a, 9.a, and 9.b, we have that the support of $(B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}\tilde{X}_i r_i$ is bounded.

From Assumptions 8.a, 9.a, 9.b, and 8.c, we have that the support of

$$(B_{n,0}^{'}\Sigma_{n,0}^{-1}B_{n,0})^{-1}B_{n,0}^{'}\Sigma_{n,0}^{-1}C_{n,0}A_{n,\gamma,0}^{-1}Z_i^{'}M_{W_i}U_i \tag{H.38}$$

is bounded.

Assumption 7 of cross-sectional independence, Assumption 8.b that a non-vanishing share of the population be stayers, and Assumptions 8.d and 9.a imply that $\lambda_{min}(\Omega_{n,0}) \geq c$.

Therefore $\lambda_{min}(V_{n,0}) \geq c$ as long as:

$$\lambda_{min}([A_{n,1,0}, A_{n,2,0}] \begin{bmatrix} A'_{n,1,0} \\ A'_{n,2,0} \end{bmatrix}) \geq c \tag{H.39}$$

which follows from $\lambda_{min}(B'_{n,0}\Sigma_{n,0}^{-1}B_{n,0})^{-1} \geq c$, which itself follows from $\lambda_{min}(B'_{n,0}\Sigma_{n,0}^{-1}B_{n,0})^{-1} = \frac{1}{\lambda_{max}(B'_{n,0}\Sigma_{n,0}^{-1}B_{n,0})}$, and $\lambda_{max}\Sigma_{n,0}^{-1} = \frac{1}{\lambda_{min}\Sigma_{n,0}} \leq C$ by Assumption 9.a, and $\lambda_{max}(B'_{n,0}B_{n,0}) \leq C$ by Assumption 8.a.

Therefore all conditions are met to use a central limit theorem for independent observations such as Theorem 5.11 in White (2001), and we have:

$$V_{n,0}^{-\frac{1}{2}}\sqrt{n}(\begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}) \xrightarrow{d} N(0, I_2) \tag{H.40}$$

which completes the proof of the second result of Proposition 5.

We can show that the estimators of ATE for stayers, $A\hat{T}E_{S,0}$ and $A\hat{T}E_{S,1}$, have a linear influence function representation using similar steps as above. Under the assumptions of this proposition we have:

$$\sqrt{n}(A\hat{T}E_{S,0} - ATE_{S,0}) = A_{n,ATE_{s,0},0}\sqrt{n} \begin{bmatrix} \frac{\sum_{i:x_{it}=0\,\forall t} \tilde{a}_i}{\sum_{i=1,\ldots,n} P(x_{it}=0\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i r_i \end{bmatrix} + o_p(1)$$

$$\sqrt{n}(A\hat{T}E_{S,1} - ATE_{S,1}) = A_{n,ATE_{s,1},0}\sqrt{n} \begin{bmatrix} \frac{\sum_{i:x_{it}=1\,\forall t} \tilde{a+b_i}}{\sum_{i=1,\ldots,n} P(x_{it}=1\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n} \tilde{X}_i r_i \end{bmatrix} + o_p(1)$$

For concision we concentrate on $A\hat{T}E_{S,0}$ in the remainder of this proof since the asymptotic normality of $A\hat{T}E_{S,1}$ is derived in the same way.

As above, $A_{n,ATE_{s,0},0}\begin{bmatrix} \frac{\sum_{i:x_{it}=0\,\forall t}\tilde{a}_i}{\sum_{i=1,\ldots,n}P(x_{it}=0\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i \end{bmatrix}$ has bounded support under the assumptions of this proposition, so that we can apply a central limit theorem for independent observations if $Var(\sqrt{n}\begin{bmatrix} \frac{\sum_{i:x_{it}=0\,\forall t}\tilde{a}_i}{\sum_{i=1,\ldots,n}P(x_{it}=0\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i \end{bmatrix})$ is uniformly positive-definite since $A_{n,ATE_{s,0},0}A'_{n,ATE_{s,0},0} \geq c$ $\forall\, n \geq C$.

Assumptions 8.d, 9.a, and 9.d impose that the variance of each term in $\sqrt{n}\begin{bmatrix} \frac{\sum_{i:x_{it}=0\,\forall t}\tilde{a}_i}{\sum_{i=1,\ldots,n}P(x_{it}=0\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i \end{bmatrix}$ is uniformly positive definite.

As above, Assumptions 7, 8.b, and 8.d guarantee that $\frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i$ and $\frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i$ are not approximately linearly dependent.

Assumption 7 of cross-sectional independence guarantees that $\frac{\sum_{i:x_{it}=0\,\forall t}\tilde{a}_i}{\sum_{i=1,\ldots,n}P(x_{it}=0\,\forall t)}$ and $\frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i$ are independent.

Assumption 9.d guarantees that $\frac{\sum_{i:x_{it}=0\,\forall t}\tilde{a}_i}{\sum_{i=1,\ldots,n}P(x_{it}=0\,\forall t)}$ and $\frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i$ are not approximately linearly dependent.

Therefore we have $\lambda_{min} Var(\sqrt{n}\begin{bmatrix} \frac{\sum_{i:x_{it}=0\,\forall t}\tilde{a}_i}{\sum_{i=1,\ldots,n}P(x_{it}=0\,\forall t)} \\ \frac{1}{n}\sum_{i=1}^{n} Z_i' M_{W_i} U_i \\ \frac{1}{n}\sum_{i\in M_n}\tilde{X}_i r_i \end{bmatrix}) \geq c > 0\ \forall\, n \geq C$ and applying a central limit theorem for independent observations such as Theorem 5.11 in White (2001) we obtain:

$$V_{n,ATE_{s,0},0}^{-\frac{1}{2}}\sqrt{n}(\hat{ATE}_{S,0} - ATE_{S,0}) \xrightarrow{d} N(0,1) \tag{H.41}$$

H.7 Proof of Proposition 6

As in the proofs of Propositions 4 and 5, convergence in mean-squared error and the continuous mapping theorem imply:

$$C_n - C_{n,0} = o_p(1), \qquad \Sigma_n - \Sigma_{n,0} = o_p(1) \quad \Sigma_{n,\gamma} - V_{n,\gamma,0} = o_p(1) \quad \Sigma_{n,\alpha\gamma} - V_{n,\alpha\gamma,0} = o_p(1)$$
$$A_{n,1} - A_{n,1,0} = o_p(1) \quad A_{n,2} - A_{n,2,0} = o_p(1) \qquad \Omega_n - \Omega_{n,0} = o_p(1) \qquad V_n - V_{n,\alpha,0} = o_p(1)$$

so that Slutsky's theorem implies:

$$V_n^{-\frac{1}{2}}\sqrt{n}(\begin{bmatrix}\hat{\alpha}_0\\ \hat{\alpha}_1\end{bmatrix} - \begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix}) \xrightarrow{d} N(0, I_2) \tag{H.42}$$

H.8 Proof of Proposition 7

The proof of this proposition follows the same steps as the proof of Proposition 5 albeit with different definitions and dependence being indexed by v_i rather than i.

H.9 Proof of Proposition 8

The proof of this proposition follows the same steps as the proof of Proposition 6 albeit with different definitions and dependence being indexed by v_i rather than i.

I Learning and the Extrapolation Identifying Assumption

In this section we briefly discuss the possibility that farmers in our empirical application do not know exactly what their returns are prior to adopting hybrid seeds for the first time, and learn about their returns as they use the technology. Learning could create a feedback from past shocks to productivity, u_{is} $\forall\, s < t$, to current technology use, x_{it}, if positive (negative)

shocks to productivity while using hybrid seeds are misinterpreted as high (low) returns to using hybrid seeds, invalidating the CRC model (2.4) used in the first step of our estimation approach.

If farmers learn directly about their returns, without confounding their adoption of the new technology with past productivity shocks, learning may not invalidate the extrapolation assumption (2.11). Suppose for simplicity that immediate learning takes place, where before her first adoption, a farmer bases her decision on whether to use hybrid seeds on the rule $x_{it} = 1[\tilde{b}_i \geq c_{it}]$ where $\tilde{b}_i = b_i + \varsigma_i$, where ς_i is measurement error, and c_{it} is the cost of using hybrid seeds, while after having used hybrid seeds at least once she bases her decision on the rule $x_{it} = 1[b_i \geq c_{it}]$. If ς_i and c_{it} are independent of a_i and b_i, this model of selection satisfies the condition (2.15) discussed in section 2.4 in the main text.

More generally, a farmer may base her selection decision on an information set, $\mathcal{I}_{it}$, and the selection rule $x_{it} = 1[E(b_i|\mathcal{I}_{it}) \geq c_{it}]$, as in D'Haultfœuille and Maurel (2013) and references therein. If $\mathcal{I}_{it}$ and c_{it} are independent of a_i conditional on b_i, then the extrapolation identifying assumption (2.11) holds (if the assumption of linearity (2.14) also holds). In section 4.3 in the main text, we allow for part of a farmer's information set to be correlated with her baseline productivity a_i or returns b_i as long as it corresponds to information that is shared by all farmers in a village.

References

Abbring, J. H. and J. J. Heckman (2007): "Chapter 72 Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5145–5303.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Re-

turns to Schooling and Measurement of the Effects of Uncertainty on College Choice*," *International Economic Review*, 44, 361–422.

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

Cunha, F., J. Heckman, and S. Navarro (2005): "The 2004 Hicks Lecture: Separating Uncertainty from Heterogeneity in Life Cycle Earnings," *Oxford Economic Papers*, 57, 191–261.

D'Haultfœuille, X. and A. Maurel (2013): "Inference on an extended Roy model, with an application to schooling decisions in France," *Journal of Econometrics*, 174, 95 – 106.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

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.

Newey, W. K. and F. Windmeijer (2009): "Generalized Method of Moments With Many Weak Moment Conditions," *Econometrica*, 77, 687–719.

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

White, H. (2001): *Asymptotic theory for econometricians*, San Diego: Academic Press.