Responses to the incidental parameter problem

Pua, A.A.Y.

Publication date
2016
Document Version
Final published version

Chapter 2

On IV estimation of a dynamic linear probability model with fixed effects

2.1 Introduction

Many researchers still use the dynamic linear probability model (LPM) with fixed effects when analyzing a panel of binary choices. Several applications of the dynamic LPM with fixed effects can be found in papers published in top journals. Applications include assessing the magnitude of state dependence in female labor force participation (Hyslop, 1999), examining the factors that affect exporting decisions (Bernard and Jensen, 2004), determining the effect of income on transitions in and out of democracy (Acemoglu et al., 2009), and determining how overnight rates affect a bank's decision to provide loans (Jiménez et al., 2014). A more suitable approach, however, is to use limited dependent variable (LDV) models when analyzing discrete choice. Unfortunately, the inclusion of fixed effects creates an incidental parameter problem that complicates the estimation of average marginal effects, especially when the time dimension is small (see the survey by Arellano and Bonhomme (2011)). Resorting to a random effects or correlated random effects approach may require specifying the full distribution of the fixed effects and initial conditions[1] – something that researchers may be unwilling to do because of the lack of specific subject

[1] Typically, only the first two moments of the full distribution are required in the case of linear models. In contrast, nonlinear models would typically require the full distribution because we use this distribution to integrate the fixed effects out of the distribution. There are some approaches that can be thought of as being in the middle of correlated random effects approaches and fixed-effects approaches. A prominent example is using a special regressor to consistently estimate common parameters without imposing a parametric assumption on the distribution of the fixed effects and initial conditions, as proposed by Honoré and Lewbel (2002).

matter knowledge to construct such a distribution. Linear dynamic panel data methods present an alternative that allows for fixed effects, dynamics, predetermined regressors, fewer functional form restrictions, and even allow for heteroscedasticity. Therefore, using methods intended for linear dynamic panel data models seems to be an attractive alternative in this setting.

In contrast, my results provide arguments against a commonly held sentiment among researchers expressed quite forcefully in Angrist and Pischke (2009, p.107):

> The upshot of this discussion is that while a nonlinear model may fit the CEF for LDVs more closely than a linear model, when it comes to marginal effects, this probably matters a little. This optimistic conclusion is not a theorem, but, as in the empirical example here, it seems to be fairly robustly true.
>
> Why, then, should we bother with nonlinear models and marginal effects? One answer is that the marginal effects are easy enough to compute now that they are automated in packages like Stata. But there are a number of decisions to make along the way (e.g., the weighting scheme, derivatives versus finite differences), while OLS is standardized. Nonlinear life also gets considerably more complicated when we work with instrumental variables and panel data. Finally, extra complexity comes into the inference step as well, since we need standard errors for marginal effects.

In this paper, I explain why usual dynamic panel data methods, specifically instrumental variable (IV) estimation, are inappropriate for estimating average marginal effects if the conditional expectation function (CEF) is truly nonlinear. In particular, I show the large-n limit of the Anderson-Hsiao (1981; 1982) IV estimator (henceforth AH) is an average marginal effect but subject to incorrect weighting. Given that the AH estimator is a special case of GMM, estimators in the spirit of Arellano and Bond (1991) may be subject to the same problem. I also show that the effect of this incorrect weighting does not disappear even when T is large. Furthermore, I give examples to show that there are certain parameter configurations and fixed effect distributions for which the large-n limit of the AH estimator is outside the nonparametric bounds derived by Chernozhukov et al. (2013).

Much research has been done on whether using the LPM is suitable. A particularly eye-catching example was provided by Lewbel, Dong, and Yang (2012). They show, in a toy example, that OLS applied to the LPM cannot even get the correct sign of the treatment effect even in the situation where there is just a binary exogenous regressor and a high signal-to-noise ratio. Horrace and Oaxaca (2006) show that the linear predictor for the probability of success should be in $[0,1]$ for all observations for the OLS estimator to be consistent for the regression coefficients because the zero conditional mean assumption does not hold when there are observations (whether in the sample or in the population) that produce success probabilities outside $[0,1]$. On the other hand, Wooldridge (2010) argues that "the case for the LPM is even stronger if most the regressors are discrete and take on only a few values". Problem 15.1 of his book asks the reader to show that we need not worry about success probabilities being outside $[0,1]$ in a saturated model. If we specialize the results in Wooldridge

(2005a) and Murtazashvili and Wooldridge (2008) to the LPM, then they show that fixed-effects estimation applied to the LPM with strictly exogenous regressors can be used to consistently estimate average marginal effects under a specific correlated random coefficients condition.

I organize the rest of the chapter as follows. In Section 2.2, I present an example to show that it is possible to use the LPM to recover an average treatment effect under very special assumptions that researchers are unwilling to make. In Section 2.3, I derive analytically the consequences of not meeting these special assumptions when interest centers on the average marginal effect of state dependence for the cases of $T = 3$ and $T \to \infty$. Next, I examine the practical implications of these results using a numerical example and an empirical application on female labor force participation and fertility in Section 2.4. The last section contains concluding remarks followed by a technical appendix.

2.2 A situation where the LPM is a good idea

Suppose we have a two-period panel binary choice model with a strictly exogenous binary regressor:

$$\Pr(y_{it} = 1|x_i, \alpha_i) = \Pr(y_{it} = 1|x_{it}, \alpha_i) = H(\alpha_i + \beta x_{it}), \tag{2.2.1}$$

where $H : [0,1] \to \mathbb{R}$ is some inverse link function that is increasing, $y_{it} \in \{0,1\}$ and $x_i = (x_{i1}, x_{i2}) = (0,1)$ for all $i = 1, \ldots, n$ and $t = 1, 2$. Assume that for all i, we have $y_{i1} \perp y_{i2} | x_i, \alpha_i$.

The regressor x is a strictly exogenous treatment indicator such that all individuals are treated in the second period but not in the first period. In other words, specification (2.2.1) is basically a before-and-after analysis. In this setting, α_i is an individual-specific fixed effect drawn from some unspecified density $g(\alpha)$.

Suppose one ignores the binary nature of the outcome variable y_{it} and one starts with an LPM with fixed effects, i.e., $y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}$ instead. The within estimator for β, which is equivalent to the first-difference estimator for $T = 2$, is then given by

$$\widehat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{i2} - y_{i1})(x_{i2} - x_{i1})}{\frac{1}{n}\sum_{i=1}^{n}(x_{i2} - x_{i1})^2} = \frac{1}{n}\sum_{i=1}^{n}(y_{i2} - y_{i1})\,\mathbf{1}(y_{i1} + y_{i2} = 1) = \frac{1}{n}(n_{01} - n_{10}),$$

where $\mathbf{1}(\cdot)$ is the indicator function. The second equality follows from the definition of x and the implication that $y_{i2} - y_{i1} = 0$ for all i such that $y_{i1} = y_{i2}$. The third equality follows from defining $n_{ab} = \sum_{i=1}^{n} \mathbf{1}(y_{i1} = a, y_{i2} = b)$ as the number

of observations for which we observe the sequence ab. Thus, only those i for which $y_{i1} \neq y_{i2}$ enter into the calculation of $\widehat{\beta}$.

When we calculate the large-n limit of the within estimator, we have

$$\begin{aligned}
\widehat{\beta} \xrightarrow{p} & \int \Pr(y_{i2}=1|x_i,\alpha)\Pr(y_{i1}=0|x_i,\alpha)\,g(\alpha)\,d\alpha \\
& -\int \Pr(y_{i2}=0|x_i,\alpha)\Pr(y_{i1}=1|x_i,\alpha)\,g(\alpha)\,d\alpha \\
= & \int [(1-H(\alpha))H(\alpha+\beta)-H(\alpha)(1-H(\alpha+\beta))]\,g(\alpha)\,d\alpha \\
= & \int [H(\alpha+\beta)-H(\alpha)]\,g(\alpha)\,d\alpha
\end{aligned}$$

In the situation I have described, the average marginal effect $\Delta = \mathbb{E}[y_{i2}-y_{i1}|x_i=(0,1)]$ can be written as:

$$\begin{aligned}
\Delta &= \mathbb{E}[\mathbb{E}(y_{i2}|x_i=(0,1),\alpha)-\mathbb{E}(y_{i1}|x_i=(0,1),\alpha)] \\
&= \mathbb{E}[\mathbb{E}(y_{i2}|x_{i2}=1,\alpha)-\mathbb{E}(y_{i1}|x_{i1}=0,\alpha)] && (2.2.2) \\
&= \mathbb{E}[H(\alpha+\beta)-H(\alpha)] && (2.2.3)
\end{aligned}$$

Despite the inability of the within estimator to consistently estimate β,[2] the within estimator does coincide with Δ even if the true model is nonlinear. In addition, the sample analog of Δ is exactly the within estimator.

Notice that the result arises because of a lucky coincidence of factors – (a) the strict exogeneity of x (allowing us to obtain (2.2.2)), (b) the independence of α_i and x_i (allowing us to obtain (2.2.3)), and (c) the time homogeneity assumption because H does not depend on time (which follows from (2.2.1)). Despite starting from a fixed effects treatment of α_i, one has no choice but to assume independence of α_i and x_i in order to obtain (2.2.3). This already violates the need to allow for arbitrary correlation between α_i and x_i. It is as if an omniscient Nature did not use the knowledge of α_i to assign a corresponding treatment vector x_i to every unit.

Hahn's (2001) discussion of Angrist (2001) has already pointed out the special conditions under which the within estimator is able to estimate an average treatment effect. In addition, he emphasizes that the simple strategies suggested by Angrist (2001) require knowledge of the "structure of treatment assignment and careful re-expression of the new target parameter". Chernozhukov et al. (2013) also make the same point and further show that the within estimator converges to some weighted average of individual difference of means for a specific subset of the data. They also show that this weighted average is not the average marginal effect of interest.

[2] Incidentally, Chamberlain (2010) shows that β is not even point identified in this example unless H is logistic. The result of Manski (1987a) does not apply here. He shows that β is identified up to scale when one of the strictly exogenous regressors has unbounded support.

Despite all these concerns, researchers still insist on estimating LPMs with fixed effects. One may argue that the example above does not really arise in empirical applications but the example already gives an indication that complicated binary choice models estimated through an LPM are unlikely to produce intended results. In particular, the lucky coincidence of factors mentioned earlier does not hold at all for the dynamic LPM which I discuss next.

2.3 Main results

2.3.1 The case of three time periods

Consider the following specification of a dynamic discrete choice model with fixed effects and no additional regressors:

$$\begin{aligned} \Pr\left(y_{it}=1|y_i^{t-1},\alpha_i\right) &= \Pr\left(y_{it}=1|y_{i,t-1},\alpha_i,y_{i0}\right) \\ &= H\left(\alpha_i+\rho y_{i,t-1}\right),\ i=1,\ldots,n,\ t=1,2,3, \end{aligned} \tag{2.3.1}$$

where y_i^{t-1} is the past history of y, α_i is an individual-specific fixed effect, y_{i0} is an observable initial condition, and $H:\mathbb{R}\to[0,1]$ is some inverse link function. Assume that $(y_{i0},y_{i1},y_{i2},y_{i3},\alpha_i)$ are independently drawn from their joint distribution for all i. I leave the joint density of (α_i,y_{i0}), denoted by f, unspecified. This data generating process satisfies Assumptions 1, 3, 5, and 6 of Chernozhukov et al. (2013).

If H is the logistic function, then ρ can be estimated consistently using conditional logit (Chamberlain, 1985). If H happens to be the standard normal cdf, then ρ is not even point-identified (Honoré and Tamer, 2006).[3] In both these cases, we also cannot point-identify the average marginal effect Δ:

$$\Delta=\int\left[\Pr\left(y_{it}=1|y_{i,t-1}=1,\alpha,y_0\right)-\Pr\left(y_{it}=1|y_{i,t-1}=0,\alpha,y_0\right)\right]f\left(\alpha,y_0\right)\,d\alpha dy_0 \tag{2.3.2}$$

even if we know H but leave the density of (y_{i0},α_i) unspecified. This average marginal effect is of practical interest because it measures the effect of state dependence in the presence of individual-specific unobserved heterogeneity.

Despite these negative results, researchers still insist on using a dynamic LPM on the grounds that linearity still provides a good approximation even if the true H is nonlinear.[4] I use this as a starting point and determine the large-n limit of IV estimators for the dynamic LPM. The linear model researchers have in mind can be expressed as:

$$y_{it}=\alpha_i+\rho y_{i,t-1}+\epsilon_{it},\ i=1,\ldots,n,\ t=1,2,3,$$

[3] Honoré and Tamer (2006) actually show that the sign of ρ is identified for any strictly increasing cdf H and unrestricted distribution of (y_{i0},α_i).

[4] The dynamic LPM is really a special case of (2.3.1), where H is the identity function.

where $\epsilon_{it} = y_{it} - E\left(y_{it}|y_i^{t-1}, \alpha_i\right)$. We now take first-differences to eliminate α_i:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta \epsilon_{it},\ i = 1, \ldots, n,\ t = 2, 3.$$

Because the differenced regressor $\Delta y_{i,t-1}$ is correlated with the differenced error $\Delta \epsilon_{it}$, IV or GMM estimators have been used to estimate ρ. Using lagged differences as instruments, the AH estimator can be written as

$$\widehat{\rho}_{AHd} = \frac{\sum_{i=1}^{n} \Delta y_{i1} \Delta y_{i3}}{\sum_{i=1}^{n} \Delta y_{i1} \Delta y_{i2}}.$$

Because of the binary nature of the sequences $\{(y_{i0}, y_{i1}, y_{i2}, y_{i3}) : i = 1, \ldots, n\}$, it is certainly possible for some of the first differences to be equal to zero. Therefore, there are only certain types of sequences that enter into the expression above. If we enumerate all these 16 possible sequences, we can rewrite the estimator as

$$\widehat{\rho}_{AHd} = \frac{n_{0110} + n_{1001} - n_{1010} - n_{0101}}{n_{0100} + n_{1010} + n_{0101} + n_{1011}},$$

where $n_{abcd} = \sum_{i=1}^{n} \mathbf{1}(y_{i0} = a, y_{i1} = b, y_{i2} = c, y_{i3} = d)$ denotes the number of observations in the data for which we observe the sequence $abcd$.[5]

It can be shown[6] that the large-n limit of $\widehat{\rho}_{AHd}$ is

$$\begin{aligned} \widehat{\rho}_{AHd} \quad &\xrightarrow{p} \quad \frac{\int H(\alpha)(1 - H(\alpha + \rho))(H(\alpha + \rho) - H(\alpha))\, g(\alpha)\, d\alpha}{\int H(\alpha)(1 - H(\alpha + \rho))\, g(\alpha)\, d\alpha} \\ &= \quad \int w_d(\alpha, \rho)(H(\alpha + \rho) - H(\alpha))\, d\alpha \\ &= \quad \int \int w_d(\alpha, \rho)\left[\Pr\left(y_{it} = 1 | y_{i,t-1} = 1, \alpha, y_0\right) - \Pr\left(y_{it} = 1 | y_{i,t-1} = 0, \alpha, y_0\right)\right] d\alpha\, dy_0 \end{aligned} \tag{2.3.3}$$

where

$$w_d(\alpha, \rho) = \frac{H(\alpha)(1 - H(\alpha + \rho))\, g(\alpha)}{\int H(\alpha)(1 - H(\alpha + \rho))\, g(\alpha)\, d\alpha}.$$

Note that the weighting function $w_d(\alpha, \rho)$ depends on the true value of ρ and the

[5] Note that we cannot just drop those sequences for which $y_{i1} + y_{i2} \neq 1$, like in conditional logit. If we do this, the resulting AH estimator becomes

$$\widetilde{\rho}_{AHd} = \frac{-n_{1010} - n_{0101}}{n_{0100} + n_{1010} + n_{0101} + n_{1011}},$$

which is always negative regardless of the sign of ρ or Δ. Observe that identification arguments based on the conditional logit do not necessarily translate to other inverse link functions, including that of the identity function.

[6] A part of the derivation can be found in the appendix.

marginal distribution of the fixed effects $g(\alpha)$. The correct weighting function should have been the joint density of (y_0, α) as in (2.3.2). Therefore, $\widehat{\rho}_{AHd}$ is inconsistent for Δ because of the incorrect weighting of the individual marginal dynamic effect $H(\alpha+\rho)-H(\alpha)$.

It is difficult to give a general indication of whether we overestimate or underestimate Δ, because the results depend on the joint distribution of (y_0, α). If it happens that $\rho = 0$ (so that $\Delta = 0$), then $\widehat{\rho}_{AHd}$ is consistent for Δ.

The analysis above can be extended to the AH estimator which uses levels as the instrument set. It can be shown that this AH estimator has the following form:

$$\begin{aligned}\widehat{\rho}_{AHl} &= \frac{\sum_{i=1}^{n}\sum_{t=2}^{3} y_{i,t-2}\Delta y_{it}}{\sum_{i=1}^{n}\sum_{t=2}^{3} y_{i,t-2}\Delta y_{i,t-1}} \\ &= \frac{n_{0110}-n_{0101}+n_{1110}-n_{1010}+n_{1100}-n_{1011}}{n_{1010}+n_{1000}+n_{1001}+n_{1011}+n_{0100}+n_{1100}+n_{0101}+n_{1101}}.\end{aligned}$$

Calculations similar to (2.3.3) allow us to derive the large-n limit of $\widehat{\rho}_{AHl}$, i.e.

$$\begin{aligned}\widehat{\rho}_{AHl} \xrightarrow{p} & \frac{\int (1-H(\alpha+\rho))(1+H(\alpha+\rho))(H(\alpha+\rho)-H(\alpha))f(\alpha,1)\,d\alpha}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha} \\ & + \frac{\int (1-H(\alpha+\rho))H(\alpha)(H(\alpha+\rho)-H(\alpha))f(\alpha,0)\,d\alpha}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha} \\ = & \int\int w_l(\alpha,\rho,y_0)(H(\alpha+\rho)-H(\alpha))\,dy_0\,d\alpha \\ = & \int\int w_l(\alpha,\rho,y_0)\big[\Pr\big(y_{it}=1|y_{i,t-1}=1,\alpha,y_0\big)-\Pr\big(y_{it}=1|y_{i,t-1}=0,\alpha,y_0\big)\big]\,d\alpha\,dy_0,\end{aligned}$$

where

$$\begin{aligned}w_l(\alpha,\rho,0) &= \frac{(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha}, \\ w_l(\alpha,\rho,1) &= \frac{(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)}{\int [(1-H(\alpha+\rho))(1+H(\alpha+\rho))f(\alpha,1)+(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)]\,d\alpha},\end{aligned}$$

I denote $f(\alpha,0) = \Pr(y_0=0|\alpha)\,g(\alpha)$ and $f(\alpha,1) = \Pr(y_0=1|\alpha)\,g(\alpha)$. Note that the weighting function $w_l(\alpha,\rho,y_0)$ depends on the true value of ρ and the joint distribution of (y_0,α). Once again, we have an incorrect weighting function $w_l(\alpha,\rho,y_0)$ instead of the joint distribution of (y_0,α). As a result, $\widehat{\rho}_{AHl}$ is inconsistent for Δ.[7]

[7] For the case where we have one less time period, i.e. we observe sequences of the form $\{(y_{i0}, y_{i1}, y_{i2}) : i = 1,\ldots,n\}$, the large-$n$ limit of $\widehat{\rho}_{AHl}$ depends only on $f(\alpha,1)$.

I was able to obtain neat analytical expressions because there are no other regressors aside from the lagged dependent variable. However, the results above can be extended to the case where we have a predetermined binary regressor (at the cost of more complicated notation). Furthermore, the results can also be extended to regressors with richer support. But these exercises will also point to the same inconsistency of IV estimators for the average marginal effect.

2.3.2 Large-T case

A natural question to ask is whether the inconsistency results extend to the case where the number of time periods T is large. An intuitive response would be to say that as $T \to \infty$, the fixed effects α_i can be estimated consistently. Therefore, we should be able to estimate average marginal effects consistently. Unfortunately, this intuition may be mistaken.

To address this issue, I use sequential asymptotics where I let $T \to \infty$ and then $n \to \infty$ (see Phillips and Moon (1999)). I first derive the large-T limits of the two AH estimators ($\widehat{\rho}_{AHd}$ and $\widehat{\rho}_{AHl}$) and the first-difference OLS estimator $\widehat{\rho}_{FD}$ for the dynamic LPM:

$$y_{it} = \alpha_i + \rho y_{i,t-1} + \epsilon_{it},\ i = 1,\dots,n,\ t = 1,\dots,T.$$

Recall that these estimators are given by the following expressions:

$$\widehat{\rho}_{AHd} = \frac{\sum_{i=1}^{n}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{it}}{\sum_{i=1}^{n}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{i,t-1}}, \widehat{\rho}_{AHl} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{it}}{\sum_{i=1}^{n}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{i,t-1}}, \widehat{\rho}_{FD} = \frac{\sum_{i=1}^{n}\sum_{t=2}^{T}\Delta y_{it}\Delta y_{i,t-1}}{\sum_{i=1}^{n}\sum_{t=2}^{T}\left(\Delta y_{i,t-1}\right)^2}.$$

It can be shown that as $T \to \infty$,[8]

$$\begin{aligned}
\frac{1}{T}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{it} &\xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)(H(\alpha+\rho)-H(\alpha))\,g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{i,t-1} &\xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)\,g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{it} &\xrightarrow{p} -\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}(H(\alpha+\rho)-H(\alpha))\,g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=2}^{T} y_{i,t-2}\Delta y_{i,t-1} &\xrightarrow{p} -\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}g(\alpha)\,d\alpha,\\
\frac{1}{T}\sum_{t=2}^{T}\Delta y_{it}\Delta y_{i,t-1} &\xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)\,g(\alpha)\,d\alpha,
\end{aligned}$$

[8] Some of the calculations can be found in the Appendix. Note that even with fixed n, the inconsistency is still present.

$$\frac{1}{T}\sum_{t=2}^{T}\left(\Delta y_{i,t-1}\right)^2 \quad \xrightarrow{p} \quad -2\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)} g(\alpha)\, d\alpha.$$

Notice that the limiting quantities above do not depend on i. Therefore, as $n \to \infty$, we must have

$$\widehat{\rho}_{AHd} \quad \xrightarrow{p} \quad \int w_d(\alpha,\rho)(H(\alpha+\rho)-H(\alpha))\, d\alpha, \tag{2.3.4}$$

$$\widehat{\rho}_{AHl} \quad \xrightarrow{p} \quad \int w_l(\alpha,\rho)(H(\alpha+\rho)-H(\alpha))\, d\alpha, \tag{2.3.5}$$

$$\widehat{\rho}_{FD} \quad \xrightarrow{p} \quad \frac{1}{2}\left[1-\int w_l(\alpha,\rho)(H(\alpha+\rho)-H(\alpha))\, d\alpha\right], \tag{2.3.6}$$

where the weighting functions are given by

$$\begin{aligned} w_d(\alpha,\rho) &= \frac{H(\alpha)(1-H(\alpha+\rho))\, g(\alpha)}{\int H(\alpha)(1-H(\alpha+\rho))\, g(\alpha)\, d\alpha}, \\ w_l(\alpha,\rho) &= \frac{\frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)} g(\alpha)}{\int \frac{(1-H(\alpha+\rho))H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)} g(\alpha)\, d\alpha}. \end{aligned}$$

As for the behavior of the fixed effects (FE) estimator in the large-T case, I rely on Proposition 3.1 of Galvao and Kato (2014). In the context I consider, the linear probability model is misspecified and the true model is the nonlinear model (2.3.1). As a result, the conditional mean $\mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right)$ is misspecified as additive and linear when in fact it is nonlinear. Under their assumptions A1 to A3, they show that the FE estimator converges to the following pseudo-true parameter:

$$\beta_0 = \frac{\mathbb{E}\left(\widetilde{y}_{it}\widetilde{y}_{i,t-1}\right)}{\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right)},$$

where $\widetilde{y}_{it} = y_{it} - \mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right)$. Assumption A1 of their paper require that the marginal distribution of $\left(\alpha_i, y_{it}, y_{i,t-1}\right)$ is invariant with respect to (i,t). As a result, the initial condition is drawn from the stationary distribution conditional on α_i. Notice that I did not impose this assumption. In the appendix, I show that this pseudo-true parameter is given by

$$\beta_0 = \frac{\mathbb{E}\left[(H(\alpha+\rho)-H(\alpha))\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)\right]}{\mathbb{E}\left[\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)\right]}, \tag{2.3.7}$$

where the expectations are calculated with respect to the marginal distribution of α.

Clearly, the FE estimator does not converge to the correct average marginal effect and the weighting function is given by

$$w_{FE}(\alpha,\rho)=\frac{\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)}{\mathbb{E}\left[\Pr\left(y_{i,t-1}=1|\alpha\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha\right)\right)\right]}.$$

The result (2.3.6) is very troubling. When $\rho = 0$ (so that the true average marginal effect is 0), $\widehat{\rho}_{FD}$ converges to 0.5, grossly overstating the true Δ. In contrast, the other two AH estimators and the FE estimator are available to consistently estimate Δ when $\rho = 0$ (so that $\Delta = 0$). Unfortunately, for all other values of ρ, these two AH estimators and the FE estimator still cannot consistently estimate the correct Δ because of incorrect weighting in (2.3.4), (2.3.5), and (2.3.7). The appropriate weighting function is now the marginal distribution of the fixed effects $g(\alpha)$, because the effect of the initial condition disappears as $T \to \infty$. Moreover, just as in the fixed-T case considered earlier, it is still not possible to determine the direction of inconsistency. Finally, Chernozhukov et al. (2013) show in their Theorem 4 that the identified set for Δ shrinks to a singleton as $T \to \infty$. Thus, it becomes more likely that the large-T limits in (2.3.4), (2.3.5), and (2.3.6) are outside the identified set.

2.4 Practical implications

Based on the results of the previous section, we should not be using IV estimators for the dynamic LPM. Despite these negative results, the IV estimators are able to estimate a zero average marginal effect, if it were the truth. This observation may allow us to construct a test of the hypothesis that $\Delta = 0$. Unfortunately, this may not be so straightforward since the appropriate standard errors for the AH estimators still depend on the unknown joint distribution of (y_0, α). Although of practical interest, testing the hypothesis $\Delta = 0$ may be infeasible.

To further persuade researchers not to use IV for the dynamic LPM, I adopt the example in Chernozhukov et al. (2013) to show that, even in the simplest of cases, we cannot ignore the distortion brought about by the incorrect weighting function. Chernozhukov et al. (2013) consider a data generating process where H is the standard normal cdf, $y_{i0} \perp \alpha_i$, $\Pr(y_{i0} = 1) = 0.5$, and $T = 3$.

I use four distributions for the fixed effects, as described in Table 2.4.1. The first is the standard normal distribution which is a usual choice in Monte Carlo simulations and serves as a benchmark. The second is a mixture of a standard normal and a normal distribution with mean 2 and variance 0.5^2. This mixture makes it more likely for cross-sectional units to have $y_{it} = 1$ across time. The third is a distribution which favors the LPM because the support of α_i is on a bounded set $(0, 1)$. Finally,

the fourth is a mixture of two normals with negative means. This mixture achieves the opposite effect compared to the second mixture.

Table 2.4.1: Distribution of fixed effects for computations

	$N(0,1)$	$0.5N(0,1)+$ $0.5N(2,0.5)$	$Beta(4,2)$	$0.5N(-2,0.1)+$ $0.5N(-1,1)$
Mean	0	1	0.667	-1.5
Variance	1	1.625	0.032	0.755
Skewness	0	-0.543	-0.468	1.132
Kurtosis	3	2.402	2.625	4.070
Multimodal?	Unimodal	Bimodal	Unimodal	Bimodal

In Figure 2.4.1, I calculate[9] the large-n limits of the AH estimators (in blue for $\widehat{\rho}_{AHd}$ and green for $\widehat{\rho}_{AHl}$) and large-n limits of the nonparametric bounds proposed by Chernozhukov et al. (2013) (in red for the lower bound $\hat{\Delta}_l$ and orange for the upper bound $\hat{\Delta}_u$) evaluated at different values of $\rho \in [-2,2]$. I also calculate the true Δ (in black) using the true distribution of (y_0, α).

Even in the benchmark case where $\alpha_i \sim N(0,1)$, both the large-n limits of the AH estimators are larger than Δ when $\rho > 0$. Further note that when $\rho < -0.5$[10], both these large-n limits are outside the identified set. For $\alpha_i \sim 0.5N(0,1)+0.5N(2,0.5)$, both the large-$n$ limits of the AH estimators nearly coincide and are much larger than Δ even for less persistent state dependence. For $\alpha_i \sim Beta(4,2)$, the large-n limits of the AH estimators are practically the same as Δ and both can be found in the identified set. The key seems to be that the bounded support for the fixed effect, which is $(0,1)$. Finally, the large-n limit of the AH estimator using levels as the instrument set is smaller than Δ for $\alpha_i \sim 0.5N(-2,0.1)+0.5N(-1,1)$.

Although I do not have analytical results for GMM applied to the dynamic LPM, I illustrate why GMM may not be a good idea using the empirical application by Chernozhukov et al. (2013) on female labor force participation and fertility. They estimate the following model using complete longitudinal data on 1587 married women selected from the National Longitudinal Survey of Youth 1979 and observed for three years – 1990, 1992, and 1994:

$$LFP_{it} = \mathbf{1}(\beta \cdot kids_{it} + \alpha_i \geq \epsilon_{it}).$$

The parameter of interest is the average marginal effect of fertility on female labor force participation. The dependent variable is a labor force participation indicator,

[9] A Mathematica notebook containing the calculations is available at `http://andrew-pua.ghost.io`.

[10] Negative state dependence has been found in the literature on scarring effects (see references in Torgovitsky (2015)).

the regressor is a fertility indicator that takes the value 1 if the woman has a child less than 3 years old, and α_i is the individual-specific fixed effect.

Chernozhukov et al. (2013) compute nonparametric bounds for the average marginal effect under the assumption that the fertility indicator is strictly exogenous (called static bounds) and that the average marginal effect is decreasing[11] in the fertility indicator. I replicate their bounds and they can be found in row (2) of Table 2.4.3. I also include static bounds without monotonicity for comparison in row (1).[12] In addition, I compute two other nonparametric bounds with and without the monotonicity assumption under the assumption that the fertility indicator is predetermined (called dynamic bounds) in rows (3) and (4).

Table 2.4.3: Female LFP and fertility ($n = 1587, T = 3$)

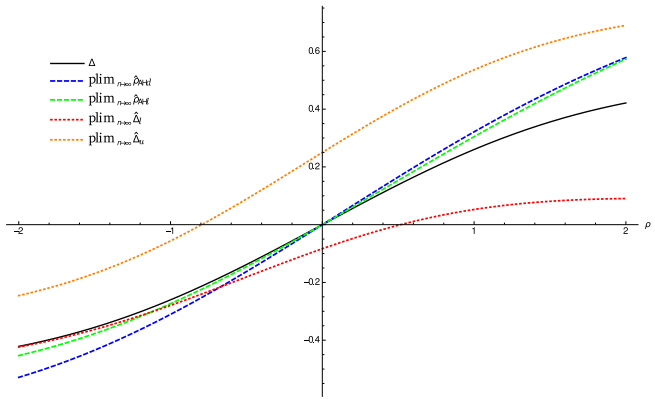
		Avg. Marginal Effect	95% CI
(1)	Static NP bounds	$[-0.40, 0.09]$	
(2)	(1) under monotonicity	$[-0.40, -0.04]$	
(3)	Dynamic NP bounds	$[-0.39, 0.11]$	
(4)	(3) under monotonicity	$[-0.39, -0.19]$	
(5)	Random effects probit	-0.11	$[-0.13, -0.08]$
(6)	Fixed effects OLS	-0.08	$[-0.11, -0.06]$
(7)	First-difference OLS	-0.08	$[-0.09, -0.04]$
(8)	AH (differences)	-0.01	$[-0.14, 0.13]$
(9)	AH (levels)	-0.02	$[-0.07, 0.03]$
(10)	Arellano-Bond	-0.02	$[-0.07, 0.03]$

I also report estimates based on the linear probability model along with the usual 95% asymptotic confidence intervals. Both the fixed effects and first-differenced estimates (rows (6) and (7)) can be found inside the static bounds. This is no longer the case when static bounds are computed under the monotonicity assumption. In contrast, the estimated average marginal effect from the random effects probit (row (5)) is inside the static bounds with or without monotonicity, despite the very incredible assumption where the fixed effects are independent of the fertility indicator. Finally, note that the AH and Arellano-Bond estimates (rows (8) to (10)), which actually assume predeterminedness, are outside the dynamic bounds under monotonicity.

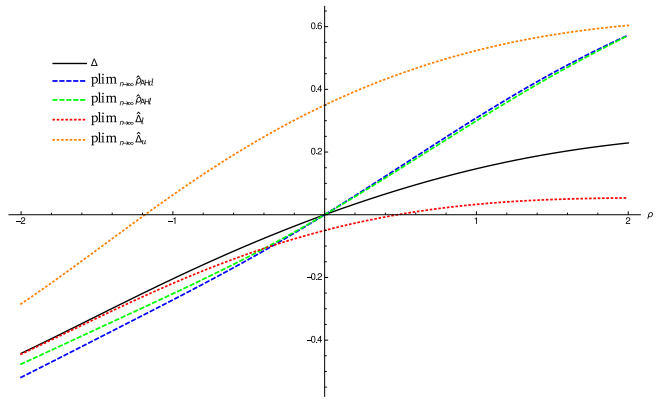
[11] Details as to how to construct the bounds under monotonicity can be found in the Supplemental Material to Chernozhukov et al. (2013).

[12] A Stata do-file is available for replication at `http://andrew-pua.ghost.io`.

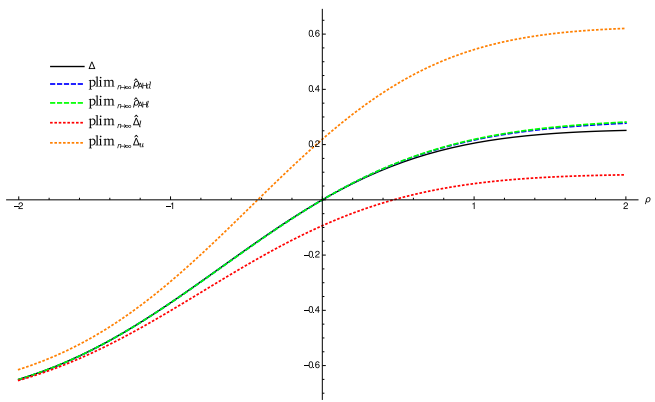
Figure 2.4.1: Large-n limits of the AH estimators under different distributions for the fixed effects



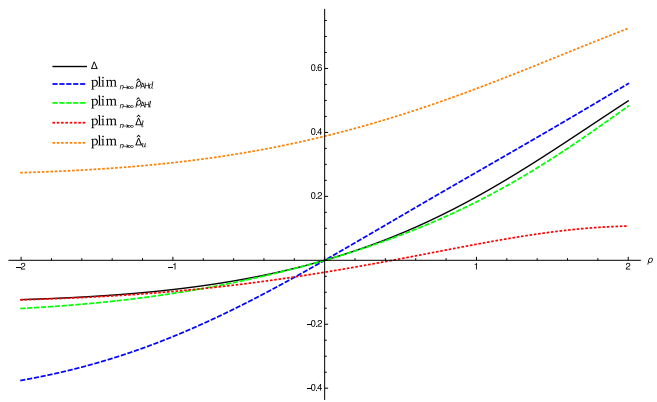
$\alpha_i \sim N(0,1)$



$\alpha_i \sim 0.5N(0,1) + 0.5N(2,0.5)$



$\alpha_i \sim Beta(4,2)$



$\alpha_i \sim 0.5N(-2,0.1) + 0.5N(-1,1)$

2.5 Concluding remarks

I show that using IV methods to estimate the dynamic LPM with fixed effects is inappropriate even in large samples (whether n or T diverge). The analytical results indicate that incorrect weighting of the individual treatment effect is the source of the problem. The numerical results indicate that the estimators may be outside the identified set in both finite and large samples. Therefore, it is more appropriate to use the nonparametric bounds proposed by Chernozhukov et al. (2013), especially if one is unwilling to specify the form for the inverse link function and the joint distribution of the initial conditions and the fixed effects.

The large-n, large-T results I obtain are based on sequential asymptotics. I conjecture that we should obtain similar inconsistency results based on joint asymptotics. It is also unclear whether bias corrections that are derived under large-n, large-T asymptotics can be directly applied to the dynamic LPM with fixed effects. The results in the paper point out that the direction of the asymptotic bias of the estimator for the average marginal effect cannot be obtained. This in stark contrast with the direction of the asymptotic bias derived by Nickell (1981). Although the Monte Carlo experiments of Fernandez-Val (2009) indicate good finite sample performance when we apply the large-T bias corrections, future work should study what exactly these corrections are doing.

It would also be interesting to derive similar analytical results for correlated random effects models so that the results in Wooldridge (2005a) and Murtazashvili and Wooldridge (2008) can be extended to the dynamic case. In the empirical application, I find that the average marginal effect from the usual random effects probit under strict exogeneity can be found in the static nonparametric bounds. Respecting the inherent nonlinearity of a discrete choice model may be responsible for this finding. Future work on this will be of practical interest.

2.6 Appendix

Some calculations for (2.3.3)

We calculate $\mathbb{E}[\mathbf{1}(y_{i0}=0,y_{i1}=1,y_{i2}=1,y_{i3}=0)]$ in detail since the other expressions follow similarly. This expression is equal to

$$\begin{aligned}
& \Pr(y_{i0}=0,y_{i1}=1,y_{i2}=1,y_{i3}=0) \\
= & \int \Pr(y_{i0}=0,y_{i1}=1,y_{i2}=1,y_{i3}=0|\alpha)\, g(\alpha)\, d\alpha \\
= & \int \Pr(y_{i3}=0|y_{i0}=0,y_{i1}=1,y_{i2}=1,\alpha)\Pr(y_{i2}=1|y_{i0}=0,y_{i1}=1,\alpha)\times \\
& \Pr(y_{i1}=1|y_{i0}=0,\alpha)\Pr(y_{i0}=0|\alpha)\, g(\alpha)\, d\alpha
\end{aligned}$$

$$
\begin{aligned}
&= \int \Pr(y_{i3}=0|y_{i2}=1,\alpha)\Pr(y_{i2}=1|y_{i1}=1,\alpha) \\
&\quad \times \Pr(y_{i1}=1|y_{i0}=0,\alpha) f(\alpha,0)\,d\alpha \\
&= \int (1-H(\alpha+\rho))H(\alpha+\rho)H(\alpha)f(\alpha,0)\,d\alpha, \qquad (2.6.1)
\end{aligned}
$$

where f is the joint density of (α, y_0). Similarly, we have the following:

$$
\begin{aligned}
\mathbb{E}[\mathbf{1}(y_{i0}=1,y_{i1}=0,y_{i2}=0,y_{i3}=1)] &= \int H(\alpha)(1-H(\alpha))(1-H(\alpha+\rho))f(\alpha,1)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=1,y_{i1}=0,y_{i2}=1,y_{i3}=0)] &= \int (1-H(\alpha+\rho))H(\alpha)(1-H(\alpha+\rho))f(\alpha,1)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=0,y_{i1}=1,y_{i2}=0,y_{i3}=1)] &= \int H(\alpha)(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=0,y_{i1}=1,y_{i2}=0,y_{i3}=0)] &= \int (1-H(\alpha))(1-H(\alpha+\rho))H(\alpha)f(\alpha,0)\,d\alpha \\
\mathbb{E}[\mathbf{1}(y_{i0}=1,y_{i1}=0,y_{i2}=1,y_{i3}=1)] &= \int H(\alpha+\rho)H(\alpha)(1-H(\alpha+\rho))f(\alpha,1)\,d\alpha
\end{aligned}
$$

Assembling these expressions together in the expression for the large-sample limit of $\widehat{\rho}_{AHd}$ gives (2.3.3).

Some calculations for the large-T case

Note that $\Delta y_{i,t-2}\Delta y_{it} = y_{i,t-2}y_{it} - y_{i,t-3}y_{it} - y_{i,t-2}y_{i,t-1} + y_{i,t-3}y_{i,t-1}$. Observe that the binary nature of y allows us to write

$$
\frac{1}{T}\sum_{t=3}^{T} y_{i,t-2}y_{it} \xrightarrow{p} \lim_{T\to\infty} \frac{1}{T}\sum_{t=3}^{T} \Pr\big(y_{it}=1, y_{i,t-2}=1\big).
$$

By the law of total probability, the definition of conditional probability, and calculations similar to (2.6.1), we are able to express $\Pr\big(y_{it}=1, y_{i,t-2}=1\big)$ as

$$
\begin{aligned}
& \Pr\big(y_{it}=1, y_{i,t-2}=1\big) \\
= \; & \Pr\big(y_{it}=1, y_{i,t-1}=0, y_{i,t-2}=1\big) + \Pr\big(y_{it}=1, y_{i,t-1}=1, y_{i,t-2}=1\big) \\
= \; & \int H(\alpha)(1-H(\alpha+\rho))\Pr\big(y_{i,t-2}=1|\alpha\big)g(\alpha)\,d\alpha \\
& + \int H(\alpha+\rho)^2 \Pr\big(y_{i,t-2}=1|\alpha\big)g(\alpha)\,d\alpha.
\end{aligned}
$$

As a result, we have

$$
\frac{1}{T}\sum_{t=3}^{T} y_{i,t-2}y_{it} \xrightarrow{p} \int \big[H(\alpha+\rho)^2 + H(\alpha)(1-H(\alpha+\rho))\big]\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\big(y_{i,t-2}=1|\alpha\big)\right] g(\alpha)\,d\alpha.
$$

Finally observe that $\Pr\left(y_{i,t-2}=1|\alpha\right)$ obeys a first-order nonhomogeneous difference equation. In particular, note that

$$\begin{aligned}
\Pr(y_{i1}=1|\alpha) &= \Pr(y_{i1}=1|y_{i0}=1,\alpha)\Pr(y_{i0}=1|\alpha) \\
&\quad +\Pr(y_{i1}=1|y_{i0}=0,\alpha)\Pr(y_{i0}=0|\alpha) \\
&= [H(\alpha+\rho)-H(\alpha)]\Pr(y_{i0}=1|\alpha)+H(\alpha) \\
\Pr(y_{i2}=1|\alpha) &= \Pr(y_{i2}=1|y_{i1}=1,\alpha)\Pr(y_{i1}=1|\alpha) \\
&\quad +\Pr(y_{i2}=1|y_{i1}=0,\alpha)\Pr(y_{i1}=0|\alpha) \\
&= [H(\alpha+\rho)-H(\alpha)]\Pr(y_{i1}=1|\alpha)+H(\alpha) \\
&\vdots \\
\Pr(y_{it}=1|\alpha) &= [H(\alpha+\rho)-H(\alpha)]\Pr\left(y_{i,t-1}=1|\alpha\right)+H(\alpha)
\end{aligned}$$

The solution to the above difference equation can be written as

$$\Pr(y_{it}=1|\alpha)=[H(\alpha+\rho)-H(\alpha)]^t\Pr(y_{i0}=1|\alpha)+\sum_{s=0}^{t-1}[H(\alpha+\rho)-H(\alpha)]^s H(\alpha).$$

Note that $|H(\alpha+\rho)-H(\alpha)|<1$. As a result, the effect of the initial condition disappears as $T\to\infty$:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\left(y_{i,t-2}=1|\alpha\right)=\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}.$$

Thus, we have

$$\frac{1}{T}\sum_{t=3}^{T}y_{i,t-2}y_{it}\xrightarrow{p}\int\left[H(\alpha+\rho)^2+H(\alpha)(1-H(\alpha+\rho))\right]\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha.$$

Following similar calculations, we can derive the large-T limits of the other components. In particular,

$$\begin{aligned}
\frac{1}{T}\sum_{t=3}^{T}y_{i,t-2}y_{i,t-1} &\xrightarrow{p} \lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\left(y_{i,t-1}=1,y_{i,t-2}=1\right) \\
&= \int H(\alpha+\rho)\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha. \\
\frac{1}{T}\sum_{t=3}^{T}y_{i,t-3}y_{it} &\xrightarrow{p} \lim_{T\to\infty}\frac{1}{T}\sum_{t=3}^{T}\Pr\left(y_{it}=1,y_{i,t-3}=1\right) \\
&= \int H(\alpha+\rho)^3\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha \\
&\quad +\int 2H(\alpha+\rho)H(\alpha)(1-H(\alpha+\rho))\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha
\end{aligned}$$

$$+\int H(\alpha)(1-H(\alpha))(1-H(\alpha+\rho))\left[\frac{H(\alpha)}{1-H(\alpha+\rho)+H(\alpha)}\right]g(\alpha)\,d\alpha.$$

Observe that the last term

$$\frac{1}{T}\sum_{t=3}^{T} y_{i,t-3}y_{i,t-1}$$

has the same probability limit as

$$\frac{1}{T}\sum_{t=3}^{T} y_{i,t-2}y_{i,t}$$

as $T\to\infty$. Assembling all the results together, we have as $T\to\infty$,

$$\frac{1}{T}\sum_{t=3}^{T}\Delta y_{i,t-2}\Delta y_{it} \xrightarrow{p} -\int (1-H(\alpha+\rho))H(\alpha)(H(\alpha+\rho)-H(\alpha))\,g(\alpha)\,d\alpha.$$

The other large-T results now follow similar computations.

Derivation of the large-n, large-T limit of the fixed effects estimator

Galvao and Kato (2014) impose assumptions A1 to A3 to derive the large-n, large-T limit of the fixed effects estimator. Assumption A1 is about independence across cross-sectional units and a mild form of time series dependence conditional on α_i. For my case, I needed to impose the assumption that the initial condition is drawn from its stationary distribution conditional on α_i, unlike the derivations for the AH estimators.

Let $\widetilde{y}_{it} = y_{it} - \mathbb{E}(y_{it}=1|\alpha_i) = y_{it} - \Pr(y_{it}=1|\alpha_i)$ for $t=1,\ldots,T$. Assumption A2 is about the existence and boundedness of the moments of $\widetilde{y}_{it}$. These moments are guaranteed to exist and be bounded because $\widetilde{y}_{it}$ has a Bernoulli distribution with probability $\Pr(y_{it}=1|\alpha_i)\in(0,1)$. They show that the fixed effects estimator converges to the following pseudo-true parameter:

$$\beta_0 = \frac{\mathbb{E}\left(\widetilde{y}_{it}\widetilde{y}_{i,t-1}\right)}{\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right)}.$$

I now calculate the denominator explicitly. First, note that

$$\begin{aligned}\widetilde{y}_{i,t-1}^2 &= y_{i,t-1}^2 - 2y_{i,t-1}\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2 \\ &= y_{i,t-1} - 2y_{i,t-1}\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2.\end{aligned}$$

Taking expectations, we have

$$\begin{aligned}
\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right) &= \mathbb{E}\Big[y_{i,t-1} - 2y_{i,t-1}\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2\Big] \\
&= \mathbb{E}\Big[\mathbb{E}\left(y_{i,t-1}|\alpha_i\right) - 2\mathbb{E}\left(y_{i,t-1}|\alpha_i\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right) + \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2\Big] \\
&= \mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right) - \left(\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)^2\Big] \\
&= \mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big].
\end{aligned}$$

Note that $\mathbb{E}\left(\widetilde{y}_{i,t-1}^2\right) > 0$ and satisfies assumption A3 of Galvao and Kato (2014). As for the numerator, note that

$$\begin{aligned}
\widetilde{y}_{it}\widetilde{y}_{i,t-1} &= y_{it}y_{i,t-1} - y_{it}\Pr\left(y_{i,t-1}=1|\alpha_i\right) - y_{i,t-1}\Pr\left(y_{it}=1|\alpha_i\right) \\
&\quad + \Pr\left(y_{it}=1|\alpha_i\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right). \qquad (2.6.2)
\end{aligned}$$

Take the first two terms of the right hand side of (2.6.2). Applying law of iterated expectations and $\mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right) = \Pr\left(y_{it}=1|y_{i,t-1},\alpha_i\right)$ gives

$$\begin{aligned}
&\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)y_{it}\right) \\
=\ &\mathbb{E}\Big[\mathbb{E}\left(\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)y_{it}|y_{i,t-1},\alpha_i\right)|\alpha_i\right)\Big] \\
=\ &\mathbb{E}\Big[\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\mathbb{E}\left(y_{it}|y_{i,t-1},\alpha_i\right)|\alpha_i\right)\Big] \\
=\ &\mathbb{E}\Big[\mathbb{E}\left(\left(y_{i,t-1} - \Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)H\left(\alpha_i+\rho y_{i,t-1}\right)|\alpha_i\right)\Big] \\
=\ &\mathbb{E}\Big[\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)H\left(\alpha_i+\rho\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right)\Big] \\
&-\mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right)H\left(\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big].
\end{aligned}$$

The last two terms of the right hand side of (2.6.2) is equal to zero. As a result, we obtain

$$\mathbb{E}\left(\widetilde{y}_{it}\widetilde{y}_{i,t-1}\right) = \mathbb{E}\Big[\left(H\left(\alpha_i+\rho\right)-H\left(\alpha_i\right)\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big].$$

Combining all these findings give us the final form for the pseudo-true parameter:

$$\beta_0 = \frac{\mathbb{E}\Big[\left(H\left(\alpha_i+\rho\right)-H\left(\alpha_i\right)\right)\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big]}{\mathbb{E}\Big[\Pr\left(y_{i,t-1}=1|\alpha_i\right)\left(1-\Pr\left(y_{i,t-1}=1|\alpha_i\right)\right)\Big]}.$$