cemmap
centre for microdata methods and practice

Binarization for panel models with fixed effects

Irene Botosaru
Chris Muris

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP31/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Binarization for Panel Models with Fixed Effects

Irene Botosaru and Chris Muris*

June 17, 2017

Abstract

In nonlinear panel models with fixed effects and fixed-T, the incidental parameter problem poses identification difficulties for structural parameters and partial effects. Existing solutions are model-specific, likelihood-based, impose time homogeneity, or restrict the distribution of unobserved heterogeneity. We provide new identification results for the large class of Fixed Effects Linear Transformation (FELT) models with unknown, time-varying, weakly monotone transformation functions. Our results accommodate continuous and discrete outcomes and covariates, require only two time periods and no parametric distributional assumptions. First, we provide a systematic solution to the incidental parameter problem in FELT via *binarization*, which transforms FELT into many binary choice models. Second, we identify the distribution of counterfactual outcomes and a menu of time-varying partial effects. Third, we obtain new results for nonlinear difference-in-differences with discrete and censored outcomes, and for FELT with random coefficients. Finally, we propose rank- and likelihood-based estimators that achieve $\sqrt{n}$ rate of convergence.

Keywords: panel data, fixed effects, incidental parameter, time-varying transformation model, partial effects, random coefficients, nonlinear difference-in-differences.
JEL classification: C14; C23; C41.

*Emails: ibotosar@sfu.ca, cmuris@sfu.ca. Address: Department of Economics, Simon Fraser University, Burnaby, BC V5A1S6, Canada. We would like to thank Federico Bugni, Sylvain Chabé-Ferret, Raj Chetty, Iván Fernández-Val, Jinyong Hahn, Bo Honoré, Hiro Kasahara, Toru Kitagawa, Vadim Marmer, David Pacini, Krishna Pendakur and Mark Pickup for useful discussions and suggestions, and seminar participants at the 2016 Seattle-Vancouver Econometrics Conference, Queen's University, Bristol University, Toulouse School of Economics, University City London, University of British Columbia, UCLA, UC Davis, for comments and suggestions. We gratefully acknowledge support from SSHRC under grant IDG 430-2015-00073.

1 Introduction

Panel models used in microeconometrics include individual-specific parameters, or *fixed effects,* in order to account for individual-specific unobservables that are correlated with the regressors. If the number of observations per individual is small, then two problems arise in nonlinear panel models. First, common parameters may fail to be point-identified, which is known as the *incidental parameter problem*. Second, even when the common parameters can be identified, interesting partial effects may not be.

This paper addresses both of these issues for a large class of nonlinear panel models with fixed effects and a short time horizon, or *fixed-T*. Leading examples of models nested by our framework are the binary choice model, the ordered choice model, the linear model, the Box-Cox transformation model, some duration models, transformation models with (unknown) monotone transformation functions, and extensions of all these models to time-varying censoring and time-varying link functions. We refer to the class of models studied in this paper as Fixed Effects Linear Transformation (FELT), building on the terminology of Abrevaya (1999, 2000).

Our first contribution is to provide a systematic solution to the incidental parameter problem in the FELT class. We do so via *binarization*, a method that exploits the connection between transformation models and binary choice models. Our results accommodate discrete and continuous outcomes and covariates, eliminating the need for case-by-case identification studies of common (structural) parameters in FELT models. We provide both nonparametric and parametric (logistic) identification results that build on Manski (1987); Chamberlain (2010); Muris (2017). Our approach can be applied to models which currently have no solution, such as models with time-varying censoring or ordered choice models with time-varying link functions.

Our second contribution is to present identification results for the distribution of counterfactual outcomes in models belonging to the FELT class. This leads to the identification of a menu of marginal and partial effects that are time-varying in an unrestricted way. This result is relevant since recent work on partial effects in nonlinear panel models restricts time variation of partial effects by relying on time-homogeneity assumptions, see e.g. Chernozhukov et al. (2013a, 2015).

Our third contribution is to obtain new identification results for the distribution of counterfactual outcomes for the treated in the context of nonlinear difference-in-

differences. Our results are invariant to unknown monotone transformations of the outcome variable, and extend those for the panel data version of the changes-in-changes method of Athey and Imbens (2006) to allow for continuous outcomes with censoring, and for discrete outcomes and fixed effects.

As a fourth contribution, we propose four estimators for the finite- and infinite-dimensional common parameters. Estimation depends on the nature of the outcome (discrete or continuous), and on whether the distribution of the error term is left unspecified (nonparametric) or assumed logistic. Three out of these four estimators are $\sqrt{n}$-consistent and asymptotically normal; the parametric rate is not attained only for the model for discrete outcomes and nonparametric errors. Our estimators for the parameters in the model with continuous outcome and nonparametric errors build on the rank estimators of Abrevaya (1999) and Chen (2002), extending the latter to panel data. For the logistic case, we extend the conditional maximum likelihood estimator in Muris (2017). None of our estimators require smoothing parameters.

Our primary identification results are obtained for the following FELT model:

$$Y_{it} = h_t\left(\alpha_i + X_{it}\beta - U_{it}\right),\ i = 1, \cdots, n,\ t = 1, \cdots, T, \tag{1.1}$$

$$U_{it}|\alpha_i, X_i \overset{d}{=} F\left(u|\alpha_i, X_i\right), \tag{1.2}$$

where Y_{it} is a discrete or continuous scalar dependent variable, α_i is a scalar unobserved fixed effect for individual i, $X_i = (X_{i1}, \cdots, X_{iT})$, where X_{it} is a vector of explanatory variables, β is the vector of corresponding coefficients, and U_{it} is a stochastic error term with (unknown) cumulative distribution function (cdf), $F\left(u|\alpha, X\right)$.[1] Importantly, the functions $(h_t,\ t = 1, \cdots, T)$ are unknown, time-varying, and weakly monotone. In particular, the transformation functions are allowed to have flat parts and jumps.[2]

We first transform FELT into *many* binary choice models, a method which we call binarization. We then show that binarization obtains identification of β and h_t without imposing any parametric assumptions on $F\left(u|\alpha, X\right)$ or on the conditional distribution of the fixed effects given the covariates X_i. Two time periods are sufficient for our results and serial dependence in U_{it} is allowed. These assumptions

[1] The stationarity of the error terms and strict exogeneity of the covariates are standard assumptions in the static nonlinear panel model literature, see e.g. the review by Arellano and Bonhomme (2012).

[2] In the Appendix, we extend our results to FELT with random coefficients.

set binarization apart from other identification strategies that have been applied in the context of nonlinear panel models, such as the functional differencing method of Bonhomme (2012), Kotlarski's lemma, see e.g. Evdokimov (2011), or more general operator diagonalization techniques, see e.g. Freyberger (2012).

Given identification of (β, h_t), we then show that the distribution of counterfactual outcomes at time t is identified *without* knowledge of the conditional distribution of α_i. The distribution of counterfactual outcomes is either point or partially identified depending on whether h_t is invertible. Additionally, we show identification of time-varying partial effects for both continuous and discrete outcomes. These results are surprising for two reasons. First, in nonlinear panel models with fixed-T, knowledge of the structural function is typically not sufficient to identify the effect of counterfactual changes, see e.g. Arellano and Honoré (2001). Second, as mentioned above, identification of partial effects in nonlinear panel models typically relies on time-homogeneity assumptions. The only exception that we are aware of is Chernozhukov et al. (2013a) who allow for time-varying location and scale effects. Their specification, however, is incompatible with outcomes with fixed, discrete support.

Our estimation strategy for (β, h_t) depends on whether $F(u|\alpha, X)$ is nonparametric or logistic, and on whether the outcomes are continuous or discrete. When $F(u|\alpha, X)$ is nonparametric and the outcomes are discrete, a maximum score estimator for the common parameters is pointwise consistent and attains cube root-n rate of convergence, see e.g. Sherman (2010). When $F(u|\alpha, X)$ is nonparametric and the outcomes are continuous, we propose a two-step rank-based estimator. In the first step, we estimate β via the leapfrog procedure of Abrevaya (1999). In the second step, we obtain an estimator for h_t by extending the rank-based estimator of Chen (2002) to the panel data case. Both estimators achieve $\sqrt{n}$ rate of convergence and are asymptotically normal, with the results for the estimator of h_t holding uniformly over compact intervals. When $F(u|\alpha, X)$ is logistic, we extend the conditional maximum likelihood estimator in Muris (2017), and we show that the estimation procedure attains the parametric rate and asymptotic normality of both estimators, whether the outcomes are continuous or discrete. Hence, the logistic estimator is interesting in that it preserves the parametric rate of convergence for both estimators when the outcomes are discrete (as opposed to the nonparametric estimator).

A key assumption driving our results is additivity in the unobservables inside the transformation function. This functional form restriction implies that the marginal

rate of substitution is invariant with respect to unobserved heterogeneity. To alleviate possible concerns about this assumption, we offer the following arguments. First, the imposed structure avoids the curse of dimensionality, which is key for its use in applied work. Second, we show that binarization obtains the same type of identification results when applied to a random coefficient version of FELT, which allows for a cross-sectionally varying marginal rate of substitution. Third, the constant marginal rate of substitution assumption is testable by exploiting over-identification of β. Fourth, we note that in spite of the strong functional form restriction, the model described by (1.1)-(1.2) nests a large class of models used in applied work, and our analysis of FELT provides several new results that are relevant to the literature on transformation models, on nonlinear panel models with fixed effects, and on partial effects (as we explain in the literature review).

The remainder of this paper is organized as follows. Section 2 provides an overview of the literature on transformation models, on the incidental parameter problem in nonlinear panel models with fixed-T, and on partial effects in nonlinear panel models with fixed-T. Section 3 presents our identification results for two non-nested cases: $F(u|\alpha, X)$ is nonparametric and $F(u|\alpha, X)$ is the standard logistic distribution function. Then Section 3.2 provides identification results for the distribution of the counterfactual outcomes and for the partial effects. Subsection 3.3 applies binarization to a nonlinear difference-in-differences model. Section 4 presents four estimators and shows their large sample properties, while Section 5 illustrates the small sample properties of the estimator that we propose for applied work via simulation studies. Finally, Section 6 concludes. All proofs are in Appendix A, Appendix B applies binarization to FELT with random coefficients, and Appendix C elaborates on our rank estimator.

2 Existing literature

Our paper contributes to four active literatures: the literature on identification and estimation of transformation models; the literature on the incidental parameter problem in nonlinear panel models with fixed effects; the literature on partial effects in nonlinear panel models with fixed effects and fixed-T; and the literature on nonlinear difference-in-differences.

2.1 Transformation models

A large literature in statistics and econometrics studies the identification and estimation of *cross-sectional* linear transformation models. The results in this literature rely on strict monotonicity of the transformation function. Important contributions for the cross-sectional case are Horowitz (1996), Chen (2002), Chiappori et al. (2015), and Florens and Sokullu (2016). We contribute to the literature by studying the identification and estimation of a linear transformation model in the *panel data* context. Importantly, we study identification of a linear transformation model with individual fixed effects and a time-varying weakly monotone transformation function, which cannot be studied in a cross-sectional context.

Within the class of linear transformation models with fixed effects, Abrevaya (1999, 2000) studies estimation of the *finite-dimensional* parameter. Abrevaya (1999) considers a time-varying linear transformation model that is similar to ours, but he restricts his analysis to transformation functions that are strictly increasing. He introduces a leapfrog estimator for the finite-dimensional parameter and shows that it is consistent and $\sqrt{n}$-asymptotically normal. Chen (2010) extends the results for the finite-dimensional parameter in Abrevaya (1999) to allow for censoring. Abrevaya (2000) considers a generalized regression model, where the transformation function does not change over time, but that otherwise nests FELT. He shows that a maximum score estimator consistently estimates the finite-dimensional parameter.

In a recent working paper, Pakes and Porter (2016) provide partial identification results for the finite dimensional parameter in a nonlinear transformation model with fixed effects. Their analysis relies critically on the *time-invariance* of the transformation function.

2.2 The incidental parameter problem

Model-specific solutions to the incidental parameter problem are available for some of the models nested by FELT.[3] The focus in the panel literature concerned with the incidental parameter problem has been on consistent estimation of the *finite dimensional* parameter, rather than on the transformation function or on the partial effects. For example, Chamberlain (1985) provides results for some duration mod-

[3]For a discussion of the incidental parameter problem in nonlinear panel models, see e.g. Neyman and Scott (1948) and Lancaster (2000).

els; Manski (1987) discusses the nonparametric binary choice model; Honoré (1992) provides a solution for censored regression; Kyriazidou (1997) for sample selection; Machado (2004) for binomial regression; Chamberlain (2010) for the binary choice logistic model; Magnac (2004) for a generalization of the conditional logit model; Shi et al. (2016) for multinomial choice models; and Muris (2017) for logit ordered choice. For reviews of the literature, see Chamberlain (1984) (Section 3), Honoré (1992), Arellano and Honoré (2001) (Section 4), Arellano (2003), and Arellano and Bonhomme (2012) (Sections 2 and 4).

There are very few solutions to the incidental parameter problem that are not model-specific. Two exceptions that we are aware of are Lancaster (2002) and Bonhomme (2012). Both papers provide likelihood-based solutions. For example, Lancaster (2002) proposes an estimator based on the integrated likelihood after an orthogonal reparametrization of the fixed effects.[4] The functional differencing approach of Bonhomme (2012) requires finding a projection that yields a set of moment conditions for the common parameter that is free of the incidental parameters. The framework in Bonhomme (2012) covers some models that our framework does not, such as models with a dynamic structure. However, it requires a parametric structure, i.e. that the distribution of the error terms and the (time-invariant) transformation function be known up to a finite-dimensional parameter. In contrast, our approach is not likelihood-based, it allows for a time-varying transformation function, and it does not require a parametric structure.[5]

2.3 Partial effects

An emerging literature studies the identification of partial effects in nonlinear panel models with fixed-T. This literature can be divided into two strands, see also Ghanem (2017). Strand 1 bypasses the identification of the structure and focuses on identifying a partial effect of interest. Strand 2 identifies the entire structure, and obtains partial

[4] Arellano and Bonhomme (2009) generalize Lancaster's approach to the case when orthogonal reparametrizations are not available. Consistency of the resulting estimators requires $n, T \to \infty$.

[5] There are other approaches that provide consistent estimators for common parameters in nonlinear panel models, but they either: do not deal with fixed effects, i.e. they restrict the joint distribution of (α_i, X_i) so they consider (correlated) random effects, see e.g. Alvarez and Arellano (2003); or they assume large T, see e.g. Hahn and Kuersteiner (2002), Hahn and Newey (2004), Arellano and Hahn (2007), Arellano and Bonhomme (2009), Fernández-Val (2009), Fernández-Val and Lee (2013), Fernández-Val and Weidner (2016). Compared to these approaches, we impose no restrictions on the distribution of (α_i, X_i) ("fixed effects") and require that $T = 2$ ("small-T").

effects as a result.

Our approach is in between these two strands. We identify some of the structure (the structural function, but not the distribution of the unobservables) and we show that this is sufficient for the identification of the distribution of counterfactual outcomes at time t and, therefore, of a menu of partial effects that are time-varying.

Strand 1. When X is discrete, Chernozhukov et al. (2013a) provide point identification of partial effects for the subpopulation of movers, and partial identification of partial effects for the entire population. When X is continuous, Chernozhukov et al. (2015) provide point identification of marginal effects for stayers. Hoderlein and White (2012) identify local partial effects for stayers. These papers invoke time-homogeneity, which implies that the partial effects are not time-varying.[6]

In contrast, our approach allows for continuous as well as discrete covariates, and obtains partial effects that vary over time in an unrestricted way. We identify the distribution of counterfactual outcomes, obtaining point identification if the transformation functions are invertible and partial identification otherwise.

Strand 2. The papers in this literature obtain identification of partial effects as a byproduct of identifying the entire structure of the model. However, using the terminology of Graham and Powell (2012), the α_i in this literature are correlated random effects (rather than fixed effects).[7]

For example, Altonji and Matzkin (2005) identify the structural function and the local average response in a nonseparable model, assuming strict monotonicity and time-invariance of the transformation function, and an exchangeability condition. Bester and Hansen (2009) identify and estimate marginal effects under an exchangeability restriction similar to that in Altonji and Matzkin (2005). These exchangeability conditions impose restrictions on the distribution of the unobserved heterogeneity that are not required by a fixed effects approach.

Another approach is to use Kotlarski's lemma or operator diagonalization techniques in order to identify the entire structure, see Evdokimov (2011) and Freyberger (2012). The identification strategies based on these techniques are substantially dif-

[6]Chernozhukov et al. (2013a) also consider a slightly weaker assumption on the transformation function that allows for time-varying scale and location effects. Nonetheless, as they note, this weaker assumption is incompatible with outcomes that have a fixed, discrete support.

[7]The analyses in this strand impose conditions on the support of (the distribution of) α_i. In contrast, fixed effects analyses do not impose such restriction, imposing instead assumptions on the transformation functions.

ferent from ours and require different assumptions. For example, at least three time periods are required (at least three for Evdokimov (2011), at least five for Freyberger (2012)); the stochastic errors are required to be serially independent; and α_i is required to either be continuously distributed with full support (when the outcomes are continuous, see Evdokimov (2011)) or have a known number of support points (when the outcomes are discrete, see Freyberger (2012)).

A crucial assumption underlying our analysis, that is not required for the results in any of the papers mentioned in this subsection, is the single-index structure on the latent variable. In return, we require only two time periods, we do not restrict the support of α_i given the covariates, we allow for serial dependence in the error terms, and impose weak restrictions on the transformation function h_t which allow our analysis to apply to both continuous and discrete outcomes. Additionally, we show that identification of the structural function is sufficient for the identification of partial effects. Finally, our identification approach suggests estimators that achieve parametric rates of convergence, and which require no smoothing parameters. For this reason, binarization provides an alternative, complementary approach to identification in nonlinear panel models.

2.4 Nonlinear difference-in-differences

Recent papers on nonlinear difference-in-differences are Athey and Imbens (2006), Bonhomme and Sauder (2011), and D'Haultfoeuille et al. (2015). The results of these papers apply to both cross-sectional and panel data, while ours apply only to panel data. However, the results of Bonhomme and Sauder (2011) and D'Haultfoeuille et al. (2015) apply only to continuous outcomes. Our approach, just as that of Athey and Imbens (2006), allows for both discrete and continuous outcomes. However, the method of Athey and Imbens (2006) does not allow for censored continuous outcomes or for fixed effects when the outcomes are discrete. In contrast, our results allow for these two cases that are relevant to applied work. Finally, just as the changes-in-changes approach, our method identifies the distribution of the counterfactual outcomes of the treated and is invariant to monotone transformations of the outcome variable.

3 Identification

This section considers identification of the common parameters in FELT, (β, h_t), of the distribution of the counterfactual outcomes at time t, and of the resulting partial effects. Section 3.1 presents the identification (β, h_t). We distinguish two non-nested cases: one with nonparametric errors, and one with logistic errors. Section 3.2 discusses the identification of the distribution of counterfactual outcomes in FELT, leading to the identification of a menu of time-varying partial effects. Section 3.3 illustrates the relevance of those results by considering the identification of the distribution of the counterfactual outcomes (and hence of the average treatment effect on the treated) in a nonlinear difference-in-differences setting. Appendix B considers the identification of the common parameters and partial effects in an extension of FELT to random coefficients.

3.1 Common parameters in FELT

Dropping the i subscript, we let $Y = (Y_1, ..., Y_T)'$ and $X = (X_1', ..., X_T')'$. We rewrite the model in (1.1)-(1.2) as the following latent variable model for $t = 1, ..., T$:

$$\begin{aligned} Y_t^* &= \alpha + X_t\beta - U_t, \\ Y_t &= h_t(Y_t^*), \\ U_t|\alpha, X &\sim F_t(u|\alpha, X), \end{aligned} \tag{3.1}$$

where Y_t^* is the latent outcome variable at time t. Denote the supports of Y_t, Y_t^*, X_t by $\mathcal{Y} \subseteq \mathbb{R}$, $\mathcal{Y}^* = \mathbb{R}$, and $\mathcal{X} \subseteq \mathbb{R}^K$, respectively.[8]

Here, we provide sufficient conditions for identification of (β, h_t). We consider two non-nested cases. The first case allows for nonparametric $F_t(u|\alpha, X)$, requiring only that it is stationary. In this case, the idiosyncratic errors are allowed to be serially dependent. The second case assumes that U_t, $t = 1, \cdots, T$, are serially independent, standard logistic. For both cases, we maintain the assumption below:

Assumption 1. *[Weak monotonicity] For each t, the transformation function $h_t : \mathcal{Y}^* \to \mathcal{Y}$ is unknown, non-decreasing, and right continuous.*

[8] The supports may be indexed by t. We omit this index here for the sake of notation.

Define the generalized inverse $h_t^- : \mathcal{Y} \to \mathcal{Y}^*$ as

$$h_t^-(y) \equiv \inf\left\{y^* \in \mathcal{Y}^* : \ y \leq h_t(y^*)\right\}$$

with the convention that $\inf(\emptyset) = \inf(\mathcal{Y})$. Additionally, let $\underline{\mathcal{Y}} \equiv \mathcal{Y} \backslash \inf \mathcal{Y}$. For an arbitrary $y \in \underline{\mathcal{Y}}$, define the binary random variable

$$\begin{aligned} D_t(y) &\equiv 1\{Y_t \geq y\} \\ &= 1\left\{U_t \leq \alpha + X_t\beta - h_t^-(y)\right\}, \end{aligned} \tag{3.2}$$

where the equality follows from specification (3.1) and weak monotonicity. Here, we use $\underline{\mathcal{Y}}$ instead of $\mathcal{Y}$ because $D_t(\inf \mathcal{Y}) = 1$ almost surely for all t.

3.1.1 Identification strategy: binarization

Two time periods are sufficient for our identification results, so we let $T = 2$ in what follows.

For any two points $(y_1, y_2) \in \underline{\mathcal{Y}}^2$, define the following vector of binary variables

$$D(y_1, y_2) \equiv (D_1(y_1), D_2(y_2)).$$

Our identification strategy for (β, h_1, h_2), which we call binarization, is based on the observation that $D(y_1, y_2)$ follows a panel data binary choice model for *any* $(y_1, y_2) \in \underline{\mathcal{Y}}^2$.

The identification proof proceeds in three steps. First, we show identification of β and of $h_2^-(y_2) - h_1^-(y_1)$ for arbitrary $(y_1, y_2) \in \underline{\mathcal{Y}}^2$. In the resulting binary choice model, the difference $h_2^-(y_2) - h_1^-(y_1)$ shows up as the coefficient on the differenced time dummy, while β shows up as the regression coefficient on $X_2 - X_1$. For a given binary choice model, identification of β and of $h_2^-(y_2) - h_1^-(y_1)$ follows Manski (1987) for the nonparametric version of our model, and Chamberlain (2010) for the logistic version.

Second, we show that varying the pair (y_1, y_2) over $\underline{\mathcal{Y}}^2$ obtains identification of

$$\left\{h_2^-(y_2) - h_1^-(y_1), \ (y_1, y_2) \in \underline{\mathcal{Y}}^2\right\}.$$

Third, we show that identification of this set of differences obtains identification

of the functions h_1 and h_2 under a normalization assumption similar to that made in the literature on transformation models.

In summary, we show that FELT can be converted into a collection of binary choice models, which is precisely what allows us to identify the transformation functions. Omitting the fact that FELT can be transformed into *many* binary choice models obtains identification of β only.

3.1.2 Nonparametric errors

In this section, we provide nonparametric identification results for (β, h_1, h_2). Parts of our identification proof build on Manski (1987).

Assumption 2. *[Error terms]*

(i) $F_1(u|\alpha, X) = F_2(u|\alpha, X) \equiv F(u|\alpha, X)$ *for all* (α, X)*;*

(ii) The support of $F(u|\,\alpha, X)$ *is* $\mathbb{R}$ *for all* (α, X)*.*

Assumption 2(i) is a stationarity assumption, requiring that the distribution of the error terms conditional on observed and unobserved heterogeneity be time-invariant. This assumption excludes lagged dependent variables as covariates, but it allows for serial dependence in U_t. Additionally, as noted by e.g. Chamberlain (2010) and Pakes and Porter (2016), this assumption allows for a particular type of heteroskedasticity, which requires that even when $X_1 \neq X_2$, U_1 and U_2 have equal skedasticities. This type of stationarity assumption is common in linear and nonlinear panel models, see e.g. Chernozhukov et al. (2013b), Pakes and Porter (2016), and references therein. Assumption 2(ii) requires full support of the error terms. This assumption is similar to Assumption 1 in Manski (1987). It guarantees that, for any pair $(y_1, y_2) \in \underline{\mathcal{Y}}^2$, the probability of being a *switcher* is positive. In our context, being a switcher refers to the event $D_1(y_1) + D_2(y_2) = 1$, so that Assumption 2 guarantees that $P(D_1(y_1) + D_2(y_2) = 1) > 0$.

Let $\Delta X \equiv X_2 - X_1$ and for an arbitrary pair $(y_1, y_2) \in \underline{\mathcal{Y}}^2$, define

$$\gamma(y_1, y_2) \equiv h_2^-(y_2) - h_1^-(y_1). \tag{3.3}$$

Lemma 1. *Suppose that* (Y, X) *follows the model in* (3.1)*. Let Assumptions 1 and 2 hold. Then for all* $(y_1, y_2) \in \underline{\mathcal{Y}}^2$*,*

$$med(D_2(y_2) - D_1(y_1)\,|X,\, D_1(y_1) + D_2(y_2) = 1) = sgn(\Delta X\beta - \gamma(y_1, y_2)). \tag{3.4}$$

Proof. The proof builds on Manski (1987), and is presented in Appendix A.1. □

Let $W \equiv (\Delta X, -1)'$ and $\theta(y_1, y_2) \equiv (\beta, \gamma(y_1, y_2))$, so that (3.4) can be written as

$$\operatorname{med}(D_2(y_2) - D_1(y_1) | X, D_1(y_1) + D_2(y_2) = 1) = \operatorname{sgn}(W\theta(y_1, y_2)).$$

For identification of $\theta(y_1, y_2)$ we impose the following additional assumptions.

Assumption 3. *[Covariates]*

(i) The distribution of ΔX is such that at least one component of ΔX has positive Lebesgue density on $\mathbb{R}$ conditional on all the other components of ΔX with probability one. The corresponding component of β is non-zero.

(ii) The support of W is not contained in any proper linear subspace of $\mathbb{R}^{K+1}$.

Assumption 3(i) requires that the change in one of the regressors be continuously distributed conditional on the other components. Assumption 3(ii) is a full rank assumption. These assumptions are standard in the binary choice literature concerned with point identification of the parameters.

Assumption 3 resembles Assumption 2 in Manski (1987), the difference being that our assumption concerns W, which includes a constant that captures a time trend. The presence of this constant requires sufficient variation in X_t over time. No linear combination of the components of X_t can equal the time trend.

Assumption 4. *[Normalization-β] For any $(y_1, y_2) \in \underline{\mathcal{Y}}^2$, $\theta(y_1, y_2) \in \Theta = \mathcal{B} \times \mathbb{R}$, where $\mathcal{B} = \{\beta : \beta \in \mathbb{R}^K, \|\beta\| = 1\}$.*

Assumption 4 imposes a normalization on the parameter of interest, namely that the norm of the regression coefficient equals 1. Scale normalizations are standard in the binary choice literature, and are necessary for point identification when the distribution of the error terms is not parametrized. Normalizing β (instead of θ) avoids a normalization that would otherwise depend on the choice of (y_1, y_2). In this way, the scale of β remains constant across different choices of (y_1, y_2). Alternatively, one can normalize the coefficient on the continuous covariate (cf. Assumption 3(i)) to be equal to one.

Theorem 1. *Suppose that (Y, X) follows the model in (3.1), and let the distribution of (Y, X) be observed. Let Assumptions 1, 2, 3, and 4 hold. Then, for an arbitrary pair $(y_1, y_2) \in \underline{\mathcal{Y}}^2$, $\theta(y_1, y_2)$ is identified.*

Proof. The proof proceeds by showing that FELT can be converted into a binary choice model for an arbitrary pair (y_1, y_2), and then builds on Theorem 1 in Manski (1987). See Appendix A.2. □

So far, we have identified the regression coefficient β and the difference in the generalized inverses at arbitrary pairs $(y_1,\ y_2)$. We consider now identification of the functions h_1 and h_2 on $\underline{\mathcal{Y}}$.

Assumption 5. *[Normalization-h_1] For some $y_0 \in \underline{\mathcal{Y}}$, $h_1^-(y_0) = 0$.*

Such a normalization is standard in transformation models, see e.g. Horowitz (1996). Without this normalization, all identification results hold up to $h_1^-(y_0)$. We only normalize the function in the first time period, imposing no restrictions on the function in the second period beyond that of weak monotonicity (cf. Assumption 1).

Theorem 2. *Suppose that (Y, X) follows the model in* (3.1)*, and let the distribution of (Y, X) be observed. Under Assumptions 1, 2, 3, 4, and 5, the transformation functions $h_1(\cdot)$ and $h_2(\cdot)$ are identified.*

Proof. The proof proceeds by identifying the generalized inverses of monotone functions, which obtains identification of the pre-images of h_1 and h_2. This obtains identification of the functions themselves. See Appendix A.3. □

3.1.3 Logistic errors

In this section, we show identification of (β, h_1, h_2) when the error terms are assumed to follow a logistic distribution. The logistic case is not nested by the nonparametric case. In particular, when the errors are logistic, we do not require a continuous regressor. However, we require conditional serial independence of the error terms.[9]

Assumption 6. *[Logit] (i) $F_1(u|\alpha, X) = F_2(u|\alpha, X) = \Lambda(u) = \frac{\exp(u)}{1+\exp(u)}$, and U_1 and U_2 are independent; (ii) $E(W'W)$ is invertible.*

Assumption 6(i) strengthens Assumption 2 by imposing serial independence and specifying a distribution for the error terms. In particular, it specifies that the variance of the error terms is equal to 1, which eliminates the need to normalize β. Assumption 6(ii) is similar in spirit to Assumption 3(ii): it requires sufficient variation in ΔX.

[9]See Chamberlain (2010) and Magnac (2004) for more details about identification under nonparametric versus logistic errors in the panel data binary choice context.

Theorem 3. *Suppose that* (Y, X) *follow the model in* (3.1)*, and let the distribution of* (Y, X) *be observed. Let Assumptions 1 and 6 hold. Then, for an arbitrary pair* $(y_1, y_2) \in \underline{\mathcal{Y}}^2$, $\theta(y_1, y_2)$ *is identified. Additionally, letting Assumption 5 hold, then the transformation functions* $h_1(\cdot)$ *and* $h_2(\cdot)$ *are identified.*

Proof. See Appendix A.4. □

3.2 Partial effects in FELT

Let $Y_t(x)$ represent the counterfactual outcome at time t under the treatment status $X_t = x$,[10] i.e.

$$Y_t(x) \equiv h_t(\alpha + x\beta - U_t), \, t = 1, 2. \tag{3.5}$$

The previous subsections show identification of (β, h_1, h_2). In the present section, we show that this is sufficient for the identification of the distributions of $Y_1(x)$ and $Y_2(x)$ conditional on $X = (X_1, X_2)$. This obtains the identification of a menu of time-varying partial effects.

Let $h_t^+ : \mathcal{Y} \to \mathcal{Y}^*$ be the right inverse of h_t defined as

$$h_t^+(y) \equiv \sup\{y^* \in \mathcal{Y}^* : y \geq h_t(y^*)\}.$$

Furthermore, we extend the domain of h_t to the extended real line, and set $h_t(-\infty) = \inf \mathcal{Y}$ and $h_t(+\infty) = \sup \mathcal{Y}$.

The following corollary shows that bounds on the distribution of the counterfactual outcomes at time t can be obtained from the observed distribution of (Y, X). We consider below the nonparametric case (the results hold trivially for the logistic case).

Corollary 1. *Let the conditions of Theorem 1 hold. Then, for* $s, t \in \{1, 2\}$,

$$\max_s L_s(x, y; \beta, h_s, h_t) \leq P(Y_t(x) \leq y | X) \leq \min_s U_s(x, y; \beta, h_s, h_t),$$

[10] In the terminology of Blundell and Powell (2003), $Y_t(x)$ is the structural function.

where

$$L_s(x, y; \beta, h_s, h_t) \equiv P\left(Y_s \leq h_s\left(h_t^-(y) + (X_s - x)\beta\right) \middle| X\right),$$
$$U_s(x, y; \beta, h_s, h_t) \equiv P\left(Y_s \leq h_s\left(h_t^+(y) + (X_s - x)\beta\right) \middle| X\right),$$

are lower and upper bounds on the distribution of counterfactual outcomes in period t, based on the observed distribution of outcomes in period s.

Proof. For $s, t \in \{1, 2\}$ consider:

$$\begin{aligned}
P(Y_t(x) \leq y | X) &= P(h_t(\alpha + x\beta - U_t) \leq y | X) \\
&= P(h_t(\alpha + x\beta - U_s) \leq y | X) && (3.6) \\
&\geq P\left(\alpha + x\beta - U_s \leq h_t^-(y) \middle| X\right) && (3.7) \\
&= P\left(\alpha + X_s\beta - U_s \leq h_t^-(y) + (X_s - x) \middle| X\right) && (3.8) \\
&= P\left(Y_s \leq h_s\left(h_t^-(y) + (X_s - x)\right) \middle| X\right), && (3.9)
\end{aligned}$$

where 3.6 follows by the stationarity assumption on the stochastic errors; 3.7 follows by weak monotonicity of h_t; 3.8 follows by adding and subtracting $X_s\beta$ on both sides of the inequality; and 3.9 follows by the model specification and weak monotonicity of the transformation functions.

Similarly, using the definition of the right inverse, obtains:

$$\begin{aligned}
P(Y_t(x) \leq y | X) &= P(h_t(\alpha + x\beta - U_t) \leq y | X) \\
&\leq P\left(Y_s \leq h_s\left(h_t^+(y) + (X_s - x)\right) \middle| X\right).
\end{aligned}$$

For $Y_t(x)$, the inequality holds for $s \in \{1, 2\}$, and therefore for the maximum (minimum) of the lower (upper) bounds.

Finally, the bounds are identified because the distribution of (Y, X) is observed, so that the distributions of Y_1 and Y_2 given X are observed. The parameters (β, h_1, h_2) are identified by Theorem 1. □

Remark 1. When h_t is invertible, $h_t^+(y) = h_t^-(y) = h_t^{-1}(y)$ so that the bounds above are equal to each other. This obtains point identification of the distribution of the

counterfactuals. For example, the counterfactual outcome at time $t = 1$ is given by:

$$\begin{aligned} P\left(Y_1\left(x\right) \leq y \middle| X\right) &= P\left(Y_1 \leq h_1\left(h_1^{-1}\left(y\right) + \left(X_1 - x\right)\beta\right) \middle| X\right) \\ &= P\left(Y_2 \leq h_2\left(h_1^{-1}\left(y\right) + \left(X_2 - x\right)\beta\right) \middle| X\right). \end{aligned}$$

This result shows that FELT satisfies rank similarity (for a definition of rank similarity see Chernozhukov and Hansen (2005)).

Remark 2. Once the distribution of the counterfactual outcomes has been identified, the average structural function (ASF) is identified. For example, when h_t is invertible, the ASF is given by

$$ASF_t\left(x\right) \equiv E\left(Y_t\left(x\right)\right),$$

and the marginal effect is given by

$$M_t\left(x, x^{'}\right) \equiv \frac{1}{d\left(x, x^{'}\right)}\left[E\left(Y_t\left(x\right)\right) - E\left(Y_t\left(x^{'}\right)\right)\right]$$

where $d\left(x, x^{'}\right)$ is a measure of distance between x and $x^{'}$. When h_t is assumed to be differentiable, the marginal effect for the k^{th} component of X_t is given by $\frac{\partial}{\partial x_k} E\left(Y_t\left(x\right)\right)$.

3.3 Nonlinear difference-in-differences

In this section, we illustrate the relevance of our identification results by extending them to the case of non-linear difference-in-differences (DID) with heterogeneous treatment effects. We show identification results for both continuous outcomes (with or without censoring) and discrete outcomes. All cases that we consider allow for fixed effects.

Let $t = 1$ correspond to the pre-treatment period, and $t = 2$ to the post-treatment period. As in the standard DID framework, let D_t be a binary random variable indicating treatment status. An individual belongs to one of two groups: the treated group ($D_1 = 0$ and $D_2 = 1$) or the control group ($D_1 = D_2 = 0$). The researcher observes (Y_t, X_t, D_t) in each period t. Define $Y_t\left(0\right)$ as the potential outcome in the absence of treatment and $Y_t\left(1\right)$ as the potential outcome under treatment. Then $Y_t = D_t Y_t\left(1\right) + \left(1 - D_t\right) Y_t\left(0\right)$ for $t = 1, 2$.

We assume that, for each group, the potential outcomes in the absence of treatment follow FELT, i.e.:

$$Y_t(0) = h_t(\alpha + X_t\beta - U_t(0)), \; t = 1, 2, \tag{3.10}$$

$$U_t(0)|X, \alpha \overset{d}{=} F(u|\alpha, X). \tag{3.11}$$

To analyze the effect of the treatment on the treated, it is not necessary to specify a model for $Y_t(1)$.[11]

The parameter of interest is the distribution of the counterfactual outcome for the treated, i.e.

$$\tau(y; X) \equiv P(Y_2(0) \leq y | D_1 = 0, D_2 = 1, X), \tag{3.12}$$

where

$$Y_2(0) = h_2(\alpha + X_2\beta - U_2(0)). \tag{3.13}$$

The following corollary obtains partial identification of $\tau(y; X)$.

Corollary 2. *Let (Y, X, D) satisfy the conditions of Theorem 1 and let the counterfactual outcomes for the treated, $Y_2(0)$, be given by* (3.13)*. Then*

$$P\left(\widetilde{Y_2^l}(0) \leq y \middle| D_1 = 0, D_2 = 1, X\right) \tag{3.14}$$

$$\leq \tau(y; X) \tag{3.15}$$

$$\leq P\left(\widetilde{Y_2^u}(0) \leq y \middle| D_1 = 0, D_2 = 1, X\right), \tag{3.16}$$

where

$$\widetilde{Y_2^l}(0) \equiv h_2\left(h_1^-(Y_1) + (X_2 - X_1)\beta\right),$$
$$\widetilde{Y_2^u}(0) \equiv h_2\left(h_1^+(Y_1) + (X_2 - X_1)\beta\right).$$

Proof. The subpopulation with $D_1 = D_2 = 0$ provides identification of (β, h_1, h_2) by

[11] One particular example of how the treatment could impact outcomes is

$$Y_{it}(1) = h_t(\alpha_i + X_{it}\beta + \gamma_i t - U_{it}(1)),$$

where γ_i is an individual-specific coefficient. This specification mirrors that of $Y_t(1)$ in the standard DID framework, where the treatment has an additive effect on the potential outcome.

Theorem 1, since their observed outcomes follow FELT:

$$\begin{aligned} Y_1 = Y_1(0) &= h_1(\alpha + X_1\beta - U_1(0)) \\ Y_2 = Y_2(0) &= h_2(\alpha + X_2\beta - U_2(0)). \end{aligned}$$

The remainder of this proof is concerned with the identification of the distribution of counterfactual outcomes $Y_2(0)$ in the subpopulation of treated units, i.e. $D_1 = 0, D_2 = 1$. We have that

$$\begin{aligned} &P(Y_2(0) \leq y | X, D_1 = 0, D_2 = 1) \\ &= P(h_2(\alpha + X_2\beta - U_2(0)) \leq y | X, D_1 = 0, D_2 = 1) && (3.17) \\ &= P(h_2(\alpha + X_2\beta - U_1(0)) \leq y | X, D_1 = 0, D_2 = 1) && (3.18) \\ &= P(h_2(\alpha + X_1\beta - U_1(0) + (X_2 - X_1)\beta) \leq y | X, D_1 = 0, D_2 = 1) \\ &\geq P\left(h_2\left(h_1^-(Y_1) + (X_2 - X_1)\beta\right) \leq y | X, D_1 = 0, D_2 = 1\right) && (3.19) \\ &= P\left(\widetilde{Y_2^l}(0) \leq y | X, D_1 = 0, D_2 = 1\right) && (3.20) \end{aligned}$$

where (3.18) follows from the stationarity of $U_t(0)$, $t = 1, 2$, (3.19) follows from weak monotonicity of h_t, $t = 1, 2$. The result in (3.16) can be obtained by similar arguments. □

Remark 3. When the outcomes are continuous and the transformation functions are invertible, $\tau(y; X)$ is point identified and given by

$$\tau(y; X) = P\left(\widetilde{Y}_2(0) \leq y \middle| D_1 = 0, D_2 = 1, X\right) \tag{3.21}$$

where $\widetilde{Y}_2(0) = h_2\left(h_1^{-1}(Y_1) + (X_2 - X_1)\beta\right)$, which is observed given identification of (β, h_1, h_2).

Remark 4. Note that 3.14 obtains identification of the ATT, where

$$\begin{aligned} ATT \equiv &E(Y_2(1) - Y_1(0) | D_1 = 0, D_2 = 1) \\ = &E(Y_2 | D_1 = 0, D_2 = 1) - E(Y_2(0) | D_1 = 0, D_2 = 1). \end{aligned}$$

Since the distribution of X is observed in the treatment group, the bounds of Corollary

2 can be integrated over it, to obtain:

$$E\left(Y_2 \middle| D_1 = 0, D_2 = 1\right) - E\left(\widetilde{Y_2^u}\left(0\right) \middle| D_1 = 0, D_2 = 1\right) \tag{3.22}$$

$$\leq ATT \tag{3.23}$$

$$\leq E\left(Y_2 \middle| D_1 = 0, D_2 = 1\right) - E\left(\widetilde{Y_2^l}\left(0\right) \middle| D_1 = 0, D_2 = 1\right). \tag{3.24}$$

Our bounds on the ATT above are valid for the case of discrete outcomes and fixed effects, as well as for the case of continuous outcomes with or without censoring.

4 Estimation

This section considers estimation of the identified parameters in model (3.1). For arbitrary $(y_1, y_2) \in \mathcal{Y}^2$, denote the true value of the parameter by

$$\theta_0\left(y_1, y_2\right) \equiv \left(\beta_0, h_{1,0}^{-}\left(y_1\right), h_{2,0}^{-}\left(y_2\right)\right).$$

To parallel our identification results, we consider two cases: one where the errors are nonparametric and one where the errors are logistic. For each one of those cases, we distinguish between continuous and discrete outcomes. Hence we introduce four estimators, three of which are shown to be $\sqrt{n}$-consistent and asymptotically normal (the results hold uniformly over compact intervals for the estimator of the transformation function).

When the errors are nonparametric, and the outcome is discrete, a maximum score estimator for $\theta_0\left(y_1, y_2\right)$ is shown to be consistent. It is well-known that the convergence rate of such estimators is slower than $\sqrt{n}$. When the outcome is continuous, we propose a two-step rank estimator for $\theta_0\left(y_1, y_2\right)$. We show that the estimator is $\sqrt{n}$-consistent and asymptotically normal, uniformly over compact intervals.

For the logistic case, we show that conditional maximum likelihood estimation leads to $\sqrt{n}$-consistent and asymptotically normal estimators for discrete and continuous outcomes. We propose estimation via integrated maximum likelihood, extending Baetschmann et al. (2015) and Muris (2017).

For all cases, we assume that a random sample is available:

Assumption 7. *[Random sample] A sample of n independent realizations*

$$\{(Y_i, X_i), i = 1, \cdots, n\}$$

is drawn from the distribution of (Y, X) generated by (1.1)-(1.2).

4.1 Nonparametric errors

4.1.1 Discrete outcomes

Our proposed estimator for $\theta_0(y_1, y_2)$ is the following maximum score estimator:

$$\widehat{\theta}(y_1, y_2) = \text{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (D_{i2}(y_2) - D_{i1}(y_1)) \operatorname{sgn}\{W_i \theta\}.$$

Theorem 4. *Let Assumptions 1, 2, 3, 4, 7, and let the parameter space Θ be compact. Then, as $n \to \infty$,*

$$\widehat{\theta}(y_1, y_2) \xrightarrow{p} \theta_0(y_1, y_2).$$

Proof. The proof consists of showing that the conditions of Theorem 2.1 and Lemma 2.4 in Newey and McFadden (1994) hold. See Appendix A.5. □

Remark 5. From the literature on maximum score estimation, we know that $\widehat{\theta}(y_1, y_2)$ obeys cube root asymptotics, see Kim and Pollard (1990), and that the bootstrap fails, see Abrevaya and Huang (2005). Smoothing the indicator function as in Horowitz (1992) may yield convergence rates that are arbitrarily close to $\sqrt{n}$ at the cost of an additional tuning parameter. For more details, see e.g. Sherman (2010).

4.1.2 Continuous outcomes

When the outcomes are continuous, we assume that the transformation functions are invertible. For this case, we propose a two-step rank estimation procedure. In the first step, we estimate the finite dimensional parameter via the leapfrog procedure in Abrevaya (1999). In the second step, we extend the cross-sectional estimation procedure in Chen (2002) to the panel data case. The estimators are shown to be consistent and $\sqrt{n}$-asymptotically normal. For the estimators of the inverse transformation functions, those properties hold uniformly over compact intervals in their domain $\mathcal{Y}$.

Let $y_t \in \underline{\mathcal{Y}}$ be a generic threshold and, for $(y_1, y_1', y_2, y_2') \in \underline{\mathcal{Y}}^4$ and $i \neq j$, define the following binary random variables:

$$\begin{aligned}\overline{D}_i(y_1, y_2) &\equiv 1\{D_{i1}(y_1) + D_{i2}(y_2) = 1\}, \\ Z_{ij}(y_1, y_1', y_2, y_2') &\equiv \overline{D}_i(y_1, y_2)\,\overline{D}_j(y_1', y_2')\,1\{D_{i2}(y_2) > D_{j2}(y_2')\}.\end{aligned}$$

Conditional on being a switcher, i.e. $\overline{D}_i(y_1, y_2) = 1$, an observation can be of two types: start low and end high ($Y_{i1} < y_1, Y_{i2} \geq y_2$) or start high and end low ($Y_{i1} \geq y_1, Y_{i2} < y_2$). Then $Z_{ij} \neq 1$ if both i, j are switchers and if i ends high and j ends low. This information is captured by the sample criterion function:

$$\begin{aligned}&Q_n(y_1, y_1', y_2, y_2', \beta, \gamma_1, \gamma_2) \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} Z_{ij}(y_1, y_1', y_2, y_2')\,1\{(\Delta X_i - \Delta X_j)\beta > \gamma_2 - \gamma_1\},\end{aligned} \tag{4.1}$$

where $g_t \equiv h_t^{-1}$ and $\gamma_t \equiv g_t(y_t) - g_t\left(y_t'\right)$, $t = 1, 2$.[12]

We propose a two step estimation procedure for $(\beta_0, g_{1,0}, g_{2,0})$, where zero subscripts refer to the true values of the parameters. Given the strict monotonicity of the transformation functions, we first estimate β_0 via the leapfrog procedure in Abrevaya (1999).[13] We then obtain estimators for $(g_{1,0}, g_{2,0})$ by extending the cross-sectional estimation procedure in Chen (2002) to the panel data case, as described below.

Let $y \in \left[\underline{y}, \overline{y}\right] \subset \underline{\mathcal{Y}}$, and let G_t, $t = 1, 2$ be (known) compact intervals such that they contain $g_{t,0}$, i.e. for $\epsilon > 0$,

$$\left[g_{t,0}\left(\underline{y} - \epsilon\right), g_{t,0}(\overline{y} + \epsilon)\right] \subset G_t.$$

[12] Appendix C provides a detailed motivation for this criterion function. The idea is to use information from pairs of observations where one individual leapfrogs (cf. Abrevaya (1999)) a chosen pair of thresholds (y_1, y_2), whereas the other individual stays within a different pair of thresholds $\left(y_1', y_2'\right)$. This can be done for any combination $\left(y_1, y_2, y_1', y_2'\right)$.

[13] That is,

$$\widehat{\beta} = \text{argmax}_{\beta \in \mathcal{B}} M_n(\beta)$$

where

$$M_n(\beta) \equiv \frac{1}{n(n-1)} \sum_{i \neq j} 1\{\Delta X_i \beta > \Delta X_j \beta\}\left(1\{Y_{i1} < Y_{j1}, Y_{i2} > Y_{j2}\} - 1\{Y_{i1} > Y_{j1}, Y_{i2} < Y_{j2}\}\right).$$

Let $\widetilde{y} \in \underline{\mathcal{Y}}$ be fixed, and let $y_0 \in \underline{\mathcal{Y}}$ be as in Assumption 5 so that $g_{1,0}(y_0) = 0$, and further let $g_{2,0}(y_0) = 0$.[14] Then for thresholds $(y_1, y_1', y_2, y_2') = (y, y_0, \widetilde{y}, \widetilde{y})$, the estimator of $g_{1,0}$ is defined as:

$$\begin{aligned}\widehat{g}_1(y) &= \text{argmax}_{g_1 \in G_1} Q_n\left(y, y_0, \widetilde{y}, \widetilde{y}, \widehat{\beta}, g_1, 0\right) \\ &= \text{argmax}_{g_1 \in G_1} \frac{1}{n(n-1)} \sum_{i \neq j} Z_{ij}(y, y_0, \widetilde{y}, \widetilde{y}) \, 1\left\{(\Delta X_i - \Delta X_j)\widehat{\beta} > g_1\right\}.\end{aligned}$$

For thresholds $(y_1, y_1', y_2, y_2') = (\widetilde{y}, \widetilde{y}, y, y_0)$, the estimator of $g_{2,0}$ is defined as:

$$\begin{aligned}\widehat{g}_2(y) &= \text{argmax}_{g_2 \in G_2} Q_n\left(\widetilde{y}, \widetilde{y}, y, y_0, \widehat{\beta}, 0, g_2\right) \\ &= \text{argmax}_{g_2 \in G_2} \frac{1}{n(n-1)} \sum_{i \neq j} Z_{ij}(\widetilde{y}, \widetilde{y}, y, y_0) \, 1\left\{(\Delta X_i - \Delta X_j)\widehat{\beta} > g_2\right\}.\end{aligned}$$

Remark 6. There are a few advantages to the proposed estimation strategy. First, monotonicity of the transformation functions is sufficient for estimation of the finite dimensional parameter in the first step. Second, estimation of the transformation functions does not involve any smoothing parameters (as it would be the case for sieve or kernel estimation). Third, we obtain $\sqrt{n}$ asymptotic normality for all three estimators. However, the disadvantage of this two-step procedure is a possible loss of efficiency.

To derive the asymptotic properties of $\widehat{g}_1$ and $\widehat{g}_2$, define the following functions:

$$\begin{aligned}&\mu_{g_1}\left(y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}, x^{(1)}, x^{(2)}, y, b, g\right) \\ &\equiv \left(1\left\{y^{(1)} \geq y\right\} + 1\left\{y^{(2)} \geq \widetilde{y}\right\}\right)\left(1\left\{y^{(3)} \geq y_0\right\} + 1\left\{y^{(4)} \geq \widetilde{y}\right\}\right) \\ &\times \left(1\left\{y^{(2)} \geq \widetilde{y}\right\} > 1\left\{y^{(4)} \geq \widetilde{y}\right\}\right) 1\left\{\left(x^{(1)} - x^{(2)}\right) b > g\right\}\end{aligned}$$

and

$$\begin{aligned}&\mu_{g_2}\left(y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}, x^{(1)}, x^{(2)}, y, b, g\right) \\ &\equiv \left(1\left\{y^{(1)} \geq \widetilde{y}\right\} + 1\left\{y^{(2)} \geq y\right\}\right)\left(1\left\{y^{(3)} \geq \widetilde{y}\right\} + 1\left\{y^{(4)} \geq y_0\right\}\right) \\ &\times \left(1\left\{y^{(2)} \geq \widetilde{y}\right\} > 1\left\{y^{(4)} \geq y_0\right\}\right) 1\left\{\left(x^{(1)} - x^{(2)}\right) b > g\right\}\end{aligned}$$

[14] This additional normalization is without loss of generality, as discussed in Appendix C.2.

Additionally, for $t = 1, 2$, define:

$$\begin{aligned}
\tau_{g_t}\left(y^{(1)}, y^{(2)}, \Delta x, y, b, g\right) &\equiv E\mu_{g_t}\left(y^{(1)}, y^{(2)}, Y_1, Y_2, \Delta x, \Delta X, y, b, g\right) \\
&\quad + E\mu_{g_t}\left(Y_1, Y_2, y^{(1)}, y^{(2)}, \Delta X, \Delta x, y, b, g\right), \\
R_t(y) &\equiv E\left[\frac{\partial}{\partial g_{t,0}}\tau_{g_t}\left(Y_1, Y_2, \Delta X, y, \beta_0, g_{t,0}(y)\right)\right], \\
V_t(y) &\equiv \frac{1}{2}E\left[\frac{\partial^2}{\partial g_{t,0}^2}\tau_{g_t}\left(Y_1, Y_2, \Delta X, y, \beta_0, g_{t,0}(y)\right)\right], \\
\Omega_t(y) &\equiv E\left[\frac{\partial^2}{\partial\beta\partial g_{t,0}}\tau_{g_t}\left(Y_1, Y_2, \Delta X, y, \beta_0, g_{t,0}(y)\right)\right].
\end{aligned}$$

Consider now the following assumption:

Assumption 8. *(i) The support of $\Delta X \in \mathbb{R}^K$ is not contained in any proper linear subspace of $\mathbb{R}^K$ a.s. The distribution of the first component of ΔX conditional on the other covariates, $\Delta\widetilde{X}$, is absolutely continuous with respect to the Lebesgue measure;*

(ii) The parameter space $\mathcal{B}$ is a compact subset of $\left\{\beta \in \mathbb{R}^K : \beta_{1,0} = 1\right\}$ and the true parameter β_0 is an interior point of $\mathcal{B}$;

(iii) For $t = 1, 2$, the transformation functions $h_{t,0}$ are strictly increasing with $h_{t,0}^{-1}(y_0) \equiv g_{t,0}(y_0) = 0$, $y_0 \in \underline{\mathcal{Y}}$, and $g_{t,0} \in G_t$ where G_t are (known) compact intervals. That is, for $\left[\underline{y}, \overline{y}\right] \subset \underline{\mathcal{Y}}$ and some $\epsilon > 0$,

$$\left[g_{t,0}\left(\underline{y}\right) - \epsilon, g_{t,0}(\overline{y}) + \epsilon\right] \subset G_t;$$

(iv) The density of $Z \equiv \Delta X\beta$ conditional on $\Delta\widetilde{X} = \widetilde{x}$, $p(z|\widetilde{x})$, is twice continuously differentiable with respect to z, with uniformly bounded derivative. The density of U is twice continuously differentiable with a uniformly bounded derivative. The third order moments of $\Delta\widetilde{X}$ are finite;

(v) For $y \in \left[\underline{y}, \overline{y}\right]$ and $t = 1, 2$, $V_t(y) < 0$, $V_t(y)$ is uniformly bounded away from zero, and $R_t(y)$ and $\Omega_t(y)$ are finite;

(vi) The estimator for β_0, $\widehat{\beta}$, has the following asymptotic representation

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) = \frac{1}{\sqrt{n}}\sum\psi\left(Y_{i1}, Y_{i2}, \Delta X_i\right) + o_p(1),$$

where $E\left(\psi\left(Y_1, Y_2, \Delta X\right)\right) = 0$ and $E\left(\psi\left(Y_1, Y_2, \Delta X\right)\psi\left(Y_1, Y_2, \Delta X\right)'\right) < \infty$ under the true distribution of (Y, X).

Theorem 5. *Let assumptions 7 and 8 hold. Then uniformly over* $y \in \left[\underline{y}, \overline{y}\right] \subset \underline{\mathcal{Y}}$ *and* $t = 1, 2$*, as* $n \to \infty$ *we have that*

$$\widehat{g}_t(y) = g_{t,0}(y) + o_p(1) \tag{4.2}$$

and

$$\sqrt{n}\left(\widehat{g}_t(y) - g_{t,0}(y)\right) \xrightarrow{d} N\left(0, E\left[I_{t,y}(Y_1, Y_2, \Delta X) I'_{t,y}(Y_1, Y_2, \Delta X)\right]\right) \tag{4.3}$$

where

$$I_{t,y}(Y_1, Y_2, \Delta X) \equiv -V_t^{-1}(y)\left[R_t(y) + \frac{1}{2}\Omega_t(y)\psi(Y_1, Y_2, \Delta X)\right]. \tag{4.4}$$

Proof. We show the asymptotic properties of $\widehat{g}_t(y)$, $t = 1, 2$, by building on the results in Chen (2002) and Jochmans (2012). For the proof, see Appendix A.6. □

4.2 Logistic errors

In this section, we propose an estimator for the common parameters in the logistic version of FELT and derive its large sample properties.

In Appendix A.4 we show that

$$P\left(D_{i2}(y_2) = 1 | \overline{D}_i(y_1, y_2) = 1, X_i\right) = \Lambda\left(W_i \theta_0(y_1, y_2)\right). \tag{4.5}$$

This suggests that estimation of $\theta_0(y_1, y_2)$ can be based on

$$l_i(\theta, y_1, y_2) \equiv \overline{D}_i(y_1, y_2)\left[D_{i2}(y_2) \ln \Lambda(W_i\theta) + (1 - D_{i2}(y_2)) \ln\left(1 - \Lambda(W_i\theta)\right)\right], \tag{4.6}$$

which is the conditional log-likelihood contribution for individual i associated with 4.5. This contribution is different from zero only if individual i is a switcher, i.e. $\overline{D}_i(y_1, y_2) = 1$.

Define the conditional maximum likelihood estimator as

$$\widehat{\theta}_n(y_1, y_2) = \text{argmax}_{\theta \in \mathbb{R}^{K+1}} \frac{1}{n} \sum_{i=1}^{n} l_i(\theta, y_1, y_2). \tag{4.7}$$

The associated score contribution is the $(K+1)$ vector

$$s_i(\theta, y_1, y_2) = \overline{D}_i(y_1, y_2)\left[D_{i2}(y_2) - \Lambda(W_i\theta)\right]W_i', \tag{4.8}$$

and the associated Hessian contribution is the $(K+1)\times(K+1)$ matrix

$$H_i(\theta, y_1, y_2) = -\overline{D}_i(y_1, y_2)\,\Lambda(W_i\theta)\,(1-\Lambda(W_i\theta))\,W_iW_i'. \tag{4.9}$$

By (4.5), $E\left(D_{i2}(y_2)|\,W_i, \overline{D}_i(y_1, y_2) = 1\right) = \Lambda(W_i\theta_0(y_1, y_2))$ so that the information matrix equality holds:

$$\begin{aligned} J(y_1, y_2) &\equiv E\left[s_i(\theta_0(y_1, y_2), y_1, y_2)\, s_i(\theta_0(y_1, y_2), y_1, y_2)'\right] \\ &= E\left(\overline{D}_i(y_1, y_2)\,\Lambda\left(W_i'\theta_0(y_1, y_2)\right)\left(1-\Lambda\left(W_i'\theta_0(y_1, y_2)\right)\right)W_iW_i'\right) \\ &= -E\left(H_i(\theta_0(y_1, y_2), y_1, y_2)\right). \end{aligned} \tag{4.10}$$

Theorem 6. *Let Assumptions 1, 6, and 7 hold, and let* $(y_1, y_2) \in \underline{\mathcal{Y}}^2$ *. Then* $\widehat{\theta}_n(y_1, y_2) \xrightarrow{p} \theta_0(y_1, y_2)$ *as* $n \to \infty$.

Proof. See Appendix A.7. □

We now establish $\sqrt{n}$-asymptotic normality of $\widehat{\theta}_n(y_1, y_2)$, as well as some intermediate results that are useful for the weak convergence results that we present later on in this section.

Theorem 7. *If the conditions of Theorem 6 hold, then: (i)* $J(y)$ *is positive definite; (ii) the score follows a central limit theorem:*

$$\sqrt{n}s_n(\theta_0(y_1, y_2), y_1, y_2) \xrightarrow{d} \mathcal{N}(0, J(y_1, y_2));$$

and (iii) the conditional maximum likelihood estimator is $\sqrt{n}$*-asymptotically normal,*

$$\sqrt{n}\left(\widehat{\theta}(y_1, y_2) - \theta_0(y_1, y_2)\right) \xrightarrow{d} \mathcal{N}\left(0, J^{-1}(y_1, y_2)\right),$$

as $n \to \infty$.

Proof. See Appendix A.8. □

Consider the joint behavior for a pair of estimators $\left(\widehat{\theta}\left(y_1, y_2\right), \widehat{\theta}\left(y_1^{'}, y_2^{'}\right)\right)$ for two pairs of cut points (y_1, y_2) and $\left(y_1^{'}, y_2^{'}\right)$. Define

$$s_n\left(\theta, y_1, y_2\right) = \frac{1}{n}\sum_i s_i\left(\theta, y_1, y_2\right)$$

and

$$H_n\left(\theta, y_1, y_2\right) = \frac{1}{n}\sum_i H_i\left(\theta, y_1, y_2\right)$$

Stacking the estimators, Taylor-expanding around θ_0, and defining

$$\Xi_n \equiv \begin{bmatrix} H_n\left(\widetilde{\theta}_n\left(y_1, y_2\right), y_1, y_2\right) & 0 \\ 0 & H_n\left(\widetilde{\theta}_n\left(y_1^{'}, y_2^{'}\right), y_1^{'}, y_2^{'}\right) \end{bmatrix}$$

obtains:

$$\begin{aligned} &\sqrt{n}\begin{pmatrix} \widehat{\theta}_n\left(y_1, y_2\right) - \theta_0\left(y_1, y_2\right) \\ \widehat{\theta}_n\left(y_1^{'}, y_2^{'}\right) - \theta_0\left(y_1^{'}, y_2^{'}\right) \end{pmatrix} \\ &= -\Xi_n^{-1}\sqrt{n}\begin{pmatrix} s_n\left(\theta_0\left(y_1, y_2\right), y_1, y_2\right) \\ s_n\left(\theta_0\left(y_1^{'}, y_2^{'}\right), y_1^{'}, y_2^{'}\right) \end{pmatrix} \\ &\xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} J^{-1}\left(y_1, y_2\right) & \Sigma\left(y_1, y_2, y_1^{'}, y_2^{'}\right) \\ \Sigma\left(y_1, y_2, y_1^{'}, y_2^{'}\right) & J^{-1}\left(y_1^{'}, y_2^{'}\right) \end{pmatrix}\right) \end{aligned} \tag{4.11}$$

where, assuming without loss of generality that $y_2 \geq y_2^{'}$,

$$\Sigma\left(y_1, y_2, y_1^{'}, y_2^{'}\right) = J^{-1}\left(y_1, y_2\right)\Omega_0\left(y_1, y_2, y_1^{'}, y_2^{'}\right)J^{-1}\left(y_1^{'}, y_2^{'}\right) \tag{4.12}$$

$$\begin{aligned} \Omega_0\left(y_1, y_2, y_1^{'}, y_2^{'}\right) = E\Big[&\overline{D}_i\left(y_1, y_2\right)\overline{D}_i\left(y_1^{'}, y_2^{'}\right)\Lambda\left(W_i\theta_0\left(y_1, y_2\right)\right) \\ &\times\left(1 - \Lambda\left(W_i\theta_0\left(y_1^{'}, y_2^{'}\right)\right)\right)W_iW_i^{'}\Big]. \end{aligned} \tag{4.13}$$

This provides the covariance function $\Sigma\left(y_1, y_2, y_1^{'}, y_2^{'}\right)$ for the weak convergence result that follows.[15]

[15]Here, we treat the two parameters $\theta_0\left(y_1, y_2\right)$ and $\theta_0\left(y_1^{'}, y_2^{'}\right)$ as if functionally independent. We ignore the linear relationship between $\theta_0\left(y_1, y_2\right)$ and $\theta_0\left(y_1^{'}, y_2^{'}\right)$, e.g. we ignore that their first K components are identical. The dependence between the two parameter vectors can be exploited for statistical and computational efficiency. We discuss this in the next section.

Assumption 9. *(i)* $E\left\|\Delta X_i\right\|^{2+\epsilon} < \infty$ for some $\epsilon > 0$; (ii) the conditional density $f_{Y_t}\left(y|\Delta X_i = x\right), t = 1, 2$, exists, and it is bounded and uniformly continuous in y, uniformly in x over the support of ΔX_i; *(iii)* h_t *is continuous for each* $t = 1, 2$.

We now establish that a functional central limit theorem holds for the conditional maximum likelihood estimator. The result holds over a compact interval $\left[\underline{y}, \overline{y}\right]$, which was introduced in Section 4.1.2.

Theorem 8. *Assume that the conditions for Theorem 7 hold, and let Assumption 9 hold. Then*

$$\sqrt{n}\left(\widehat{\theta}\left(\cdot\right) - \theta\left(\cdot\right)\right) \Rightarrow z\left(\cdot\right) \text{ in } \ell^{\infty}\left(\left[\underline{y}, \overline{y}\right]\right)^2$$

as $n \to \infty$ *where* $z\left(\cdot\right)$ *is a Gaussian process with covariance function* $\Sigma\left(y_1, y_2, y_1', y_2'\right)$.

Proof. See Appendix A.9. □

4.3 Implementation

For applied practice, we recommend using logit estimation via the integrated log-likelihood contribution

$$\widetilde{l}_i\left(\beta, h_2^-\left(\cdot\right), h_1^-\left(\cdot\right)\right) = \int_{\left[\underline{y}, \overline{y}\right]} \int_{\left[\underline{y}, \overline{y}\right]} w\left(y_1, y_2\right) l_i\left(\theta, y_1, y_2\right) dy_1 dy_2 \tag{4.14}$$

where $w\left(y_1, y_2\right)$ is a positive, integrable weight function, e.g. any continuous pdf. A criterion function based on (4.14) integrates the log-likelihood contribution (4.7) over all pairs $y = \left(y_1, y_2\right) \in \left[\underline{y}, \overline{y}\right]^2$.

The first advantage of the integrated objective function is related to the functional dependence of the parameters across different pairs $\left(y_1, y_2\right) \neq \left(y_1', y_2'\right)$. Decomposing $\theta_0\left(y_1, y_2\right) = \left(\beta\left(y_1, y_2\right), \gamma\left(y_1, y_2\right)\right)$, we have $\beta\left(y_1, y_2\right) = \beta\left(y_1', y_2'\right)$ for any choice of thresholds. The integrated objective function (4.14) automatically imposes such relationships.[16]

A second advantage of the integrated likelihood approach is that it has excellent small sample performance. In particular, the simulation results in Baetschmann et al. (2015) and Muris (2017) document excellent performance of the version with equal

[16] An alternative approach would be to use a minimum distance approach. See e.g. Baetschman et al. (2015) and Muris (2017) for a discussion in the context of the time-invariant fixed effects ordered logit model.

weights, $w(y_1, y_2) = 1$, in the context of ordered choice. The integrated likelihood estimator performs well regardless of the number of time periods, or the choice of (number of) pairs (y_1, y_2). Alternatives based on (optimal) minimum distance and GMM have poor performance when the number of pairs is high, or when there are many time periods, see Muris (2017).

A third advantage is ease of implementation (see Baetschmann et al. (2015) for the ordered choice case, and the extension in Muris (2017)). For the present case, we can extend that procedure on an arbitrarily precise grid for (y_1, y_2). We describe this procedure below.

First, duplicate the original data set once for each point on the grid. For each duplicate, compute $\overline{D}_i(y_1, y_2)$ and $D_{i2}(y_2)$. Second, in the subsample with $\overline{D}_i = 1$, run a logit regression of D_{i2} on ΔX_i augmented with a set of dummy variables indicating which grid values (y_1, y_2) were used for that duplicate observation. In the resulting output, the coefficient on ΔX_i is $\widehat{\beta}$. The coefficients on the y_1 dummies are $\widehat{h_1^-}(y_1)$. The coefficients on the y_2 dummies are $\widehat{h_2^-}(y_2)$. All of these steps use only standard software, and require only a few lines of code.

To see that the weak convergence results in Theorem 8 continue to hold, rewrite the scores in (4.8) as:

$$\begin{aligned} s_i(\theta, y_1, y_2) &= \overline{D}_i(y_1, y_2)\left[D_{i2}(y_2) - \Lambda(W_i\theta)\right]W_i' \\ &\equiv t_i\left(\beta, h_2^-, h_1^-, y_1, y_2\right)\begin{pmatrix}\Delta X_i \\ -1\end{pmatrix}. \end{aligned}$$

Then, the set of scores for the integrated likelihood contribution is

$$\frac{\partial \widetilde{l}_i\left(\beta, h_2^-(\cdot), h_1^-(\cdot)\right)}{\partial \beta} = \int_{[\underline{y},\overline{y}]}\int_{[\underline{y},\overline{y}]} w(y_1, y_2)\, t_i\left(\beta, h_2^-, h_1^-, y_1, y_2\right)\Delta X_i dy_1 dy_2, \quad (4.15)$$

$$\frac{\partial \widetilde{l}_i\left(\beta, h_2^-(\cdot), h_1^-(\cdot)\right)}{\partial h_1^-(y_1)} = \int_{[\underline{y},\overline{y}]} w(y_1, y_2)\, t_i\left(\beta, h_2^-, h_1^-, y_1, y_2\right) dy_2, \quad (4.16)$$

$$\frac{\partial \widetilde{l}_i\left(\beta, h_2^-(\cdot), h_1^-(\cdot)\right)}{\partial h_2^-(y_2)} = -\int_{[\underline{y},\overline{y}]} w(y_1, y_2)\, t_i\left(\beta, h_2^-, h_1^-, y_1, y_2\right) dy_1. \quad (4.17)$$

Weak convergence of the Z-estimator process associated with the scores (4.15)-(4.17) follows if the weight function w is positive and integrable on $\left[\underline{y}, \overline{y}\right]^2$. The weight

function is chosen by the researcher, so this condition is easily satisfied. In particular, it will be satisfied for our recommended estimator, which uses $w(y_1, y_2) = 1$. Given Donskerity of the scores in (4.8), a Donsker preservation theorem then applies to the scores in (4.15)-(4.17), given positivity and integrability of the weight function. Then the proof of Theorem 8, applied to the Z-estimator based on (4.15)-(4.17) follows with only a minor modification.

5 Simulation study

This section provides some evidence on the finite sample properties of the integrated likelihood estimator with uniform weights, described in Section 4.3. This is the estimator that we recommend for applied practice given its small sample performance and ease of implementation.

First, we consider estimation of the common parameters in the fixed effects ordered logit model with time-varying link function. We are not aware of any other consistent estimators for the parameters in a model of this type. Second, we consider estimation of the transformation functions and of the ATT in a nonlinear difference-in-differences setting.

5.1 Ordered choice with time-varying link function

Our first simulation results are for ordered logit with a time-varying link function. We consider the logistic version of FELT with transformation function given by:

$$h_t(y^*) = \begin{cases} 1 & \text{if } y^* < \gamma_{1t} \\ 2 & \text{if } \gamma_{1t} \leq y^* < \gamma_{2t} \\ 3 & \text{if } y^* \geq \gamma_{2t} \end{cases}$$

and with the normalization $\gamma_{11} = 0$. The unknown parameters are $(\gamma_{21}, \gamma_{12}, \gamma_{22}, \beta)$.

We generate the univariate regressors from a standard normal, $X_{it} \sim \mathcal{N}(0,1)$ and the unobserved heterogeneity according to $\alpha_i \sim \mathcal{N}(0,1) + \frac{1}{2}(X_{i1} + X_{i2})$. We set $\gamma_{21} = 1$, $\beta = 1$, and vary the remaining parameters $(n, \gamma_{12}, \gamma_{22})$ across designs. We use $S = 1000$ replications for each design.

We consider seven designs. The first four designs vary the sample size across

$n \in \{100, 250, 500, 1000\}$ while keeping $h_2 = h_1$, i.e. an ordered choice model without time-varying link function. For the remaining designs, we use $n = 1000$. The fifth design expands the width of the intervals between categories in period 2 by setting $h_2 = 2h_1$. The sixth design shifts $h_2 = 1 + h_1$. The seventh design combines the location and scale shifts of designs (5) and (6) by setting $h_2 = 2h_1 + 1$.

The results of our simulation study can be found in Table 1. We only report results for our integrated likelihood estimator, as we are not aware of any other consistent estimator for the parameters in the time-varying ordered choice model. The following paragraph describes our findings.

First, the effect of sample size is as expected, see the results across the first four rows. The bias decreases with n. For example, for β, the bias drops from 0.077 to 0.007. Biases are similarly small and decreasing for the other parameters. The RMSE also decreases monotonically in n for all parameters. Second, changing the link function (designs (4)-(7)) does not affect the RMSE. Third, when we change the link function, the bias remains of small order. However, there is some effect of varying the link function on the bias. We conjecture that these differences in bias are due to the frequency of switchers decreasing as γ_{22} increases. For example, the bias for γ_{22} in designs (5) through (7) occurs when γ_{22} is 2 and 3, respectively. In our simulation designs, the distribution of the latent variable is centered at zero and symmetric. The probability $P(Y_{it}^* > \gamma_{22})$ sharply decreases in γ_{22}, thus reducing the number of observations with $(Y_{i1} < 3, Y_{i2} = 3)$. Additional simulations show that $P(Y_{i1} < 3, Y_{i2} = 3)$ drops from 0.23 in design 4 to 0.05 in design 7.

5.2 Nonlinear difference-in-differences

In this section, we considers a simulation study in the context of nonlinear difference-in-differences (the theoretical development can be found in Section 3.3). We consider estimation of the common parameters as well as of the ATT.

The control outcomes are generated by FELT, with

$$\begin{aligned} h_1(y^*) &= y^* \\ h_2(y^*) &= \Phi\left(\frac{y^* - 1}{\sigma}\right), \end{aligned}$$

with $\sigma = 0.5$ unless mentioned otherwise, and all other design parameters are as in

Design		β		γ_{12}		γ_{22}		γ_{21}	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\gamma_{12} = 0, \gamma_{22} = 1$	100	0.0766	0.45	−0.0020	0.43	0.0428	0.46	0.0570	0.31
	250	0.0486	0.26	0.0211	0.27	0.0491	0.27	0.0351	0.18
	500	0.0169	0.18	0.0060	0.19	0.0139	0.18	0.0088	0.12
	1000	0.0066	0.12	−0.0054	0.13	0.0015	0.13	0.0035	0.09
$\gamma_{12} = 0, \gamma_{22} = 2$	1000	0.0060	0.12	−0.0057	0.13	0.0059	0.15	0.0029	0.09
$\gamma_{12} = 1, \gamma_{22} = 2$	1000	0.0046	0.12	0.0010	0.13	0.0052	0.15	0.0029	0.09
$\gamma_{12} = 1, \gamma_{22} = 3$	1000	0.0038	0.12	0.0016	0.14	0.0169	0.19	0.0048	0.09

Table 1: Simulation results for ordered logit with time-varying link function.

the previous subsection (i.e. the generation of α_i, X_{it} and Y_{it}^*). We use a sample size of $n = 500$, $S = 1000$ simulations for each result, and a linear spline with 12 knots at equispaced quantiles of Y_{i1} (for h_1^{-1}) and Y_{i2} (for h_2^{-1}), with uniform weights w.

First, our simulations yield a bias for the regression coefficient of 0.01 and an RMSE of 0.1. The estimated functions are displayed in Figure 5.1. The solid line represents the true function. The dotted line represents the simulated average of the estimated functions. The dashed black lines represent the simulated (point-wise) interquartile range. The estimator captures most of the nonlinearity in h_2. Further increasing the number of evaluation points, or shifting them further towards the edges of the support, would capture even better the nonlinearity in the transformation function.

Second, in order to compute the ATT and to compare our estimator to that of the standard DID, we generate counterfactual outcomes for a treated group according to:

$$h_2^{treat}\left(Y_{it}^*\right) = \Phi\left(\frac{Y_{it}^* + \gamma_i - 1}{\sigma}\right),$$
$$\gamma_i \sim \mathcal{N}\left(1, 1\right).$$

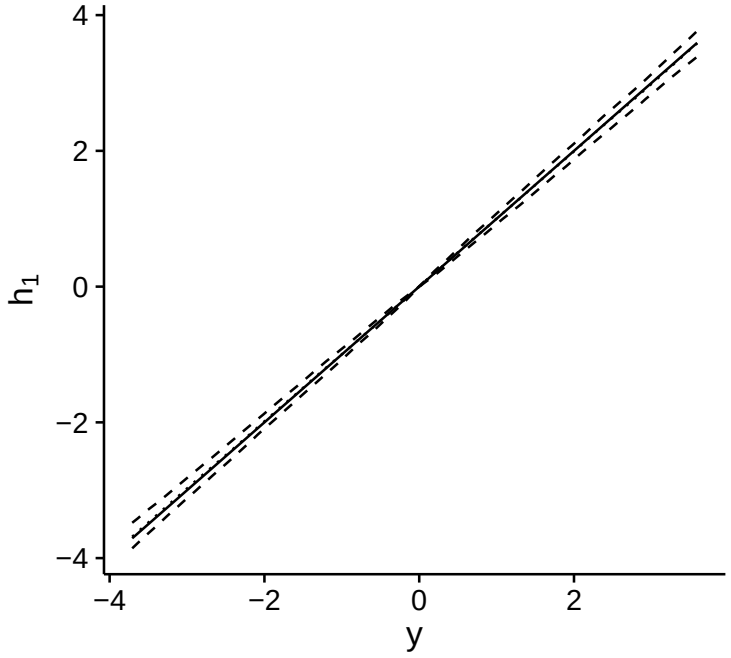
That is, we add a heterogeneous treatment effect linearly, and draw it from a normal distribution with unit mean and standard deviation.[17] The control outcomes for this treatment group are generated in the same way as for the control group, except that we shift the mean unobserved heterogeneity by 1, i.e.

$$\alpha_i \sim N\left(\mu, 1\right) + \frac{1}{2}\left(X_{i1} + X_{i2}\right),$$
$$\mu = 1.$$

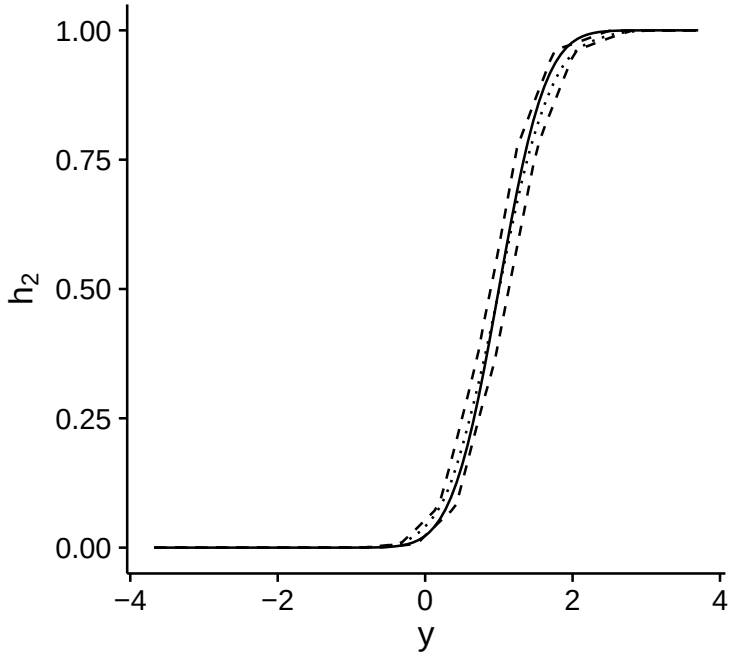
The nonlinearity in h_2 and the location shift in Y_{it}^* between the control and the treatment group pose difficulties for the standard DID estimator. In our design, the true ATT equals $\tau = 0.140$. The standard DID estimator obtains -0.713 while FELT produces a mean ATT estimate of 0.140 across the replications.

Table 2 presents simulation results for the regression coefficient and the ATT in a number of designs. Design (0) is the benchmark design outlined above. Design (1) reduces the number of knots to 6. Design (2) sets $\sigma = 0.25$. Design (3) sets $\mu = 0$.

[17] Consistency for the ATT does not require additivity in the treatment indicator.



h_1



h_2

Figure 5.1: Simulation results for the difference-in-differences study: control outcome equations. True functions are solid lines.The simulated means and pointwise interquartile ranges are reported.

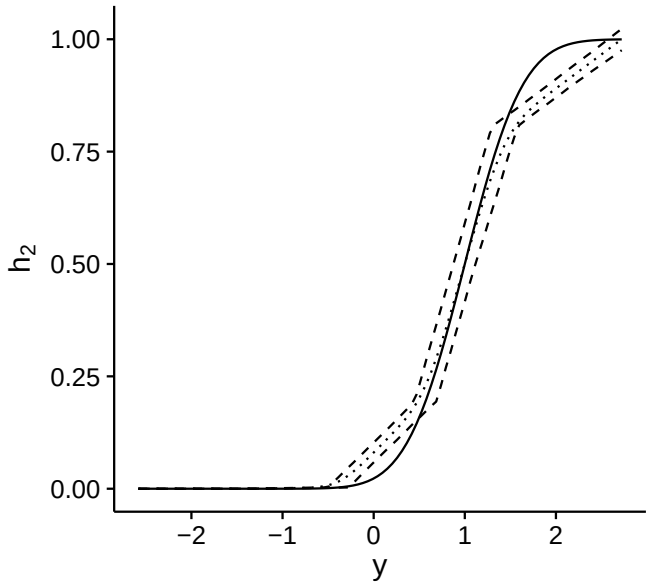


Figure 5.2: Estimate of h_2 in design (1).

Design (4) sets $h_2(y^*) = y^*$.

In design (1), by reducing the number of knots, function h_2 is not approximated as closely, see Figure 5.2. This shows up in Table 2 as an increased bias for the FELT ATT estimator. In design (2), the increased steepness in h_2 does not affect the relative performance substantively. Designs (3) and (4) remove the reasons for the poor performance of the linear DID estimator. In design (3), we impose that the distribution of Y_{i1}^* is the same in treatment and control group, so that the DID estimator correctly estimates the trend despite the nonlinearity. In design (4), we set $h_2 = h_1$ to be linear, so that the DID estimator is consistent. This is reflected in Table 2 by an improvement in the performance of the DID estimator. Even in design (4), the FELT estimator is competitive with the linear DID estimator, although it is slightly outperformed in terms of bias and RMSE.

6 Conclusion

In this paper, we consider identification and estimation in a fixed-T, fixed effects linear transformation (FELT) model, where the transformation function is unknown, weakly monotone, and time-varying. FELT nests a large number of fixed effects panel

Design	β				ATT		
			ATT	DID		**FELT**	
	Bias	RMSE	True	Bias	RMSE	Bias	RMSE
(0)	0.0100	0.10	0.140	−0.850	0.15	0.001	0.03
(1)	0.0114	0.10	0.140	−0.850	0.15	0.058	0.04
(2)	0.0110	0.10	0.142	−0.851	0.15	−0.001	0.03
(3)	0.0156	0.10	0.150	−0.024	0.15	−0.002	0.03
(4)	0.0157	0.10	1.000	−0.015	0.13	−0.039	0.18

Table 2: Simulation results for difference-in-differences: regression coefficient and ATT.

models for discrete and continuous outcomes that are used in applied work, such as binary choice, ordered choice, and various transformation models.

Our approach to identification and estimation, which we call binarization, relies on the relationship between time-varying, weakly monotone transformation models and a collection of binary choice models.

We contribute to the literature on nonlinear panel models in three ways. First, we provide a general solution to the incidental parameter problem for this large class of models. Existing solutions are either model-specific or likelihood-based. Second, we obtain identification of the distribution of counterfactual outcomes, leading to a menu of time-varying partial effects. Current fixed-T results rely on time-homogeneity, which restricts the variability of the partial effects over time. Additionally, we show how our results can be used in a nonlinear difference-in-differences setting to identify the distribution of the counterfactual outcomes for the treated, as well as the ATT. Third, we provide estimators for the parameters of interest and derive their large sample properties. We discuss four estimators, depending on whether the outcome variable is discrete or continuous, and on whether the stationary distribution of the error term is nonparametric or logistic. All estimators are $\sqrt{n}$-consistent and asymptotically normal, except in the nonparametric discrete case.

Our results leave some questions for future research. First, it would be interesting to extend the findings to allow for lagged dependent variables, and for other explanatory variables that do not satisfy the strict exogeneity assumption maintained in this paper. Second, it would be interesting to establish the efficiency bound for the parameters in FELT. Third, we hope that the binarization approach may prove useful for identification in even more general panel models, for example nonlinear latent variable models.

A Proofs

A.1 Proof of Lemma (1)

Proof. Define $\overline{D} = 1\{D_1(y_1) + D_2(y_2) = 1\}$. The proof consists in showing the following:

$$\begin{aligned}
& \operatorname{med}\left(D_2(y_2) - D_1(y_1) \,|X, \overline{D} = 1\right) && \text{(A.1)}\\
= \; & \operatorname{sgn}\left(P\left(D(y_1,y_2) = (0,1) \,|X, \overline{D} = 1\right) - P\left(D(y_1,y_2) = (1,0) \,|X, \overline{D} = 1\right)\right) && \text{(A.2)}\\
= \; & \operatorname{sgn}\left(\frac{P\left(D(y_1,y_2) = (0,1), \overline{D} = 1|X\right)}{P(\overline{D} = 1|X)} - \frac{P\left(D(y_1,y_2) = (1,0), \overline{D} = 1|X\right)}{P(\overline{D} = 1|X)}\right) && \text{(A.3)}\\
= \; & \operatorname{sgn}\left(P\left(D(y_1,y_2) = (0,1), \overline{D} = 1|X\right) - P\left(D(y_1,y_2) = (1,0), \overline{D} = 1|X\right)\right) && \text{(A.4)}\\
= \; & \operatorname{sgn}\left(P\left(D(y_1,y_2) = (0,1) \,|X\right) - P\left(D(y_1,y_2) = (1,0) \,|X\right)\right) && \text{(A.5)}\\
= \; & \operatorname{sgn}\left(P\left(D_2(y_2) = 1|X\right) - P\left(D_1(y_1) = 1|X\right)\right) && \text{(A.6)}\\
= \; & \operatorname{sgn}\left(\Delta X\beta - \gamma(y_1,y_2)\right) && \text{(A.7)}
\end{aligned}$$

where $(A.2)$ follows since the random variable $D_2(y_2) - D_1(y_1) \in \{-1, 1\}$, which implies that

$$\begin{aligned}
& \operatorname{med}\left(D_2(y_2) - D_1(y_1) \,|X, \overline{D} = 1\right)\\
= \; & \begin{cases} 1 \text{ if } P\left(D(y_1,y_2) = (0,1) \,|X, \overline{D} = 1\right) > P\left(D(y_1,y_2) = (1,0) \,|X, \overline{D} = 1\right) \\ -1 \text{ if } P\left(D(y_1,y_2) = (0,1) \,|X, \overline{D} = 1\right) < P\left(D(y_1,y_2) = (1,0) \,|X, \overline{D} = 1\right) \end{cases},
\end{aligned}$$

$(A.3)$ follows from the definition of conditional probability, $(A.4)$ follows since the sign function is not affected by scaling both quantities by the same positive factor (the denominator), $(A.5)$ follows by the definition of $\overline{D}$, and $(A.6)$ follows since:

$$\begin{aligned}
P\left(D_2(y_2) = 1|X\right) &= P\left(D(y_1,y_2) = (0,1) \,|X\right) + P\left(D(y_1,y_2) = (1,1) \,|X\right)\\
P\left(D_1(y_1) = 1|X\right) &= P\left(D(y_1,y_2) = (1,0) \,|X\right) + P\left(D(y_1,y_2) = (1,1) \,|X\right)
\end{aligned}$$

Finally, $(A.7)$ follows from Assumption 2(ii), which implies that e.g.

$$P\left(D_2(y_2) = 1|\alpha, X\right) > P\left(D_1(y_1) = 1|\alpha, X\right) \Leftrightarrow \alpha + X_2\beta - h_2^-(y_2) > \alpha + X_1\beta - h_1^-(y_1).$$

Integrating both sides over the conditional distribution of α given X obtains:

$$\begin{aligned} P\left(D_2\left(y_2\right)=1|X\right) > P\left(D_1\left(y_1\right)=1|X\right) &\Leftrightarrow X_2\beta - h_2^-\left(y_2\right) > X_1\beta - h_1^-\left(y_1\right) \\ &\Leftrightarrow \Delta X\beta - \gamma\left(y_1,y_2\right) > 0. \end{aligned}$$

Result (3.4) now follows. □

A.2 Proof of Theorem 1

Proof. Following Manski (1985), it suffices to show that for an arbitrary $\theta \in \Theta$, $\theta \neq \theta_0 \equiv \theta_0\left(y_1,y_2\right)$,

$$P\left(W\theta < 0 \leq W\theta_0\right) + P\left(W\theta_0 < 0 \leq W\theta\right) > 0. \tag{A.8}$$

Our proof follows very closely that in Manski (1985), with $W\theta$ taking the role of xb and $W\theta_0$ taking the role of $x\beta$. However, our scale normalization is different.

Without loss of generality, let X_K be the continuous regressor in Assumption 3(i). Separate $\Delta X = (\Delta X_{-K}, \Delta X_K)$ where the first component ΔX_{-K} represents all covariates except the K-th one. Similarly, for any $\theta = (\beta,\gamma) \in \Theta$, separate $\beta = (\beta_{-K}, \beta_K)$. Furthermore denote $W_{-K} = (\Delta X_{-K}, -1)$ and $\theta_{-K} = (\beta_{-K}, \gamma)$.

Assume that the associated regression coefficient $\beta_{0,K} > 0$. The case $\beta_{0,K} < 0$ follows similarly. Let $\theta = (\beta,\gamma) \in \Theta$, $\theta \neq \theta_0$. As in Manski (1985, p. 318), consider three cases: (i) $\beta_K < 0$; (ii) $\beta_K = 0$; (iii) $\beta_K > 0$.

Cases (i) and (ii). $\beta_K \leq 0$. The proof is identical to that in Manski, with $X\beta$ replaced by $W\theta$. The fact that we use a different scale normalization does not come into play.

Case (iii). $\beta_K > 0$. note that

$$\begin{aligned} P\left(W\theta < 0 \leq W\theta_0\right) &= P\left(-\frac{W_{-K}\theta_{0,-K}}{\beta_{0,K}} < \Delta X_K < -\frac{W_{-K}\theta_{-K}}{\beta_K}\right). \\ P\left(W\theta_0 < 0 \leq W\theta\right) &= P\left(-\frac{W_{-K}\theta_{-K}}{\beta_K} < \Delta X_K < -\frac{W_{-K}\theta_{0,-K}}{\beta_{0,K}}\right). \end{aligned}$$

By assumption 4, $\frac{\beta_{-K}}{\beta_K} \neq \frac{\beta_{0,-K}}{\beta_{0,K}}$, which shows that the first K components of the vector θ are not a scalar multiple of the first K components of the vector θ_0. Therefore, θ is not a scalar multiple of θ_0. In particular, $\frac{\theta_{0,-K}}{\beta_{0,K}} \neq \frac{\theta_{-K}}{\beta_K}$. Additionally, assumption 3(ii)

implies that $P\left(\frac{W_{-K}\theta_{0,-K}}{\beta_{0,K}} \neq \frac{W_{-K}\theta_{-K}}{\beta_K}\right) > 0$. Hence at least one of the two probabilities above is positive so that (A.8) holds. □

A.3 Proof of Theorem 2

Proof. Under Assumption 5, $h_1^-(y_0) = 0$. Using the pair (y_0, y_2) for binarization thus obtains identification of

$$\begin{aligned}\gamma(y_0, y_2) &= h_2^-(y_2) - h_1^-(y_0) \\ &= h_2^-(y_2).\end{aligned}$$

By varying $y_2 \in \underline{\mathcal{Y}}$, we identify the function h_2^- from the binary choice models associated with $\{D(y_0, y_2) = (D_1(y_0), D_2(y_2)), y_2 \in \underline{\mathcal{Y}}\}$.

The pairs (y_0, y_2) and (y_1, y_2) identify the difference

$$\begin{aligned}\gamma(y_0, y_2) - \gamma(y_1, y_2) &= (h_2^-(y_2) - h_1^-(y_0)) - (h_2^-(y_2) - h_1^-(y_1)) \\ &= h_1^-(y_1).\end{aligned}$$

By varying $y_1 \in \underline{\mathcal{Y}}$ we therefore identify h_1^-.

Thus, the functions h_1^- and h_2^- are identified. Because of monotonicity of h_t (Assumption 1), and because $\mathcal{Y}$ is known, h_t^- contains all the information about the pre-image of h_t. Knowledge of the pre-image of a function is equivalent to knowledge of the function itself. Therefore, h_t can be identified from h_t^-. □

A.4 Proof of Theorem (3)

Proof. For the panel data binary choice model with logit errors, we obtain

$$P(D_2(y_2) = 1|\overline{D}(y_1, y_2) = 1, X, \alpha) \tag{A.9}$$

$$= \frac{P(D_2(y_2) = 1, \overline{D}(y_1, y_2) = 1|X, \alpha)}{P\left(\overline{D}(y_1, y_2) = 1|X, \alpha\right)} \tag{A.10}$$

$$= \frac{P(D_1(y_1) = 0, D_2(y_2) = 1|X, \alpha)}{P\left(\overline{D}(y_1, y_2) = 1|X, \alpha\right)} \tag{A.11}$$

$$= \frac{P(D_1(y_1) = 0, D_2(y_2) = 1|X, \alpha)}{P(D_1(y_1) = 0, D_2(y_2) = 1|X, \alpha) + P(D_1(y_1) = 1, D_2(y_2) = 0|X, \alpha)} \tag{A.12}$$

$$= \frac{1}{1 + \frac{P(D_1(y_1)=1, D_2(y_2)=0|X,\alpha)}{P(D_1(y_1)=0, D_2(y_2)=1|X,\alpha)}} \tag{A.13}$$

$$= \Lambda(\Delta X\beta - \gamma(y_1, y_2)) \tag{A.14}$$

where A.10 follows from the definition of a conditional probability; A.11 follows because $D_2 = 1$ and $\bar{D} = 1$ are equivalent to $D_1 = 0$ and $D_2 = 1$; A.12 follows because $D_1 + D_2 = 1$ happens precisely when either $(D_1, D_2) = (1, 0)$ or $(D_1, D_2) = (0, 1)$; A.13 follows by dividing by the numerator; and the final expression follows by the argument below.

Note that $\frac{P(D_1(y_1)=1, D_2(y_2)=0|X,\alpha)}{P(D_1(y_1)=0, D_2(y_2)=1|X,\alpha)}$ equals

$$\frac{P(D_1(y_1) = 1|X, \alpha)P(D_2(y_2) = 0|X, \alpha)}{P(D_1(y_1) = 0|X, \alpha)P(D_2(y_2) = 1|X, \alpha)} \tag{A.15}$$

$$= \frac{\Lambda\left(\alpha + X_1\beta - h_1^-(y_1)\right)\left[1 - \Lambda\left(\alpha + X_2\beta - h_2^-(y_2)\right)\right]}{\left[1 - \Lambda\left(\alpha + X_1\beta - h_1^-(y_1)\right)\right]\Lambda\left(\alpha + X_2\beta - h_2^-(y_2)\right)} \tag{A.16}$$

$$= \frac{\exp\left(\alpha + X_1\beta - h_1^-(y_1)\right)}{\exp\left(\alpha + X_2\beta - h_2^-(y_2)\right)} \tag{A.17}$$

$$= \exp\left((X_1 - X_2)\beta - \left(h_1^-(y_1) - h_2^-(y_2)\right)\right), \tag{A.18}$$

where A.15 follows from serial independence of (U_1, U_2) conditional on (X, α); A.16 from the logit model specification; and A.17 follows from

$$\Lambda(u) / (1 - \Lambda(u)) = \exp(u).$$

The discussion above implies that A.9 does not depend on α. Hence,

$$\begin{aligned} p\left(X, y_1, y_2\right) &\equiv P(D_2\left(y_2\right) = 1 | \overline{D}\left(y_1, y_2\right) = 1, X) \\ &= \Lambda(\Delta X\beta - \gamma(y_1, y_2)) \\ &= \Lambda\left(W\theta\left(y_1, y_2\right)\right). \end{aligned}$$

and note that $p\left(X, y_1, y_2\right)$ is identified from the distribution of (Y, X), which is assumed to be observed. Then

$$\theta(y_1, y_2) = [E(W'W)]^{-1} E(W'\Lambda^{-1}(p(X, y_1, y_2)))$$

by invertibility of Λ and the full rank assumption on $E\left[W'W\right]$. This establishes identification of β and $\gamma\left(y_1, y_2\right)$. The proof in Section A.3 applies, which shows the identification of h_1 and h_2. □

A.5 Proof of Theorem 4

Proof. Our proof consists of checking the conditions of Theorem 2.1 in Newey and McFadden (1994). That theorem requires four conditions: (i) identification; (ii) compactness of the parameter space; (iii) continuity of the population objective function; (iv) uniform convergence of the sample criterion function to the population objective function.

(i) Identification. The population objective function is defined as the limit of our sample criterion function $Q_n\left(\theta\right)$, and is given by

$$Q_0(\theta) = E\left(\operatorname{sgn}\left(W\theta\right)\left(D_2\left(y_2\right) - D_1\left(y_1\right)\right)\right). \tag{A.19}$$

This function achieves its unique maximum on Θ, as shown in Manski (1987), Lemma 3.

(ii) Compactness. Implied by Assumption 4 and compactness of Γ, which is assumed in the statement of the theorem.

(iii)+(iv) Continuity and uniform convergence. These conditions follow if the conditions of Lemma 2.4 in Newey and McFadden are verified. Define

$$a\left(D_2, D_1, W, \theta\right) = \operatorname{sgn}\left(W\theta\right)\left(D_2\left(y_2\right) - D_1\left(y_1\right)\right).$$

The Lemma requires (a) compactness of Θ; (b) continuity of $a(\cdot)$ in θ with probability 1; (c) dominance of $a(\cdot)$ by an integrable function. (a) follows by our assumptions, see (ii) above. To see that (b) holds, note that $\text{sgn}(W\theta)$ is continuous in θ unless $W\theta = 0$. By Assumption 3(i), $P(W\theta = 0) = 0$ for all θ, because W includes a continuous component ΔX_K with $\beta_K \neq 0$. To see that (c) holds, note that $|a(\cdot)| \leq 1$ for any value of (W, θ, D_2, D_1), hence the dominance condition is satisfied. □

A.6 Proof of Theorem 5

Proof. The proof follows that in Chen (2002) and Jochmans (2012), see also Horowitz (2009), and it is not repeated here. A summary of the proof is as follows. First, it is shown that $\widehat{g}_t$ is asymptotically linear by using the arguments in Sherman (1993). Then it is shown that the influence function of $\widehat{g}_t$ belongs to a Euclidean class of functions for unit envelope, so that Theorem 5.3 in Pollard (1984) applies. Notice that e.g. the only difference between our criterion function,

$$Q_n\left(y, y_0, \widetilde{y}, \widetilde{y}, \widehat{\beta}, g_1, 0\right),$$

and that of Chen (2002) is that in our framework the indicator $1\{(\Delta X_i - \Delta X_j)\beta > g_1\}$ is premultiplied by:

$$\overline{D}_i(y_1, y_2)\,\overline{D}_j\left(y_1', y_2'\right) 1\left\{D_{i2}(y_2) > D_{j2}\left(y_2'\right)\right\}$$

while in Chen (2002), it is premultiplied by:

$$D_i(y) - D_j(y_0)$$

Since these two functions are bounded from above by 1 and do not depend on the parameters of interest, the results in Chen (2002) and Jochmans (2012) apply directly. □

A.7 Proof of Theorem 6

Proof. This proof is similar to that in Muris (2017, Theorem 2) and proceeds by verifying the **conditions** in Theorem 2.7 in Newey and McFadden (1994). To verify those conditions, we connect the notation in the present paper to that in Newey and

McFadden (1994) by letting the sample objective function

$$\begin{aligned}\widehat{Q}_n(\theta) &= \frac{1}{n}\sum_i l_i(\theta, y_1, y_2) \\ &= \frac{1}{n}\sum_i \overline{D}_i(y_1, y_2)\left[D_{i2}(y_2)\ln\Lambda(W_i\theta) + (1 - D_{i2}(y_2))\ln(1 - \Lambda(W_i\theta))\right],\end{aligned}$$

population objective function

$$Q_0(\theta) = E\left[\overline{D}_i(y_1, y_2)\left[D_{i2}(y_2)\ln\Lambda(W_i\theta) + (1 - D_{i2}(y_2))\ln(1 - \Lambda(W_i\theta))\right]\right],$$

and true value of the parameter

$$\theta_0 = \theta_0(y_1, y_2),$$

which lives in $\Theta = \mathbb{R}^{K+1}$.

Because of the information inequality and the positive definiteness of $E\left(W_i W_i^{'}\right)$, Q_0 achieves its minimum uniquely at θ_0 (**condition i**).

In the main text, we established that

$$\begin{aligned}\frac{\partial^2 l_i(\theta, y_1, y_2)}{\partial\theta\left(\partial\theta\right)^{'}} &= H_i(\theta, y_1, y_2) \\ &= -\overline{D}_i(y_1, y_2)\Lambda(W_i\theta)(1 - \Lambda(W_i\theta))W_i W_i^{'}\end{aligned}$$

which is clearly non-positive since $\overline{D}_i(y_1, y_2) \in \{0, 1\}$, and $\Lambda(1 - \Lambda) \in (0, 1)$. The second derivative of $\widehat{Q}_n(\theta)$ is $\frac{1}{n}\sum_i H_i(\theta, y_1, y_2) \leq 0$ so that the sample objective function is concave (**condition ii**).

To bound the second moment of the conditional log likelihood contribution, note that:

$$\ln\Lambda(W_i\theta) = \ln\Lambda(0) + \left(1 - \Lambda\left(W_i\widetilde{\theta}\right)\right)W_i\theta$$

so that

$$\operatorname{var}\left(\overline{D}_i(y_1, y_2)D_{i2}(y_2)\ln\Lambda(W_i\theta)\right) \leq \theta^{'}\operatorname{var}(W)\theta < \infty.$$

A similar analysis applies to the second term in Q_0. Since the variance of l_i is bounded, and we have assumed a random sample, a law of large numbers applies to $Q_n(\theta)$ (**condition iii**). □

A.8 Proof of Theorem 7

Proof. This proof proceeds by verifying the conditions in Theorem 3.3 in Newey and McFadden (1994), which is an asymptotic normality result for maximum likelihood estimators. In Section A.7, we demonstrated the consistency of $\widehat{\theta}(y_1, y_2)$ for $\theta_0(y_1, y_2)$. The true value of the parameter is automatically in the interior, so that **condition i** is satisfied. The logit distribution function is twice differentiable, and the indicator functions do not depend on the parameter (**condition ii**). Denote

$$f(w|\theta) = \left[\Lambda(w\theta)^{D_{i2}(y)}(1-\Lambda(w\theta))^{(1-D_{i2}(y))}\right]^{\overline{D}_i(y_1,y_2)}$$

so that, letting Λ=$\Lambda(w\theta)$ and $\lambda \equiv \frac{\partial \Lambda}{\partial w\theta}\Lambda(w\theta)[1-\Lambda(w\theta)]$, so that

$$\begin{aligned}\partial\lambda/\partial(w\theta) &= \lambda(1-\Lambda) - \Lambda\lambda \\ &= \lambda(1-2\Lambda).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\partial f}{\partial \theta} &= \left[(\lambda w)^{D_{i2}(y)}(1-\Lambda)^{(1-D_{i2}(y))} - \Lambda^{D_{i2}(y)}(\lambda w)^{(1-D_{i2}(y))}\right]^{\overline{D}_i(y_1,y_2)} \\ &= \left[\lambda\left[(1-\Lambda)^{(1-D_{i2}(y))} - \Lambda^{D_{i2}(y)}\right]w\right]^{\overline{D}_i(y_1,y_2)}\end{aligned}$$

and

$$\frac{\partial^2 f}{\partial\theta(\partial\theta)'} = \left[\left(\lambda(1-2\Lambda)\left[(1-\Lambda)^{(1-D_{i2}(y))} - \Lambda^{D_{i2}(y)}\right] - \lambda^2\right)ww'\right]^{\overline{D}_i(y_1,y_2)}.$$

Let $\mathcal{N} = (\theta_0 - 1, \theta_0 + 1)$. Then

$$\begin{aligned}\int \sup_{\theta\in\mathcal{N}}\left\|\frac{\partial f(z|\theta)}{\partial\theta}\right\| dz &= E\left[\sup_{\theta\in\mathcal{N}}\|\left[\lambda\left[(1-\Lambda)^{(1-D_{i2}(y))} - \Lambda^{D_{i2}(y)}\right]W_i\right]^{\overline{D}_i(y_1,y_2)}\|\right] \\ &\leq aE\left[\|W_i'\|\right] < \infty,\end{aligned}$$

where $a = \sup_{\theta\in\mathcal{N}}\left\|\lambda\left[(1-\Lambda)^{(1-D_{i2}(y))} - \Lambda^{D_{i2}(y)}\right]\right\|$ which is bounded for a given interval in the interior of $\mathbb{R}^{k+1}$. Uniform boundedness on that interval then follows because

the second moment of W is bounded. Similarly,

$$\begin{aligned}
&\int \sup_{\theta\in\mathcal{N}} \left\| \frac{\partial^2 f\left(z|\theta\right)}{\partial\theta\partial\theta'} \right\| dz \\
&= E\left[\sup_{\theta\in\mathcal{N}} \| \left[\left(\lambda\left(1-2\Lambda\right) \left[\left(1-\Lambda\right)^{\left(1-D_{i2}(y)\right)} - \Lambda^{D_{i2}(y)} \right] - \lambda^2 \right) W_i W_i' \right]^{\overline{D}_i(y_1,y_2)} \| \right] \\
&\leq bE\left[\|W_i W_i'\| \right] < \infty,
\end{aligned}$$

where

$$b = E\left[\sup_{\theta\in\mathcal{N}} \| \left[\left(\lambda\left(1-2\Lambda\right) \left[\left(1-\Lambda\right)^{\left(1-D_{i2}(y)\right)} - \Lambda^{D_{i2}(y)} \right] - \lambda^2 \right) \right] \| \right]$$

which is finite on any open interval in $\mathbb{R}^{k+1}$. Together with assumed boundedness of the second moment of W_i, this gives the boundedness of the expectation of the second derivative of the density. Therefore**, condition iii** is satisfied.

For the Hessian, we have

$$\begin{aligned}
&\int \sup_{\theta\in\mathcal{N}} \left\| \frac{\partial^2 \ln f\left(z|\theta\right)}{\partial\theta\partial\theta'} \right\| dz \\
&= E\left[\sup_{\theta\in\mathcal{N}} \| -\overline{D}_i\left(y_1,y_2\right) \Lambda\left(W_i\theta\right) \left(1-\Lambda\left(W_i\theta\right)\right) W_i W_i' \| \right] \\
&\leq E\left[\|W_i W_i'\| \right] < \infty,
\end{aligned}$$

which is finite because the second moment of W is assumed to be bounded. **(condition v**).

In Section 4.2we show that the information matrix equality holds:

$$\begin{aligned}
&E\left[\frac{\partial \ln f\left(z|\theta\right)}{\partial\theta} \left(\frac{\partial \ln f\left(z|\theta\right)}{\partial\theta} \right)' \Bigg|_{\theta=\theta_0} \right] \\
=&E\left[\overline{D}_i\left(y_1,y_2\right) \Lambda\left(W_i\theta\right) \left(1-\Lambda\left(W_i\theta\right)\right) W_i W_i' \right].
\end{aligned}$$

To see that this matrix is non-singular (**condition iv**), see Muris (2017, proof of Theorem 2). □

A.9 Proof of Theorem 8

Proof. Given the multivariate CLT, and since $y \mapsto \theta(y)$ is continuous in y and $y \mapsto E\left(\overline{D}(y_1, y_2)\Lambda(W'\theta(y))(1-\Lambda(W'\theta(y)))WW'\right)$ is continuous at $\theta_0(y)$ for each $y \in \left[\underline{y}, \overline{y}\right]$, the result of Theorem 8 holds under stochastic equicontinuity of $(y, \theta) \mapsto \mathbb{G}_n\left(\overline{D}(y_1, y_2)(D_2(y) - \Lambda(W'\theta))W\right)$ where $\mathbb{G}_n(f(Z)) = \frac{1}{\sqrt{n}}\sum_i [f(Z_i) - Ef(Z_i)]$ and $Z = (Y_{i1}, Y_{i2}, W_i)$. To show this consider the following argument.

The function classes

$$\begin{aligned}
\mathcal{F}_1 &\equiv \left\{1\{Y_1 \geq y\}, y \in \left[\underline{y}, \overline{y}\right] \subset \mathcal{Y}\right\}, \\
\mathcal{F}_2 &\equiv \left\{1\{Y_2 \geq y\}, y \in \left[\underline{y}, \overline{y}\right] \subset \mathcal{Y}\right\}, \\
\mathcal{F}_3 &\equiv \left\{W'\theta, \theta \in \mathbb{R}^{k+1}\right\},
\end{aligned}$$

and $\{W_q, q = 1, \ldots, k+1\}$, where q indexes elements of the vector W, are VC classes of functions. The class

$$\mathcal{F} = \{1\{\mathcal{F}_1 + \mathcal{F}_2 = 1\}(\mathcal{F}_2 - \Lambda(\mathcal{F}_3))W_q, q = 1, \ldots, k+1\}$$

is a Lipschitz transformation of VC classes with Lipschitz constant bounded by $c\|W\|$, where c is a positive constant, and square integrable envelope $c\|W\|$. □

B FELT with random coefficients

In this Appendix, we extend the identification results to the case of FELT with random coefficients. We use a subpopulation of stayers (units whose regressors associated with the random coefficients do not change over time) to identify the common parameters. This technique is similar to that used by Graham and Powell (2012) in a linear model with random coefficients.

Denote by $Z = \left(Z_1', Z_2'\right)'$ the regressors associated with the random coefficient δ. Denote by $\mathcal{Z}$ the support of Z_t. Extend the model in (3.1) as the following latent variable model for $t = 1, 2$:

$$\begin{aligned}
Y_t^* &= \alpha + X_t\beta + Z_t\delta - U_t, \\
Y_t &= h_t(Y_t^*), \\
U_t|\alpha, \delta, X, Z &\sim F_t(u|\alpha, \delta, X, Z).
\end{aligned} \tag{B.1}$$

In this model,

$$\begin{aligned} D_t(y) &\equiv 1\{Y_t \geq y\} \\ &= 1\left\{U_t \leq \alpha + X_t\beta - Z_t\delta - h_t^-(y)\right\}, \end{aligned} \tag{B.2}$$

The presence of the additional regressors and their random coefficients δ requires us to impose slightly different assumptions:

Assumption 10. *[Error terms]*
(i) $F_1(u|\alpha,\delta,X.Z) = F_2(u|\alpha,\delta,X,Z) \equiv F(u|\alpha,\delta,X,Z)$ *for all* (α,δ,X,Z)*;*
(ii) The support of $F(u|\,\alpha,\delta,X,Z)$ *is* $\mathbb{R}$ *for all* (α,δ,X,Z)*.*

Let

$$\Delta Z \equiv Z_2 - Z_1.$$

Assumption 11. *[Positive density]* $P(\Delta Z = 0) > 0$.

This assumption can be relaxed to positive density of ΔZ in a neighborhood of 0, see Graham and Powell (2012).

Lemma 2. *Let Assumptions 1, 10, and 11 hold. Then*

$$med\left(D_2(y_2) - D_1(y_1)\,|X,\, \overline{D}(y_1,y_2) = 1, \Delta Z = 0\right) = sgn\left(\Delta X\beta - \gamma(y_1,y_2)\right). \tag{B.3}$$

Proof. The main derivation in the proof of Lemma 1 yields

$$\text{med}\left(D_2(y_2) - D_1(y_1)\,|X,\, \overline{D}(y_1,y_2) = 1, \delta, Z\right) = \text{sgn}\left(\Delta X\beta + \Delta Z\delta - \gamma(y_1,y_2)\right).$$

Further conditioning on $\Delta Z = 0$ obtains the desired result. □

Identification of $\theta(y_1,y_2) \equiv (\beta, \gamma(y_1,y_2))$ requires the following additional assumptions, which differ from 3 only by conditioning on the event $\Delta Z = 0$.

Assumption 12. *[Covariates] Denote by* $F^*_{\Delta X}$ *the distribution of* ΔX *conditional on* $\Delta Z = 0$.
(i) $F^*_{\Delta X}$ *is such that at least one component of* ΔX *has positive Lebesgue density on* $\mathbb{R}$ *conditional on all the other components, with probability one. The corresponding component of* β *is non-zero;*

(ii) The support of $F^*_{\Delta X}$ is not contained in any proper linear subspace of $\mathbb{R}^K$.

A modification of our main result 1 to the case with random coefficients now follows.

Theorem 9. *Suppose that (Y, X, Z) follow the model in* (B.1)*, and let the distribution of (Y, X, Z) be observed. Let Assumptions 1, 4, 10, 12, and 11 hold. Then β is identified. Additionally, under assumption 5, the functions $h_1(\cdot)$ and $h_2(\cdot)$ are identified.*

Proof. The proof extends that of Theorems 1 and 2. The only modification is that the probabilities are now understood to be conditional on the event $\Delta Z = 0$. □

Using the population of stayers, we have identified all the common parameters in this model. In what follows, assume for simplicity that h_1 and h_2 are invertible. Define the structural function as

$$Y_t(x) \equiv h_t(x\beta + Z_t\delta + \alpha - U_t). \tag{B.4}$$

Then, by arguments similar to those in Section 3.2, the structural function is identified and given by

$$Y_t(x) = h_t\left(h_t^{-1}(Y_t) + (x - X_t)\beta\right).$$

If interest lies in the effect of Z_t on Y_t, one can use the identification of the common parameters to transform the model to a linear panel data model with random coefficients:

$$\begin{aligned} S_t &\equiv h_t^{-1}(Y_t) - X_t\beta \\ &= \alpha + Z_t\gamma - U_t, \end{aligned}$$

One could then use the identification results in the literature on linear random coefficients models (Chamberlain (1992); Graham and Powell (2012); Arellano and Bonhomme (2012)).

C Rank estimator criterion function

In this Appendix, we discuss the intuition behind the criterion function in Section 4.1.2. We also explain that the additional scale normalization in that Section is

without loss of generality.

C.1 Motivating the criterion function

Consider an individual i and a pair of thresholds $y = (y_1, y_2)$, chosen by the researcher. The binary variables $D_{i1}(y_1)$ and $D_{i2}(y_2)$ indicate whether an individual's latent variable

$$Y_{it}^* = \alpha_i + X_{it}\beta - U_{it}$$

is above or below the threshold $h_t^-(y_t)$ on the latent variable in period t. The binary variable $\overline{D}_i(y_1, y_2)$ indicates whether the individual is a "switcher", i.e. whether the individual's outcome is above the threshold exactly once in the two time periods. Non-switchers are non-informative about the model parameters, as is well-known from the panel data binary choice literature.

There are two types of switchers. First, those with $D_{i2}(y_2) = 1$ (and $D_{i1}(y_1) = 0$) that start below the time-1 threshold and end above the time-2 threshold. In Abrevaya's (1999) language, the former type "leapfrogs" with respect to the time-varying thresholds. Second, there are those with $D_{i2}(y_2) = 0$ (and $D_{i1}(y_1) = 1$). They are anti-leapfroggers (call them "nosedivers") as they start above the threshold but end below.

Compare an observation i that leapfrogs with respect to (y_1, y_2) to an observation j that nosedives with respect to $\left(y_1', y_2'\right)$. Setting $(y_1, y_2) = \left(y_1', y_2'\right)$ leads to information about β, whereas considering different values $(y_1, y_2) \neq \left(y_1', y_2'\right)$ provides information on

$$\beta, \left(h_2^-(y_2) - h_2^-\left(y_2'\right)\right) - \left(h_1^-(y_1) - h_1^-\left(y_1'\right)\right).$$

To see this, note that i, the leapfrogger, has

$$\begin{aligned} Y_{i1}^* &= \alpha_i + X_{i1}\beta - U_{i1} < h_1^-(y_1) \\ Y_{i2}^* &= \alpha_i + X_{i2}\beta - U_{i2} > h_2^-(y_2) \end{aligned}$$

so that

$$\Delta X_i \beta > h_2^-(y_2) - h_1^-(y_1) + \Delta U_{it}. \tag{C.1}$$

Similarly, for a nosediver j, we have

$$\Delta X_j\beta < h_2^-\left(y_2^{'}\right) - h_1^-\left(y_1^{'}\right) + \Delta U_{jt}. \tag{C.2}$$

Then, subtracting C.2 from C.1 obtains

$$\Delta X_i\beta - \Delta X_j\beta > \left(h_2^-\left(y_2\right) - h_1^-\left(y_1\right)\right) - \left(h_2^-\left(y_2^{'}\right) - h_1^-\left(y_1^{'}\right)\right) + \left(\Delta U_{it} - \Delta U_{jt}\right).$$

The error terms are stationary, and independent across observations. Therefore, the inequality

$$\left(\Delta X_i - \Delta X_j\right)\beta > \left(h_2^-\left(y_2\right) - h_1^-\left(y_1\right)\right) - \left(h_2^-\left(y_2^{'}\right) - h_1^-\left(y_1^{'}\right)\right)$$

is associated with

$$D_{2i}\left(y_2\right) - D_{2j}\left(y_2^{'}\right) > 0$$

conditional on $\overline{D}_i\left(y_1, y_2\right) = \overline{D}_j\left(y_1^{'}, y_2^{'}\right) = 1$.

The sample objective function sums over all pairwise comparisons for a given choice of $\left(y_1, y_2, y_1^{'}, y_2^{'}\right)$. For $y_1 = y_1^{'}$ and $y_2 = y_2^{'}$, the objective function does not depend on h_t, and corresponds to Abrevaya's (1999) criterion function applied to a subset of the pairs. For $y_2 = y_2^{'}$ and $y_1^{'} = y_0$,it corresponds to Chen's (2002) estimator, applied to the subset of pairs with $\overline{D}_i\left(y_1, y_2\right)\overline{D}_j\left(y_1^{'}, y_2^{'}\right) = 1$, using differences rather than levels.

C.2 Normalization on h_2

The procedure in Section 4.1.2 targets β, h_1^-, and $h_2^-\left(y_2\right) - h_2^-\left(y_0\right)$ where y_0 is such that $h_1^-\left(y_0\right) = 0$. Here, we show that the scale of h_2^- is recovered by augmenting the first step regression with a time dummy.

The FELT model is given by

$$Y_{i1} = h_1\left(\alpha_i + X_{i1}\beta - U_{i1}\right)$$

$$Y_{i2} = h_2\left(\alpha_i + X_{i2}\beta - U_{i2}\right)$$

with the normalization $h_1^-\left(y_0\right)$. Here, as in Section 4.1.2, we assume invertibility of h_t.

Consider a normalized version of the transformation function in the second time period,

$$\tilde{h}_2^- (y_2) = h_2^- (y_2) - h_2^- (y_0),$$

which is exactly the quantity obtained from the rank procedure. Consider the following model:

$$\begin{aligned} Y_{i1} &= h_1 (\alpha_i + X_{i1}\beta - U_{i1}) \\ Y_{i2} &= \tilde{h}_2 (\alpha_i + X_{i2}\beta - U_{i2} - \delta). \end{aligned}$$

Then

$$\begin{aligned} \delta &= (\alpha_i + X_{i2}\beta - U_{i2}) - \tilde{h}_2^{-1} (Y_{i2}) \\ &= h_2^{-1\tilde{}} (Y_{i2}) - \tilde{h}_2^{-1} (Y_{i2}) \\ &= h_2^{-1\tilde{}} (Y_{i2}) - \left(h_2^- (y_2) - h_2^- (y_0)\right) \\ &= h_2^- (y_0), \end{aligned}$$

so that the scale of h_2 can be obtained by applying the leapfrog estimator to $(Y_{it}, (X_{it}, 1\{t = 2\}))$, i.e. a transformation model augmented with a time dummy.

References

Abrevaya, J. (1999). Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *Journal of Econometrics*, 93(2):203–228.

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics*, 95(1):1–23.

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73(4):1175–1204.

Altonji, J. G. and Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica*, 73(4):1053–1102.

Alvarez, J. and Arellano, M. (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica*, 71(4):1121–1159.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Arellano, M. and Bonhomme, S. (2009). Robust priors in nonlinear panel data models. *Econometrica*, 77(2):489–536.

Arellano, M. and Bonhomme, S. (2012). Nonlinear panel data analysis. *Annual Review of Economics*, 3:395–424.

Arellano, M. and Hahn, J. (2007). Understanding bias in nonlinear panel models: Some recent developments. In Blundell, R., Newey, W., and Persson, T., editors, *Advances in Economics and Econometrics*, pages 381–409. Cambridge University Press.

Arellano, M. and Honoré, B. (2001). Panel data models: Recent developments. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5, chapter 53, pages 3219–3296. Elsevier.

Athey, S. and Imbens, G. (2006). Identification and inference in nonlinear difference-in-differences model. *Econometrica*, 74(2):431–497.

Baetschmann, G., Staub, K. E., and Winkelmann, R. (2015). Consistent estimation of the fixed effects ordered logit model. *Journal of the Royal Statistical Society A*, 178(3):685–703.

Bester, C. A. and Hansen, C. (2009). Identification of marginal effects in a nonparametric correlated random effects model. *Journal of Business and Economic Statistics*, 27(2):235–250.

Blundell, R. and Powell, J. L. (2003). *Endogeneity in Nonparametric and Semiparametric Regression Models*, volume 2 of *Econometric Society Monographs*, pages 312–357. Cambridge University Press.

Bonhomme, S. (2012). Functional differencing. *Econometrica*, 800(4):1337–1385.

Bonhomme, S. and Sauder, U. (2011). Recovering distributions in difference-in-differences: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, 93(2):479–494.

Chamberlain, G. (1984). Panel data. In *Handbook of Econometrics*, volume 2, pages 1247–1318. Elsevier.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In Heckman, J. J. and Singer, B. S., editors, *Longitudinal Analysis of Labor Market Data*, Econometric Society Monographs, pages 3–38. Cambridge University Press.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, 60(3):567–596.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica*, 78(1):159–168.

Chen, S. (2002). Rank estimation of transformation models. *Econometrica*, 70(4):1683–1697.

Chen, S. (2010). Root-n-consistent estimation of fixed-effect panel data transformation models with censoring. *Journal of Econometrics*, 159(1):222–234.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013a). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

Chernozhukov, V., Fernández-Val, I., Hoderlein, S., Holzmann, H., and Newey, W. (2015). Nonparametric identification in panels using quantiles. *Journal of Econometrics*, 188(2):378–392.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013b). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Chiappori, P.-A., Komunjer, I., and Kristensen, D. (2015). Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22–39.

D'Haultfoeuille, X., Hoderlein, S., and Sasaki, Y. (2015). Nonlinear difference-in-differences in repeated cross sections with continuous treatments. Working paper.

Evdokimov, K. (2011). Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells. Working paper.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150(1):71–85.

Fernández-Val, I. and Lee, J. (2013). Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. *Quantitative Economics*, 4(3):453–481.

Fernández-Val, I. and Weidner, M. (2016). Individual and time effects in nonlinear panel data models with large N, T. *Journal of Econometrics*, 192(1):291–312.

Florens, J.-P. and Sokullu, S. (2016). Nonparametric estimation of semiparametric transformation models. *Econometric Theory*, pages 1–35.

Freyberger, J. (2012). Nonparametric panel data models with interactive fixed effects. Working paper.

Ghanem, D. (2017). Testing identifying assumptions in nonseparable panel data models. *Journal of Econometrics*, 197(2):202–217.

Graham, B. and Powell, J. (2012). Identification and estimation of average partial effects in irregular correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large. *Econometrica*, 70(4):1639–1657.

Hahn, J. and Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319.

Hoderlein, S. and White, H. (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, 168(2):300–314.

Honoré, B. (1992). Trimmed lad and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica*, 60(3):533–565.

Horowitz, J. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer-Verlag.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, 64(1):103–137.

Jochmans, K. (2012). The variance of a rank estimator of transformation models. *Economics Letters*, 117(1):168–169.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65(6):1335–1364.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2):391–413.

Lancaster, T. (2002). Orthogonal parameters and panel data. *The Review of Economic Studies*, 69(3):647–666.

Machado, M. (2004). A consistenct estimator for the binomial distribution in the presence of incidental parameters: an application to patent data. *Journal of Econometrics*, 119(1):73–98.

Magnac, T. (2004). Panel binary variables and suficiency: Generaling conditional logit. *Econometrica*, 72(6):1859–1876.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica*, 55(2):357–362.

Muris, C. (2017). Estimation in the fixed effects ordered logit model. *The Review of Economics and Statistics*. forthcoming.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. and McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Pakes, A. and Porter, J. (2016). Moment inequalities for multinomial choice with fixed effects. NBER Working Paper No. 21893.

Sherman, R. P. (2010). *Maximum Score Methods*, pages 122–128. Palgrave Macmillan UK, London.

Shi, X., Shum, M., and Song, W. (2016). Estimating semi-parametric panel multinomial choice models using cyclic monotonicity. Working paper.