Discussion Papers in Economics

No. 09/2019

Fixed-Effects estimation of the Linear Discrete-Time Hazard Model: an Adjusted First-Differences Estimator

Tauchmann, Harald
FAU Erlangen-Nürnberg

ISSN 1867-6707

Friedrich-Alexander-Universität Erlangen-Nürnberg
Institute for Economics
https://www.iwf.rw.fau.de/discussion-papers/

Fixed-Effects estimation of the Linear Discrete-Time Hazard Model: an Adjusted First-Differences Estimator*

Harald Tauchmann

Universität Erlangen-Nürnberg
RWI – Leibniz Institut für Wirtschaftsforschung
CINCH – Health Economics Research Center

November 2019

Abstract

This paper shows that popular linear fixed-effects panel-data estimators (first-differences, within-transformation) are biased and inconsistent when applied in a discrete-time hazard setting, that is, one with the outcome variable being a binary dummy indicating an absorbing state, even if the data generating process is fully consistent with the linear discrete-time hazard model. Besides conventional survival bias, these estimators suffer from another source of – potentially severe – bias that originates from the data transformation itself and is present even in the absence of any unobserved heterogeneity. We suggest an alternative, computationally very simple, adjusted first-differences estimator that cures the data-transformation driven bias of the classical estimators. The theoretical line of argument is supported by evidence from Monte Carlo simulations and is illustrated by an empirical application.

JEL Codes: C23, C25, C41.
Keywords: linear probability model, individual fixed effects, short panel, discrete-time hazard, duration analysis, survival analysis, non-repeated event, absorbing state, survival bias, misscaling bias.

*Address for correspondence: Harald Tauchmann, Professur für Gesundheitsökonomie, Findelgasse 7/9, 90402 Nürnberg, Germany. Email: harald.tauchmann@fau.de. Phone: +49 (0)911 5302 635. The user-written Stata® ado-file `xtlhazard` that implements the estimation procedure suggested in this paper is available from `ssc` (Boston College Statistical Software Components). I would like to thank Daniel Kühnle, Helmut Herwartz, Simon Reif, Boris Hirsch, Claus Schnabel, Stefan Pichler, the members of the dggö Health Econometrics Working Group, the participants of the 2019 German Stata Users Group Meeting, the RWI Research Seminar, the Nuremberg Research Seminar in Economics, and the Verein für Socialpolitik Annual Conference 2019 for valuable comments and suggestions. Excellent research assistance from Helene Könnecke, Sabrina Schubert, and Irina Simankova is gratefully acknowledged.

1 Introduction

Many economically relevant outcomes are non-repeated events, i.e, absorbing states. Death, retirement, firm bankruptcy, plant closure, technology adoption, and smoking initiation are just a few examples among numerous others.[1] Hazard models, also referred to as duration, failure-time, survival, and event history analysis are usually used for modeling such outcomes. If the analysis is based on panel data, in which the outcome is not continuously observed but only at a limited number of points in time[2], discrete-time hazard models are often regarded as the estimation method of choice. These models are simply stacked binary outcome models (Jenkins, 1995; Cameron and Trivedi, 2005, p. 602), such as probit, logit, or cloglog (Prentice and Gloeckler, 1978). The discrete-time hazard binary-outcome model approach has much appeal since it is not only technically simple but also intuitive as it allows thinking of a process that may lead into the absorbing state as a series of binary choices.

Following the general trend towards using linear models in applied econometrics, which put little emphasis on correctly specifying the data generating process but rely on their capability of identifying average partial effects even in the presence of non-linearities (Angrist and Imbens, 1995), the linear probability model has developed into an increasingly popular alternative to non-linear binary outcome models (cf. Angrist and Pischke, 2009, 2010). One argument in favor of the linear probability model is that it allows straightforwardly removing unobserved individual heterogeneity as a possible source of bias, using the within- or the first-differences transformation. Allowing for individual fixed effects is far less straightforward in non-linear models (e.g. Greene, 2004; Stammann et al., 2016). In fact, in recent empirical analyses the linear probability model with individual fixed effects has frequently been applied not only to repeated events, but also to non-repeated event data (e.g. Miguel et al., 2004; Ciccone, 2011; Brown and Laschever, 2012; Cantoni, 2012; Harding and Stasavage, 2014; Jacobson and von Schedvin, 2015; Fernandes and Paunov, 2015; Wang et al., 2017; Bogart, 2018). This suggests that there is little awareness that the favorable properties of the popular linear fixed-effects estimators do not in the same way apply to non-repeated event settings as they apply to other kinds of dependent variables. We are, in fact, not aware of any article that explicitly establishes the properties of the conventional linear fixed-effects estimators in a discrete-time hazard setting.[3]

[1] In empirical applications, it often depends on the institutional setting, the available data, the research question, and the economic model one has in mind, whether thinking of such events – retirement for instance – as non-repeated is appropriate.

[2] This includes both, cases in which the time structure is intrinsically discrete (e.g. termination of a rolling fixed-period contract) and cases in which thinking of time as a sequence of periods of significant length is an artifact of incompletely observing the process of interest.

[3] Allison and Christakis (2006) and Allison (2009, chap. 5) discuss obstacles to fixed-effects estimation of in non-linear hazard models but do not consider the linear model. Allison (1994) considers linear fixed-effects estimation but thinks

In this paper we show that conventional linear fixed-effects estimators (first-differences, within-transformation) fail to remove unobserved individual heterogeneity if the outcome variable indicates an absorbing state. This applies even if the true data generating process is fully consistent with the linear hazard assumption. Moreover these estimators are – contingent on the data generating process of the explanatory variables – potentially severely biased, even if the unobserved heterogeneity is uncorrelated with the explanatory variables in the population. The bias originates from two sources. One is selective survival that renders the unobserved heterogeneity correlated with the explanatory variables in the estimation sample. This bias is not specific to fixed-effects estimation but – in a somewhat different way – also applies to pooled OLS. The second source of bias is specific as it originates from the first differences and within-transformation itself that makes the exogenous variables enter the conditional mean of the disturbance. For this reason this second source of bias is present even in the absence of any unobserved individual heterogeneity. Moreover this second source of bias turns out to be the clearly dominant one in many settings.

Building on this result, we suggest a novel adjusted first-differences estimator that eliminates the second source of bias. Though the suggested estimator still suffers from survival bias, it outperforms conventional fixed estimators in almost all considered settings. This in particular applies to applications to small samples and non-stationary explanatory variables for which the bias of the conventional estimators can be of strange magnitude. The contribution of this paper is twofold. Firstly it pins down why conventional fixed-effects estimators should not be used in a discrete time hazard framework. Secondly it suggests an alternative estimator that – though not consistent – usually suffers from a smaller asymptotic bias and, more importantly, confines the asymptotic bias to survival bias. This is a source of bias researchers should anyway be aware of since also OLS is subject to it, even if the unobserved heterogeneity is uncorrelated with the explanatory variables in the population.

The remainder of this paper is organized as follows. In section 2 we establish biasedness and inconsistency of the conventional fixed-effects estimators and develop an alternative method that eliminates the data-transformation driven bias. In section 3 we use Monte Carlo simulations to compare the different estimators. Section 4 presents an empirical application, which is based on the analysis of peer effects in the timing of retirement by Brown and Laschever (2012). Section 5 concludes.

of non-repeated events as explanatory variables rather than outcome variables. Horowitz and Lee (2004) suggest a fixed-effects estimator for continuous-time proportional hazard model with multiple-spells. Horowitz (1999) proposes a random-effects estimator for a similar setting with single-spell data.

2 Model and Theory

2.1 The Data Generating Process

Consider a linear probability model in a panel data context. We observe N units i in a panel of T waves t, i.e. $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The units i are independently sampled from the population. The number of panel waves is finite and fixed and is small compared to the number of cross-sectional units. Any argument regarding asymptotic properties is hence in terms of $N \to \infty$. y_{it} denotes a binary outcome variable. The scalar a_i denotes unobserved, time-invariant individual heterogeneity, with $\mathrm{E}(a_i) = \alpha$. $\mathbf{x}_{it}$ is a $1 \times k$ row vector of exogenous explanatory variables observed for unit i in period t, which does not include a constant. β is a $k \times 1$ column vector of coefficients subject to estimation. We assume $a_i + \mathbf{x}_{it}\beta \in [0, 1]$ for any i and any t. That is, the argument of Horrace and Oaxaca (2006) that the least squares linear probability estimator is biased and inconsistent if this condition is violated, by assumption, does not apply.

$y_{it} = 1$ represents an absorbing state and, in consequence, only a single spell at risk is observed for any unit i.[4] In other words, after observing $y_{it} = 1$ for the first time, any possible available subsequent observations of i do not contain any additional information about the data generating process, since for $s \geq 1$, y_{it+s} equals one, irrespective of $\mathbf{x}_{it+s}$. In many applications one may not even observe $\mathbf{x}_{it+s}$.[5] The number of periods $T_i \leq T$ for which unit i is (effectively) observed is hence not fixed but endogenous. By thinking of T as fixed, we implicitly allow for right censoring, i.e. we may not observe the (first) occurrence of $y_{it} = 1$ for some units. We use $\mathbf{y}_{it^-}$ to denote the vector of outcomes for all periods prior to t. The data generating process (DGP) of y_{it} reads as

$$y_{it} = a_i + \mathbf{x}_{it}\beta + \varepsilon_{it} \tag{1}$$

and for the disturbance term $\varepsilon_{it} = y_{it} - a_i - \mathbf{x}_{it}\beta$ necessarily holds

$$\varepsilon_{it} = \begin{cases} 1 - a_i - \mathbf{x}_{it}\beta & \text{if} \quad t = T_i \quad \text{and } i \text{ is not censored} \\ -a_i - \mathbf{x}_{it}\beta & \text{if} \quad t = T_i \quad \text{and } i \text{ is censored} \\ -a_i - \mathbf{x}_{it}\beta & \text{if} \quad t < T_i \end{cases} \tag{2}$$

since y_{it} equals one for the final observation of a noncensored unit and is otherwise zero.

[4] If the spell is considered the genuine unit of observation and, correspondingly, the constant α is spell rather than unit (individual, firm, country, etc.) specific, the line of argument likewise applies to cases that allow for multiple spells at risk being observed for one unit.

[5] Events such as death or bankruptcy may render some time varying characteristics of i ill-defined or unobservable after the event has occurred and will usually result in attrition from the panel.

Assuming zero conditional mean for the disturbance

$$\mathrm{E}\left(\varepsilon_{it}|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0}\right) = 0 \tag{3}$$

renders (1) a regression model and yields a conditional probability of the event $y_{it} = 1$

$$\mathrm{P}(y_{it} = 1|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0}) = a_i + \mathbf{x}_{it}\beta \tag{4}$$

which is linear in a_i and $\mathbf{x}_{it}$.

2.2 Estimation by Ordinary Least Squares

First, using pooled ordinary least squares (OLS) to estimate β, this empirical model captures the mean of a_i conditional on entering the estimation sample $\alpha^c \equiv \mathrm{E}(a_i|t \leq T_i, \mathbf{X})$ by including a constant term in the regression.[6] However, the model does not take into account the heterogeneity in a_i. For this reason, the disturbance in this regression is not ε_{it} but $\varepsilon_{it}^{\mathrm{OLS}} \equiv y_{it} - \alpha^c - \mathbf{x}_{it}\beta$ and for its conditional mean we get

$$\begin{aligned}
\mathrm{E}\left(\varepsilon_{it}^{\mathrm{OLS}}|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0}\right) &= \mathrm{P}(y_{it} = 1|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0})\,(1 - \alpha^c - \mathbf{x}_{it}\beta) \\
&\quad + \mathrm{P}(y_{it} = 0|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0})\,(-\alpha^c - \mathbf{x}_{it}\beta) \\
&= (a_i + \mathbf{x}_{it}\beta)\,(1 - \alpha^c - \mathbf{x}_{it}\beta) \\
&\quad + (1 - a_i - \mathbf{x}_{it}\beta)\,(-\alpha^c - \mathbf{x}_{it}\beta) \\
&= a_i - \alpha^c
\end{aligned} \tag{5}$$

For $\mathrm{Cov}(a_i, \mathbf{x}_{it}) \neq \mathbf{0}$ the well known problem of unobserved heterogeneity rendering least squares biased and inconsistent arises. One may hence think of applying the within- or first-differences transformation to the data, in order to eliminate a_i and to allow unbiased and consistent estimation by least squares. The subsequent section demonstrates that this well established approach does not succeed in the considered setting, i.e. with an indicator for a non-repeated event entering the regression model on the left-hand side.

[6] The larger a_i is, the more likely it is that the event $y_{it} = 1$ occurs and that T_i is small. The estimation sample is hence selective with respect to a_i and $\alpha^c \neq \alpha$. $\mathbf{X}$ denotes the regressor matrix.

2.3 Biasedness of the First-Differences Estimator

Consider the conventional first-differences estimator for the above linear probability model. For the first-differenced dependent variable we have $\Delta y_{it} \equiv y_{it} - y_{it-1} = y_{it}$, since $y_{it-1} = 0$ follows from the fact that y_{it} denotes a non-repeated event. $\Delta \mathbf{x}_{it} \equiv \mathbf{x}_{it} - \mathbf{x}_{it-1}$ denotes the vector of the first-differenced right-hand side variables, and the first-differenced disturbance $\varepsilon_{it}^{\text{FD}}$ is just $y_{it} - \Delta \mathbf{x}_{it}\beta$. The conventional first-differences estimator that does not include a constant term[7] reads as follows:

$$b^{\text{FD}} = \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}' \Delta \mathbf{x}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}' y_{it} \right) = \beta + \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}' \Delta \mathbf{x}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}' \varepsilon_{it}^{\text{FD}} \right) \quad (6)$$

The conditional mean of the disturbance $\varepsilon_{it}^{\text{FD}}$ in this regression is

$$\begin{aligned} \mathrm{E}(\varepsilon_{it}^{\text{FD}} | a_i, \mathbf{x}_{it}, \mathbf{x}_{it-1}, \mathbf{y}_{it^-} = \mathbf{0}) &= \mathrm{P}(y_{it} = 1 | a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0}) \, (1 - \Delta \mathbf{x}_{it}\beta) \\ &\quad + \mathrm{P}(y_{it} = 0 | a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0}) \, (-\Delta \mathbf{x}_{it}\beta) \\ &= (a_i + \mathbf{x}_{it}\beta) \, (1 - \Delta \mathbf{x}_{it}\beta) + (1 - a_i - \mathbf{x}_{it}\beta) \, (-\Delta \mathbf{x}_{it}\beta) \\ &= a_i + \mathbf{x}_{it-1}\beta \end{aligned} \quad (7)$$

In this setting, taking first-differences fails to remove unobserved individual heterogeneity and, in turn, fails to generate a transformed disturbance that is conditional mean independent of the exogenous variables. This is so even if a_i is uncorrelated with the regressors in the population. The result in (7) has much intuitive appeal. Since the left-hand-side variable remains unaffected by the first-differences transformation, the disturbance needs to fully compensate for the transformation that is applied to the deterministic part of the right-hand side. This is why $\mathrm{E}(\varepsilon_{it}^{\text{FD}} | a_i, \mathbf{x}_{it}, \mathbf{x}_{it-1}, \mathbf{y}_{it^-} = \mathbf{0}) - \mathrm{E}(\varepsilon_{it}^{\text{OLS}} | a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0})$ equals $\alpha^c + \mathbf{x}_{it-1}\beta$, which is what is subtracted from the deterministic part by taking first-differences. This argument in general also holds true for the within-transformation, which for $T = 2$ is fully equivalent to taking first-differences (e.g. Wooldridge, 2002, p. 284).[8] That the first-differences estimator b^{FD} is biased follows directly

[7] A constant in a first-differences regression model is equivalent to a linear time trend in the regression in levels. Hence, one usually only includes a constant if the empirical model involves a linear time trend. Yet, since most applications of discrete-time hazard models will allow for duration dependence of the baseline hazard by including a trend or, more typically, wave indicators, the first-differences model will effectively include a constant.

[8] For the general case $T > 2$, the conditional mean of the disturbance in the within-transformation model is a much more complicated function than (7), see equation (24) in Appendix A.1. However, the crucial result that a_i and $\mathbf{x}_{is}$, with $s < t$, enter the conditional mean holds for any value of T. It is important to note that the former argument does not apply if y_{it} is a repeated event. In this alternative setting, the first-differences transformation is not immaterial for the dependent variable and yields a transformed disturbance with the desired properties, see Appendix A.2.

from (6) and (7):

$$\begin{aligned}
\mathrm{E}\left(b^{\mathrm{FD}}|\mathbf{X},\mathbf{a}\right) &= \beta + \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}'a_i\right) \\
&\quad + \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}'\mathbf{x}_{it-1}\right)\beta \\
&\neq \beta
\end{aligned} \tag{8}$$

Now consider an alternative first-differences estimator, denoted by b^{FDC}, that includes a constant[9] term. In a panel of only two waves this estimator coincides with the within-transformation estimator if the latter includes a wave indicator. The disturbance in this regression model is $\varepsilon_{it}^{\mathrm{FDC}} \equiv y_{it} - \tilde{\alpha}^c - \Delta\mathbf{x}_{it}\beta$ and its conditional mean reads as

$$\begin{aligned}
\mathrm{E}(\varepsilon_{it}^{\mathrm{FDC}}|a_i,\mathbf{x}_{it},\mathbf{x}_{it-1},\mathbf{y}_{it^-}=\mathbf{0}) &= \mathrm{P}(y_{it}=1|a_i,\mathbf{x}_{it},\mathbf{y}_{it^-}=\mathbf{0})\,(1-\tilde{\alpha}^c-\Delta\mathbf{x}_{it}\beta) \\
&\quad + \mathrm{P}(y_{it}=0|a_i,\mathbf{x}_{it},\mathbf{y}_{it^-}=\mathbf{0})\,(-\tilde{\alpha}^c-\Delta\mathbf{x}_{it}\beta) \\
&= (a_i+\mathbf{x}_{it}\beta)\,(1-\tilde{\alpha}^c-\Delta\mathbf{x}_{it}\beta) \\
&\quad + (1-a_i-\mathbf{x}_{it}\beta)\,(-\tilde{\alpha}^c-\Delta\mathbf{x}_{it}\beta) \\
&= (a_i-\tilde{\alpha}^c)+\mathbf{x}_{it-1}\beta \\
&= \tilde{a}_i+\widetilde{\mathbf{x}}_{it-1}\tilde{\beta}
\end{aligned} \tag{9}$$

with $\tilde{a}_i \equiv a_i - \tilde{\alpha}^c$, $\tilde{\beta}' \equiv [\tilde{\alpha}^c \;\; \beta']$, and $\widetilde{\mathbf{x}}_{it-1} \equiv [0 \;\; \mathbf{x}_{it-1}]$. Though including a constant term still does not remove the individual effects from the disturbance, it captures their mean, which in consequence does not enter the disturbance. Let $\widetilde{\Delta\mathbf{x}}_{it} \equiv [1 \;\; \Delta\mathbf{x}_{it}]$ denote the vector of the right-hand side variables including the constant term. For the conditional mean of the first-differences estimator with a constant term we get

$$\begin{aligned}
\mathrm{E}\left(b^{\mathrm{FDC}}|\mathbf{X},\mathbf{a}\right) &= \tilde{\beta} + \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\tilde{a}_i\right) \\
&\quad + \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{it-1}\right)\tilde{\beta}
\end{aligned} \tag{10}$$

Though (10) looks very similar to (8), the crucial difference is that the demeaned rather than the raw unobserved individual heterogeneity enters the conditional means. This will usually render the first bias term in (10), i.e. $\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\tilde{a}_i\right)$, much smaller than its counterpart in (8).

[9] Since the first panel wave is effectively excluded from the estimation sample through taking first-differences, the constant captures $\tilde{\alpha}^c = \mathrm{E}(a_i|1 < t \leq T_i, \mathbf{X})$ rather than α^c.

A natural extension to including a single constant in the first differences estimator is to include a set of $T-1$ wave specific constants. This is what one usually does in order to allow for a baseline hazard that is not flat. While this is straightforward and does not alter the nature of the model, one has to be aware that wave specific constants do not capture the genuine time effects on the baseline hazard for the same reason for which a single constant does not capture the population mean of the unobserved heterogeneity. They rather capture both, true changes in the baseline hazard and that the conditional mean $\mathrm{E}(a_i|t,\mathbf{X})$ changes from one period to the next due to selective survival. The baseline hazard is hence not identified even if period specific constants are included in the empirical model.[10]

2.4 Inconsistency of the First-Differences Estimator

To determine the asymptotic properties of b^{FDC} we define $\tilde{\varepsilon}_{it}^{\mathrm{FDC}} \equiv \varepsilon_{it}^{\mathrm{FDC}} - \tilde{a}_i$. Then we write the first-difference estimator with constant term as

$$b^{\mathrm{FDC}} = \tilde{\beta} + \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}_{it}' \widetilde{\Delta \mathbf{x}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}_{it}' \tilde{a}_i + \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}_{it}' \tilde{\varepsilon}_{it}^{\mathrm{FDC}} \right) \tag{11}$$

Based on (9) and assuming that the data are well behaved, i.e. finite first and second moments of $\mathbf{x}_{it}$ and $\mathbf{x}_{it-1}$ exist, we get

$$\operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}_{it}' \tilde{\varepsilon}_{it}^{\mathrm{FDC}} \right) = \operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}_{it}' \widetilde{\mathbf{x}}_{it-1} \right) \tilde{\beta} \tag{12}$$

which will in general deviate from $\mathbf{0}$, unless β equals zero or $\mathbf{x}_{it}$ follows a random walk. For the probability limit of the term in (11) that involves $\tilde{a}_i$ we get

$$\operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}_{it}' \tilde{a}_i \right) = \operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\mathbf{x}}_{it}' \tilde{a}_i \right) - \operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\mathbf{x}}_{it-1}' \tilde{a}_i \right) \tag{13}$$

That is, this term converges in probability to a weighted sum of differences in conditional covariances i.e. $\sum_{t=2}^{T} w_t \big(\mathrm{Cov}(a_i, \mathbf{x}_{it} | \mathbf{y}_{it^-} = \mathbf{0}) - \mathrm{Cov}(a_i, \mathbf{x}_{it-1} | \mathbf{y}_{it^-} = \mathbf{0}) \big)$. These differences may deviate from $\mathbf{0}$ for two reasons. Firstly (13) will obviously not vanish in the limit, if we have $\mathrm{Cov}(a_i, \mathbf{x}_{it}) \neq \mathrm{Cov}(a_i, \mathbf{x}_{it-1})$ in the population, which is equivalent to having $\mathrm{Cov}(a_i, \Delta \mathbf{x}_{it}) \neq \mathbf{0}$. In other words, (13) deviates from $\mathbf{0}$ if the individual heterogeneity is correlated with the changes in the explanatory variables. For the population, one may rule this out by assumption. Yet, secondly, even assuming that a_i is uncorrelated with $\Delta \mathbf{x}_{it}$ in the population does not render (13) zero. The

[10] See Appendix A.4 for simulation results that consider time effects in the data generating process and in the model specification that is estimated.

reason for this is survival bias in the sense that conditioning on $\mathbf{y}_{it^-} = \mathbf{0}$ affects the covariance of a_i and $\mathbf{x}_{it}$. This is most obvious for $\mathrm{Cov}(a_i, \mathbf{x}_{it-1}|\mathbf{y}_{it^-} = \mathbf{0})$. Conditioning on $\mathbf{y}_{it^-} = \mathbf{0}$ means that $\mathbf{x}_{it-1}$ enters the conditional covariance only if $y_{it-1} = 0$ holds. This implies that large $\mathbf{x}_{it-1}$ are more likely to enter for a small value of a_i than for a large value of a_i. Conditioning on survival thus renders a_i and $\mathbf{x}_{it-1}$ negatively correlated, unless $\mathbf{x}_{it-1}$ is immaterial for survival that is $\beta = \mathbf{0}$. This does not one-to-one apply to $\mathrm{Cov}(a_i, \mathbf{x}_{it}|\mathbf{y}_{it^-} = \mathbf{0})$, since that covariance is unconditional on the contemporaneous y_{it}. However, if $\mathbf{x}_{it}$ exhibits some persistence over time, the negative correlation with a_i carries over to $\mathbf{x}_{it}$. In the case of perfect persistence, that is if $\mathbf{x}_{it}$ follows a random walk, the conditional covariance is the same for $\mathbf{x}_{it-1}$ and $\mathbf{x}_{it}$. In consequence (13) equals $\mathbf{0}$ if $\mathbf{x}_{it}$ follows a random walk. Yet, the smaller the persistence of $\mathbf{x}_{it}$ is, the more $\mathrm{Cov}(a_i, \mathbf{x}_{it}|\mathbf{y}_{it^-} = \mathbf{0})$ deviates from $\mathrm{Cov}(a_i, \mathbf{x}_{it-1}|\mathbf{y}_{it^-} = \mathbf{0})$, rendering a_i and $\Delta\mathbf{x}_{it}$ positively correlated for a positive β. Besides the dynamic properties of $\mathbf{x}_{it}$, the variance of a_i plays an important role for the size of the survival bias. If the variance of a_i is small then survival from $t-1$ to t is hardly selective. If so, conditioning or not conditioning on the contemporaneous y_{it} makes little difference for the distribution of $\mathbf{x}_{it}$. This renders (13) close to zero and in turn renders survival bias a minor issue.

It is important to note that not only the first differences estimator but also pooled OLS suffers from survival bias, even if a_i and $\mathbf{x}_{it}$ are uncorrelated in the population. Yet, for OLS it is not the differences in conditional covariances but only the levels of $\mathrm{Cov}(a_i, \mathbf{x}_{it}|\mathbf{y}_{it^-} = \mathbf{0})$ that matter. This implies that the survival bias is to the opposite direction for OLS and increases, rather than decreases, in the degree of persistence $\mathbf{x}_{it}$ exhibits. Moreover, since a conditional covariance rather than a difference in conditional covariances generate the survival bias, between-group heterogeneity, that is differences in the level of $\mathbf{x}_{it}$ across the units i, contribute to the bias.

Denoting the identity matrix by I, from (11), (12) and (13) then follows

$$\begin{aligned}
\mathrm{plim}(b^{\mathrm{FDC}}) &= \mathrm{plim}\left(I + \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\mathbf{x}}_{it-1}\right)\right)\tilde{\beta} \\
&\quad + \mathrm{plim}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}_{it}'\tilde{a}_i\right) \\
&= \tilde{\beta} + \text{asymptotic misscaling bias} + \text{asymptotic survival bias} \qquad (14)
\end{aligned}$$

In other words, b^{FDC} is an inconsistent estimator for $\tilde{\beta}$, that suffers from two sources of asymptotic bias. Assuming that a_i and $\Delta\mathbf{x}_{it}$ are uncorrelated in the population, one is the survival bias discussed above. The other is that $\tilde{\beta}$ enters erroneously scaled. This second source of bias originates from the first differences transformation making the conditional mean of the disturbance a function of $\mathbf{x}_{it-1}\beta$. This misscaling bias is present even for $\tilde{a}_i = 0$, that is, in the absence of any

unobserved, time-invariant individual heterogeneity. It hence does not originate from a failure to remove individual heterogeneity but from the first-differences transformation itself. Both sources of asymptotic biases disappear either for $\beta = \mathbf{0}$, or for $\mathbf{x}_{it}$ following a random walk. Even in these cases b^{FDC} is not consistent for α, since the constant converges to in probability to $\tilde{a}^c$ rather than to its unconditional counterpart. This likewise holds for period specific constants that are not consistent for true time effects.

2.5 An Adjusted First-Differences Estimator

While little can be done about the survival bias, the misscaling bias can be eliminated by appropriately rescaling b^{FDC}. This is of major importance to applied work, since in many settings the survival bias turns out to be small while misscaling bias is the overwhelmingly dominant source of bias; see section 3 for Monte-Carlo simulation results. Since the misscaling bias depends only on moments of observables, one can straightforwardly derive an adjusted first-differences estimator from (14)

$$b^{\text{FDC}}_{\text{adjust}} = \left(I + \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\Delta \mathbf{x}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\mathbf{x}}_{it-1} \right) \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\Delta \mathbf{x}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} y_{it} \right) \quad (15)$$

that does not suffer from misscaling bias even in small samples. The shape of the adjustment matrix $\left(I + \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\Delta \mathbf{x}}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\mathbf{x}}_{it-1} \right) \right)^{-1}$, that we denote $\mathbf{H}$, depends strongly on the data generating process of the variables in $\mathbf{x}_{it}$. If $\mathbf{x}_{it}$ follows a random walk, $\mathbf{H}$ converges in probability to the identity matrix I. This corresponds to our earlier result that b^{FDC} is consistent for this special case and no adjustment ist required. If $\mathbf{x}_{it}$ is however covariance stationary – that is in the population we have $\text{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it}\right) = \mathbf{Q}$ and $\text{E}\left(\mathbf{x}'_{it}\mathbf{x}_{it-1}\right) = \text{E}\left(\mathbf{x}'_{it-1}\mathbf{x}_{it}\right) = \mathbf{Q}_{\Delta}$ for all t – $\mathbf{H}$ would converge to $2I$, if the moments of $\mathbf{x}_{it}$ were not affected by conditioning on $\mathbf{y}_{it^-} = \mathbf{0}$. This renders a scaling factor of simply two an important benchmark for settings in which the considered process exhibits little selectivity. In general, the elements of $b^{\text{FDC}}_{\text{adjust}}$ are just matrix weighted sums of the elements of b^{FDC}. This implies that $b^{\text{FDC}}_{\text{adjust}}$ also rescales the survival bias in b^{FDC}. The probability limit of $b^{\text{FDC}}_{\text{adjust}}$ thus reads as follows:[11]

$$\text{plim}(b^{\text{FDC}}_{\text{adjust}}) \quad = \quad \tilde{\beta} + \text{plim} \left(\mathbf{H} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\Delta \mathbf{x}}_{it} \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \tilde{a}_i \right) \right) \quad (16)$$

Equation (16) illustrates that for $b^{\text{FDC}}_{\text{adjust}}$ – unlike b^{FDC} – survival bias is the only source of asymptotic bias.

[11] In (16) one may rewrite $\mathbf{H} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} \widetilde{\Delta \mathbf{x}}_{it} \right)^{-1}$ as $\left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \widetilde{\Delta \mathbf{x}}'_{it} [1 \; \mathbf{x}_{it}] \right)^{-1}$.

2.6 Existence of the Adjusted First-Differences Estimator

$b^{\text{FDC}}_{\text{adjust}}$ only exists if $\left(I + \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}'_{it}\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}'_{it}\widetilde{\mathbf{x}}_{it-1}\right)\right)$, that we denote $\mathbf{G}$, is non-singular. This is a non-trivial condition that may be violated even if b^{FDC} exists. This issue becomes obvious by thinking of $\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}'_{it}\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\widetilde{\Delta\mathbf{x}}'_{it}\widetilde{\mathbf{x}}_{it-1}\right)$ as a matrix that is composed of the coefficient vectors one gets from regressing each lagged explanatory variable on the contemporaneous changes of all variables in $\mathbf{x}_{it}$. If, for instance, one explanatory variable is perfectly negatively correlated with its own forward change, then $\mathbf{G}$ is singular because it contains a column of zeros. A trivial example for this is when $\mathbf{x}_{it}$ includes a dummy variable age^{min} indicating the youngest age, measured in years, observed in an individual-level yearly panel. In this case, we have

$$\text{age}^{min}_{it-1} = \begin{cases} 1 & \text{if} \quad t=2 \quad \text{and} \quad \text{age}_{i1} = \min_{it}(\text{age}_{it}) \\ 0 & \text{else} \end{cases} \tag{17}$$

$$\Delta\text{age}^{min}_{it} = \begin{cases} -1 & \text{if} \quad t=2 \quad \text{and} \quad \text{age}_{i1} = \min_{it}(\text{age}_{it}) \\ 0 & \text{else} \end{cases} \tag{18}$$

In other words, age^{min}_{it-1} is perfectly predicted by $-\Delta\text{age}^{min}_{it}$. The fact that the adjusted first-differences estimator does not allow estimating some empirical models that can be estimated using the simple first-differences or the within-transformation estimator, seems, at first glance, to be a major shortcoming of $b^{\text{FDC}}_{\text{adjust}}$. However, the non-existence of $b^{\text{FDC}}_{\text{adjust}}$ just reveals that one cannot obtain information about some model parameters of interest, even if the corresponding coefficients are seemingly identified by b^{FDC}. From (14) we see that – ignoring the survival bias for a second – b^{FDC} converges in probability to a matrix-weighted sum of the true model parameters $\tilde{\beta}$. Yet, $\tilde{\beta}_l$ receives no weight in this sum if the lth column of $\mathbf{G}$ is $\mathbf{0}$ and, in consequence, there is no way to retrieve any information about $\tilde{\beta}_l$ from $\hat{\beta}^{\text{FDC}}$.

2.7 The Variance of the Adjusted First-Differences Estimator

From (9) for the variance of the disturbance in the first-differences model with constant we get

$$\begin{aligned} \text{Var}(\varepsilon^{\text{FDC}}_{it}|a_i, \mathbf{x}_{it}, \mathbf{x}_{it-1}, \mathbf{y}_{it^-} = \mathbf{0}) &= \text{P}(y_{it}=1|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-}=\mathbf{0})\,(1-\tilde{\alpha}^c - \Delta\mathbf{x}_{it}\beta - \tilde{a}_i - \mathbf{x}_{it-1}\beta)^2 \\ &\quad + \text{P}(y_{it}=0|a_i, \mathbf{x}_{it}, \mathbf{y}_{it^-}=\mathbf{0})\,(-\tilde{\alpha}^c - \Delta\mathbf{x}_{it}\beta - \tilde{a}_i - \mathbf{x}_{it-1}\beta)^2 \\ &= (a_i + \mathbf{x}_{it}\beta)\,(1-a_i-\mathbf{x}_{it}\beta)^2 \\ &\quad + (1-a_i-\mathbf{x}_{it}\beta)\,(-a_i-\mathbf{x}_{it}\beta)^2 \\ &= (a_i + \mathbf{x}_{it}\beta)\,(1-a_i-\mathbf{x}_{it}\beta) \end{aligned} \tag{19}$$

This is the error variance of a standard linear probability model (cf. Greene, 2014, p. 727), except for the constant being individual specific. This straightforwardly follows from $\varepsilon_{it}^{\text{FDC}}$ being y_{it} minus a term which is conditional on a_i, $\mathbf{x}_{it}$, and $\mathbf{x}_{it-1}$ a constant, see (9). For the disturbance covariance we get

$$\text{Cov}\left(\varepsilon_{it}^{\text{FDC}}, \varepsilon_{it-s}^{\text{FDC}} | a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \mathbf{y}_{it^-} = \mathbf{0}\right) = 0 \quad \text{for} \quad s \geq 1 \tag{20}$$

since y_{it} is only observed conditionally on $y_{it-s} = 0$, and in consequence, conditionally on $\varepsilon_{it-s}^{\text{FDC}}$ taking one specific value.[12] Therefore the standard result (e.g. Greene, 2014, p. 302) for the variance of least squares in the presence of heteroscedasticity holds:

$$\text{Var}\left(b^{\text{FDC}} | \mathbf{X}, \mathbf{a}\right) = \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i} \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i} \sigma_{it}^2 \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right) \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i} \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1} \tag{21}$$

with σ_{it}^2 denoting $\left(a_i + \mathbf{x}_{it}\beta\right)\left(1 - a_i - \mathbf{x}_{it}\beta\right)$. Since conditional on the explanatory variables $b_{\text{adjust}}^{\text{FDC}}$ is just b^{FDC} weighted by a matrix of constants, for $\text{Var}\left(b_{\text{adjust}}^{\text{FDC}} | \mathbf{X}, \mathbf{a}\right)$ we get

$$\text{Var}\left(b_{\text{adjust}}^{\text{FDC}} | \mathbf{X}, \mathbf{a}\right) = \mathbf{H}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i} \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i} \sigma_{it}^2 \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right) \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i} \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}\right)^{-1} \mathbf{H}' \tag{22}$$

For the same reason, the asymptotic normality of the least squares estimator b^{FDC} carries over to $b_{\text{adjust}}^{\text{FDC}}$; see section 3.5 for simulations corroborating this result.

With estimates of β in hand, (22) can in principle be estimated, using $\frac{1}{T_i}\sum_{t=1}^{T_i}\left(y_{it} - \mathbf{x}_{it}\hat{\beta}_{\text{adjust}}^{\text{FDC}}\right)$ to estimate a_i. For a small T_i, however, a_i and in turn σ_{it}^2 are poorly estimated by this procedure. Moreover $\left(\hat{a}_i + \mathbf{x}_{it}\hat{\beta}_{\text{adjust}}^{\text{FDC}}\right)$ may well be negative or exceed unity, leading to invalid estimates of σ_{it}^2. In applied work, calculating a heteroscedasticity robust estimate of $\text{Var}\left(b^{\text{FDC}}\right)$ and adjusting it by $\mathbf{H}$ seems to be preferable to estimating analytical standard errors based on (22) and estimates of β and a_i. This straight forward and simple method, however, ignores the survival bias b^{FDC} and $b_{\text{adjust}}^{\text{FDC}}$ may suffer from. Using White (1980) robust standard errors is nevertheless a conservative approach, since in the presence of survival bias $\frac{1}{\sum_{i=1}^{N} T_i - 1}\sum_{i=1}^{N}\sum_{t=2}^{T_i}(e_{it}^{\text{FDC}})^2 \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}$ will asymptotically overestimate $\frac{1}{\sum_{i=1}^{N} T_i - 1}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\sigma_{it}^2 \widetilde{\Delta\mathbf{x}}_{it}'\widetilde{\Delta\mathbf{x}}_{it}$ by relying on residuals e_{it}^{FDC} that capture a bias. Naturally bootstrapping provides an alternative approach to estimating (22) that circumvents the issue of survival bias. See section 3.4 for simulations addressing the estimation of standard errors.

[12] The standard argument in favor of the within-transformation as compared to the first-differences estimator that taking first-differences brings serial correlation into the model (cf. Wooldridge, 2009, p. 430), does not apply in the considered setting.

2.8 Higher-Order Differences

The result that the adjusted first-differences estimator only suffers from survival bias critically hinges on having $\text{Cov}(a_i, \Delta \mathbf{x}_{it}) = \mathbf{0}$ in the population. Contingent on the specific application, this non-testable assumption might neither be valid nor plausible. However, assuming $\text{Cov}(a_i, \Delta^j \mathbf{x}_{it}) = \mathbf{0}$ instead, with the integer j greater than unity, may possibly be less questionable. In such settings, an adjusted estimator $b^{\text{JDC}}_{\text{adjust}}$ based on higher-order differences $\Delta^j \mathbf{x}_{it}$ can be, analogously to $b^{\text{FDC}}_{\text{adjust}}$, straightforwardly constructed. More specifically, $b^{\text{JDC}}_{\text{adjust}}$ is just an adjusted jth-differences estimator with constant term

$$
\begin{aligned}
b^{\text{JDC}}_{\text{adjust}} \quad = \quad & \left(I + \left(\sum_{i=1}^{N} \sum_{t=j+1}^{T_i} \widetilde{\Delta^j \mathbf{x}_{it}}' \widetilde{\Delta^j \mathbf{x}_{it}} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=j+1}^{T_i} \widetilde{\Delta^j \mathbf{x}_{it}}' \widetilde{(\mathbf{x}_{it} - \Delta^j \mathbf{x}_{it})} \right) \right)^{-1} \\
& \times \left(\sum_{i=1}^{N} \sum_{t=j+1}^{T_i} \widetilde{\Delta^j \mathbf{x}_{it}}' \widetilde{\Delta^j \mathbf{x}_{it}} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=j+1}^{T_i} \widetilde{\Delta^j \mathbf{x}_{it}}' y_{it} \right)
\end{aligned} \tag{23}
$$

$\Delta^j \mathbf{x}_{it}$ denotes the vector of the jth-differenced explanatory variables. For $j = 2$ we have $\Delta^2 \mathbf{x}_{it} \equiv \Delta \mathbf{x}_{it} - \Delta \mathbf{x}_{it-1}$, for $j = 3$ we have $\Delta^3 \mathbf{x}_{it} \equiv (\Delta \mathbf{x}_{it} - \Delta \mathbf{x}_{it-1}) - (\Delta \mathbf{x}_{it-1} - \Delta \mathbf{x}_{it-2})$, et cetera. Here $(\mathbf{x}_{it} - \Delta^j \mathbf{x}_{it})$ is the analogue to $\mathbf{x}_{it-1}$ in (15). It originates from the conditional mean of the disturbance in the unadjusted jth-differences estimator that – analogously to (9) – involves $\Delta^j \mathbf{x}_{it} \beta$ minus $\mathbf{x}_{it} \beta$. As before, tildes indicate that the vectors of jth-differenced explanatory variables are augmented by a constant term and that $(\mathbf{x}_{it} - \Delta^j \mathbf{x}_{it})$ is augmented by a column of zeros. Following the same line of argument as above, $b^{\text{JDC}}_{\text{adjust}}$ suffers only from survival bias but from no other source of asymptotic bias, as long as $\text{Cov}(a_i, \Delta^j \mathbf{x}_{it}) = \mathbf{0}$. Naturally, $b^{\text{JDC}}_{\text{adjust}}$ coincides with $b^{\text{FDC}}_{\text{adjust}}$ for $j = 1$, and with pooled OLS for $j = 0$. Evidently, taking higher-order differences removes much variation from the explanatory variables. The price one pays in terms of precision to establish desirable asymptotic properties under alternative, possibly weaker, assumptions is, hence, likely to be high.

3 Monte Carlo Analysis

In this section we present results from Monte Carlo (MC) simulations. For y_{it} we consider the data generation process described in section 2.1, with $\mathbf{x}_{it}$ consisting of just one variable x_{it}.[13] The slope

[13] One may not feel comfortable with considering a DGP for y_{it} that is consistent with the linear hazard model, because the linear model requires strong assumptions regarding the DGPs of $\mathbf{x}_{it}$ and a_i to guarantee $\text{P}(y_{it} = 1 | a_i, x_{it}, \mathbf{y}_{it^-} = \mathbf{0}) \in [0, 1]$. For this reason, applied researchers might primarily be interested in the performance – in terms of estimating average marginal effects – of the linear estimators when applied to data that is generated by a process that is consistent with a classical nonlinear binary outcome models such as probit or logit. The simulation results presented in the Appendix A.3 consider this case.

coefficient is $\beta = 1$.[14] We specify a_i to be iid. continuously uniformly $U(0.05, 0.15)$ distributed, implying $\mathrm{E}(a_i) = \alpha = 0.1$. We consider a short panel with $T = 5$. We examine the properties of the considered estimators for three different data generating processes for x_{it}:

(i) $x_{it}^{ST} = a_i + 0.1 + \zeta_{it}$, with $\zeta_{it} \sim$ iid. $U(-0.035, 0.035)$, i.e., x_{it}^{ST} is stationary

(ii) $x_{it}^{RW} = x_{it-1}^{RW} + \nu_{it}$, with $x_{i1} = a_i + 0.1$ and $\nu_{it} \sim$ iid. $U(-0.05, 0.05)$,
i.e., x_{it}^{RW} follows a random walk without drift

(iii) $x_{it}^{TR} = a_i + 0.075 + \eta_{it}$, with $\eta_{it} \sim$ iid. $U(0, 0.025t)$,
i.e., x_{it}^{TR} exhibits a trend and increasing variance around the trend

For all three data generating processes a_i is positively correlated with x_{it} but uncorrelated with Δx_{it} in the population.[15] Besides the estimators discussed above – that is b^{FD}, b^{FDC}, and $b^{\mathrm{FDC}}_{\mathrm{adjust}}$ – we also consider pooled ordinary least squares b^{OLS} as reference and the within-transformation estimator b^{WI}, which appears to be the most popular fixed-effects estimator in applied work and does not coincide with b^{FD} for $T > 2$.

In order to assess the large sample properties of the estimators, we choose $N = 4 \cdot 10^7$. We report the point estimates from one-shot regressions using this very large artificial sample, see Table 1. Along with the point estimates we report (heteroscedasticity robust) standard errors. Note that they are not non-parametrically generated by replicating the analysis, but are calculated following the procedure suggested in section 2.7. They are hence not meant for assessing the sampling variability of the different estimation methods by means of an MC simulation. They are only reported in order to provide some intuition on 'how distant from infinite size' the artificial sample is, as the standard errors would collapse to zero in this case.

To study the estimators' small sample properties, we choose $N = 400$. Here we replicate the regressions 10 000 times. The reported coefficients are averages over the replications and the reported standard deviations are non-parametrically calculated from the simulated distribution. They, hence, illustrate the degree to which the different estimators suffer from sampling error in the considered settings. We consider two variants for the Monte Carlo experiment. In the first we redraw a_i and x_{it} in each replication, see Table 2, upper panel. In the second we keep a_i and x_{it} fixed and only vary y_{it} in each replication, see Table 2, lower panel.

[14] This choice is simply to make the simulation results more conveniently comparable to the true parameter value. It implies that x_{it} is scaled such that a one unit change is all but a marginal change. Rescaling x_{it} appropriately would hence straightforwardly yield a β-coefficient whose magnitude would be better in line with what one would consider a marginal effect in a binary outcome model.

[15] The parameter values are chosen to align $\mathrm{P}(y_{it} = 1)$ and $\mathrm{Var}(\Delta x_{it})$ across the different data generating processes and to guarantee that the condition $a_i + x_{it}\beta \in [0, 1]$ is satisfied for any i and any $t = 1, \ldots, 5$. For the unconditional correlations we have $\mathrm{Cor}(a_i, x_{it}^{ST}) = 0.82$, $\mathrm{Cor}(a_i, x_{it}^{RW}) = 0.58$, and $\mathrm{Cor}(a_i, x_{it}^{TR}) = 0.70$.

Table 1: Monte Carlo Analysis - Large Sample Estimates

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
$x_{it}^{\boldsymbol{ST}}$ **stationary**										
$\hat{\beta}$	1.6671	0.0012	0.9024	0.0025	0.7072	0.0022	0.5008	0.0019	0.9980	0.0037
$\hat{\alpha}$	−0.0345	0.0002	0.1160	0.0005			0.2899	0.0001	0.0955	0.0007
$x_{it}^{\boldsymbol{RW}}$ **follows random walk**										
$\hat{\beta}$	1.4267	0.0009	0.9472	0.0019	1.0011	0.0022	1.0000	0.0018	0.9999	0.0018
$\hat{\alpha}$	0.0134	0.0002	0.1072	0.0004			0.2882	0.0001	0.0951	0.0004
$x_{it}^{\boldsymbol{TR}}$ **with trend and increasing variance around trend**										
$\hat{\beta}$	1.5715	0.0012	6.0363	0.0019	4.4998	0.0020	0.6725	0.0019	1.0075	0.0028
$\hat{\alpha}$	−0.0180	0.0002	−0.9154	0.0004			0.2950	0.0001	0.0936	0.0006

Notes: True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$; $N = 4 \cdot 10^7$, $T = 5$; the # of observations for x_{it}^{ST} is 71 748 906, the corresponding # of observations for x_{it}^{RW} is 71 823 746, and for x_{it}^{TR} it is 72 218 321. For b^{OLS} the #s of observations are higher by $4 \cdot 10^7$ observations, since the first wave is not eliminated by the within-transformation or the first-differences transformation. See Table A4 for simulation results using a different seed for the RNG.

3.1 Large Sample Properties

Considering the large sample results[16] presented in Table 1, consistent with x_{it} being positively correlated with the unobserved individual heterogeneity a_i, the estimated β-coefficient from b^{OLS} exhibits substantial upward bias. The results for $b^{\text{FDC}}_{\text{adjust}}$ and b^{FDC} are also in line with the analytic results derived for their large sample properties. $b^{\text{FDC}}_{\text{adjust}}$ hits the true value of β almost exactly. The simulations, hence, point to survival bias being of very little importance in the considered setting. For x_{it}^{ST} the estimate of β is marginally smaller than the true parameter, though survival bias should operate in the opposite direction. Only for x_{it}^{TR} the positive deviation from the true value of β seems to indicate a non-negligible bias of this kind. However, choosing different starting values for the random number generator regularly yields estimates much closer to unity, see Table A4 in the Appendix. Simulation results for b^{FDC} also reflect what theory predicts. No bias occurs if x_{it} is generated by a random walk. If x_{it} is stationary, b^{FDC} yields a large sample estimate that is almost exactly $\beta/2$. This is approximately the value that should occur due to the misscaling bias b^{FDC} suffers from. The very small upward deviation from 0.5 may represent survival bias. Yet, taking the estimated standard error into account, it could also be attributed to sampling variability. If the mean and variance of x_{it} are functions of time, the slope coefficient of b^{FDC} is erroneously scaled by a factor between one-half and one. Consistent with our earlier argument, the large sample estimates $b^{\text{FDC}}_{\text{adjust}}$ yields for the constant are slightly smaller than α but almost perfectly coincide with the average a_i in the estimation samples.[17] b^{FD} hits the true value of β only if x_{it} is generated by a random walk. For stationary x_{it}, it exhibits a substantial bias towards

[16] The standard errors are fairly small but still clearly different from zero. While sampling error should play a minor role in the reported point estimates, the reported standard errors illustrate that a sample of more than 100 million observations is still a clearly imperfect approximation of a sample of infinite size.

[17] The estimation sample averages of a_i are 0.0951, 0.0951, and 0.0952 for x_{it} stationary, following a random walk, and exhibiting a trend.

zero. If x_{it} has a trend, b^{FD} exhibits very poor large sample properties. This is consistent with our earlier argument that b^{FD} – unlike b^{FDC} and b^{FDC}_{adjust} – let $\tilde{\alpha}^c$ enter the disturbance, which renders $\text{plim}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}'_{it}a_i\right) \neq \mathbf{0}$ a major source of asymptotic bias, even if the selective survival driven correlation of Δx_{it} and a_i is negligibly small. The reason for this is that neither Δx_{it} nor a_i have zero mean if a_i is not demeaned by including a constant and x_{it} exhibits a trend. b^{WI} exhibits large sample properties similar to those of b^{FD}, which is not too surprising since a short panel is considered and both coincide for an extremely short panel of only two waves. First of all, b^{WI} is severely biased when x_{it} has a trend. Yet, a sizable bias towards zero is also observed when x_{it} is stationary and even when x_{it} follows a random walk. The extreme biases found for b^{FD} and b^{WI} for x_{it} with a trend would be moderated if a time trend or a set of time dummies were used as additional regressors. However, the time effects themselfs are then severely biased; see Appendix A.4 for a more detailed discussion of how including time indicators affect the results the different estimators yield.

3.2 Small Sample Properties

Turning to the small sample results, we begin by considering x_{it} and a_i random. The results in the upper panel of Table 2 show that the mean coefficients are very close to the coefficients we got from the large sample. This suggests that b^{FDC}_{adjust} does not only suffer from very little large sample bias in the considered setting but, unconditionally on a_i and x_{it}, also from little small sample bias. b^{FDC} exhibits the same kind of misscaling bias in small and in large samples. The same applies to the endogeneity bias of b^{OLS} and the biases of b^{FD} and b^{WI}, which appear to be of the same size in small samples as they are asymptotically. The standard deviations indicate that all considered estimators suffer from substantial sampling error in settings similar to the one considered here. Not surprisingly, b^{OLS} does best in this respect, while b^{FDC}_{adjust} and b^{FDC} exhibit the largest variance.[18]

Now we turn to the small sample results, with x_{it} and a_i considered as fixed, presented in the lower panel of Table 2. Whether one thinks of x_{it} and a_i as random or fixed appears to make little difference for the small sample properties of b^{OLS}, b^{FDC}, and b^{FDC}_{adjust}. In particular, the slope coefficient from b^{FDC}_{adjust} appears to suffer from rather moderate small sample bias conditional on x_i and a_i. Yet, this does not apply to b^{FD} and b^{WI}. For them the bias with x^{ST}_{it} and x^{RW}_{it} is much bigger if x_{it} and a_i are fixed. For b^{WI} the bias is even in the opposite direction than when x_{it} and a_i are random. This pattern, at first glance astonishing, is easily explained by the fact that

[18] Since b^{FDC}_{adjust} just re-scales b^{FDC}, the standard errors of both estimators are the same up to the scaling factors that apply to the respective coefficient.

Table 2: Monte Carlo Analysis - Small Sample Estimates

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†
x_{it} and a_i random										
x_{it}^{ST} **stationary**										
$\hat{\beta}$	1.6755	0.3808	0.9208	0.7885	0.7240	0.7038	0.5133	0.5902	1.0167	1.1728
$\hat{\alpha}$	−0.0356	0.0746	0.1128	0.1549			0.2903	0.0171	0.0923	0.2286
x_{it}^{RW} **follows random walk**										
$\hat{\beta}$	1.4278	0.3004	0.9485	0.6089	1.0068	0.69504	1.0019	0.5862	1.0027	0.5856
$\hat{\alpha}$	0.0138	0.0582	0.1068	0.1195			0.2887	0.0170	0.0954	0.1131
x_{it}^{TR} **with trend and increasing variance around trend**										
$\hat{\beta}$	1.5763	0.3654	6.0427	0.6069	4.5072	0.67781	0.6691	0.6155	0.9940	0.9147
$\hat{\alpha}$	−0.0186	0.0733	−0.9167	0.1167			0.2950	0.0187	0.0965	0.1909
x_{it} and a_i fixed										
x_{it}^{ST} **stationary**										
$\hat{\beta}$	1.6443	0.3826	1.3168	0.7160	0.8548	0.6678	0.5351	0.5790	1.0326	1.1189
$\hat{\alpha}$	−0.0310	0.0743	0.0324	0.1390			0.2853	0.0168	0.0865	0.2161
x_{it}^{RW} **follows random walk**										
$\hat{\beta}$	1.4208	0.3227	1.6595	0.5408	1.5261	0.6514	0.9350	0.5921	0.9807	0.6203
$\hat{\alpha}$	0.0125	0.0627	−0.0344	0.1054			0.2852	0.0166	0.0969	0.1209
x_{it}^{TR} **with trend and increasing variance around trend**										
$\hat{\beta}$	1.5638	0.3795	5.9851	0.5921	4.5432	0.6561	0.6581	0.6064	0.9792	0.9023
$\hat{\alpha}$	−0.0172	0.0751	−0.8950	0.1113			0.2903	0.0177	0.0973	0.1855

Notes: True coefficient values: $\boldsymbol{\beta} = \mathbf{1}$, $\boldsymbol{\alpha} = \mathbf{0.1}$; $N = 400$, $T = 5$; 10 000 replications. †S.D. denotes the empirical standard deviation of the coefficient in the simulated sample. In order to interpret these values in terms of standard errors for the respective mean-estimator, one has to multiply the value of the S.D. by $10\,000^{-0.5} = 0.001$. See Table A5 for simulation results using a different seed for the RNG.

$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}'a_i$ may well substantially deviate from zero in a small and fixed sample of data, even if $\mathbf{x}_{it}$ does not exhibit a trend and $\Delta\mathbf{x}_{it}$ is not correlated with a_i in the population. To illustrate this, running the same simulation with a different seed for the random number generator (RNG) yields very different biases for b^{FD} and b^{WI}, see Table A5, lower panel in the Appendix. b^{FD} and b^{WI} are for this reason prone to substantial small sample bias.

3.3 Analyzing the Survival Bias

The simulation results discussed above provide little evidence for survival bias being a significant issue for the considered estimation methods, $b^{\text{FDC}}_{\text{adjust}}$ in particular. This, however, might just be an artifact of the choice of the parameters, and survival bias might be a more important issue in different settings. To make the simulation setting more prone to survival bias we increase the variance of the unobserved heterogeneity. The simulation results presented below – see Table 3 – originate from simulations in which a_i is sampled from the $U(0, 0.5)$ distribution rendering its standard deviation 0.144 in the population.[19] This value substantially exceeds 0.029, which is the corresponding value for the simulations discussed in sections 3.1 and 3.2. Allowing for larger

[19]The standard deviation of a continuous uniformly distributed random variable a is $(a^{max} - a^{min})/\sqrt{12}$. The upper limit for the standard deviation of a_i is 0.5. In this very special case, a_i is Bernoulli $b(0.5)$ distributed. For $\text{Cov}(a_i, \mathbf{x}_{it}) = \mathbf{0}$ in the population, $\beta = \mathbf{0}$ must hold to satisfy $a_i + \mathbf{x}_{it}\beta \in [0, 1]$ for all i and t. In consequence y_{it} deterministically depends on a_i.

values of a_i decreases the survival rate in the artificial sample. For this reason we adjusted the number of units, which is now $N = 10^8$, and also the length of the panel, which is now $T = 3$. Accordingly, in this section we analyze the properties of the estimators only in a large sample. The DGP for y_{it} is the same as above, with unity still being the true value of β.

As another important deviation from the previous design, we consider DGPs for which a_i and x_{it} are uncorrelated in the population. This makes survival bias the only source of bias for b^{OLS}, which allows comparing this type of bias across b^{OLS} and $b^{\text{FDC}}_{\text{adjust}}$. The discussion will focus on these two estimation methods, since the remaining three estimators are subject to more than one source of bias. As before, we consider three data generating processes for x_{it}, which however deviate from the hitherto considered ones in some respects. To make the design more flexible, we sample x_{it} from the beta distribution rather than from the continuous uniform distribution.[20] Moreover, instead of considering a linear trend in x_{it}, we distinguish two stationary process. For one, all variation in x_{it} is purely transitory, while for the other a substantial share of the variation is between the units i. Finally we also consider a random walk for the DGP of x_{it}, which in this simulation involves a drift[21]:

(i) $x_{it}^{STT} = \frac{1}{2}\zeta_{it}$, with $\zeta_{it} \sim$ iid. $B(6,2)$,

(ii) $x_{it}^{STB} = \frac{1}{4}\mu_i + \frac{1}{4}\eta_{it}$, with $\mu_i \sim$ iid. $B(6,2)$ and $\eta_{it} \sim$ iid. $B(6,2)$,

(iii) $x_{it}^{RWD} = x_{it-1}^{RWD} + \left(\frac{1}{6}\nu_{it} - \frac{1}{12}\right)$, with $x_{i1}^{RWD} = \frac{1}{6} + \frac{1}{6}\nu_{i1}$ and $\nu_{it} \sim$ iid. $B(6,2)$

The results displayed in the first panel of Table 3 (upper panel) indicate that in the considered setting $b^{\text{FDC}}_{\text{adjust}}$ is subject to a moderate, yet non-negligible, upward survival bias if x_{it} is stationary. Yet, if x_{it} follows a random walk the simulations do not reveal any large sample bias as predicted by theory. In contrast, survival bias seems only to be an issue for b^{OLS} if x_{it} follows a random walk. In this case the downward bias is however severe and substantially exceeds the survival bias $b^{\text{FDC}}_{\text{adjust}}$ exhibits with stationary x_{it}. Inconsistent with the prediction from theory, b^{OLS} seems not to suffer from survival bias if x_{it} is stationary with a major share of its variation being between groups. This puzzle can be explained by the majority of observations that enter OLS are from the initial period that by design does not suffer from survival driven selection. To address this issue, the lower panel of Table 3 displays result from regressions excluding the initial panel wave. In real data applications this corresponds to analyzing data subject to left censoring, i.e. some – or even all – units are not observed from the very beginning of the process under scrutiny. With the initial period excluded, the results for b^{OLS} exhibit the expected pattern. While no bias occurs

[20] See Appendix A.5 for results considering other beta distributions than $B(6,2)$.

[21] Since the $B(6,2)$ distribution has a mean of 0.75 the drift parameter is 0.0417.

Table 3: Monte Carlo Analysis - **Survivor Bias**, Large Sample Estimates

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†
initial period included										
x_{it}^{STT} **stationary, all variation in** x_{it} **transitory**										
$\hat{\beta}$	0.9998	0.0005	1.3172	0.0012	1.2045	0.0009	0.5418	0.0006	1.1346	0.0013
$\hat{\alpha}$	0.2264	0.0002	0.1074	0.0005			0.5500	0.0001	0.1317	0.0005
x_{it}^{STB} **stationary, between-group variation in** x_{it}										
$\hat{\beta}$	0.9991	0.0008	1.3554	0.0025	1.2403	0.0019	0.5571	0.0013	1.1352	0.0027
$\hat{\alpha}$	0.2267	0.0003	0.0936	0.0010			0.5511	0.0001	0.1322	0.0010
x_{it}^{RWD} **follows random walk with drift**										
$\hat{\beta}$	0.4653	0.0009	8.5661	0.0015	9.9046	0.0013	1.0006	0.0025	1.0004	0.0025
$\hat{\alpha}$	0.3940	0.0003	−2.1406	0.0005			0.4947	0.0001	0.1917	0.0009
initial period excluded (left censoring)										
x_{it}^{STT} **stationary, all variation in** x_{it} **transitory**										
$\hat{\beta}$	1.0002	0.0009	1.0971	0.0016	1.0971	0.0016	0.5229	0.0012	1.0918	0.0024
$\hat{\alpha}$	0.1821	0.0004	0.1458	0.0008			0.5220	0.0001	0.1189	0.0009
x_{it}^{STB} **stationary, between-group variation in** x_{it}										
$\hat{\beta}$	0.9282	0.0013	1.1271	0.0034	1.1271	0.0034	0.5405	0.0024	1.1014	0.0049
$\hat{\alpha}$	0.2089	0.0005	0.1352	0.0016			0.5214	0.0001	0.1168	0.0018
x_{it}^{RWD} **follows random walk with drift**										
$\hat{\beta}$	0.7483	0.0015	9.8897	0.0023	9.8897	0.0023	0.9961	0.0045	0.9960	0.0045
$\hat{\alpha}$	0.2786	0.0005	−2.8706	0.0010			0.4941	0.0002	0.1658	0.0017

Notes: True coefficient values: $\boldsymbol{\beta} = \mathbf{1}$, $\boldsymbol{\alpha} = \mathbf{0.25}$; $N = 10^8$, $T = 3$. The initial wave # one naturally comprises $N = 10^8$ observations; observation #s for wave two are: 37 503 419 (x_{it}^{STT}), 37 492 118 (x_{it}^{STB}), and 45 837 313 (x_{it}^{RWD}); observation #s for wave three are: 16 141 100 (x_{it}^{STT}), 16 271 318 (x_{it}^{STB}), and 21 235 298 (x_{it}^{RWD}).

for purely transitory variation in x_{it}, a downward bias is found for both, constant between-group variation and persistence in the DGP of x_{it}.

To complete this discussion, we have a brief look on the remaining estimators that – unlike b^{OLS} and $b^{\text{FDC}}_{\text{adjust}}$ – suffer from both survivor and misscaling bias in this setting. For a stationary regressor, b^{FDC} is throughout close to the one-half of the true value of β, suggesting that misscaling is the dominant source of bias in the considered setting. With x_{it} following a random walk, b^{FDC} as expected does not exhibit a significant bias. b^{WI} and b^{FD} appear to be moderately biased – yet slightly more than $b^{\text{FDC}}_{\text{adjust}}$ – for stationary x_{it}. However, with x_{it} having a drift they exhibit the same weird behavior as with the explanatory variables having a trend.

Finally, we dig deeper into the question of what role the variance of the unobserved heterogeneity plays for the bias, the survival bias in particular. To this end we run a simulation that generalizes the data generating process (ii) considered above. Now we consider a_i to be sampled from the $U(0, q)$ distribution, while we have $x_{it}^{STB} = \left((1-q)/2\right)\mu_i + \left((1-q)/2\right)\eta_{it}$ for the regressor, with μ_i and η_{it} as above being sampled from the beta $B(6, 2)$ distribution. We vary q between 0 and 0.96 and thus consider values for $\sqrt{\text{Var}(a_i)}$ in the range between 0 and 0.277. In order not to violate the condition $a_i + x_{it} \in [0, 1]$, considering a large variance for a_i requires considering a small one for x_{it}. This is why q also enters the DGP for x_{it}. Figure 1 plots the estimated slope coefficients of the considered estimators as function of $\sqrt{\text{Var}(a_i)}$. Since we consider regressions

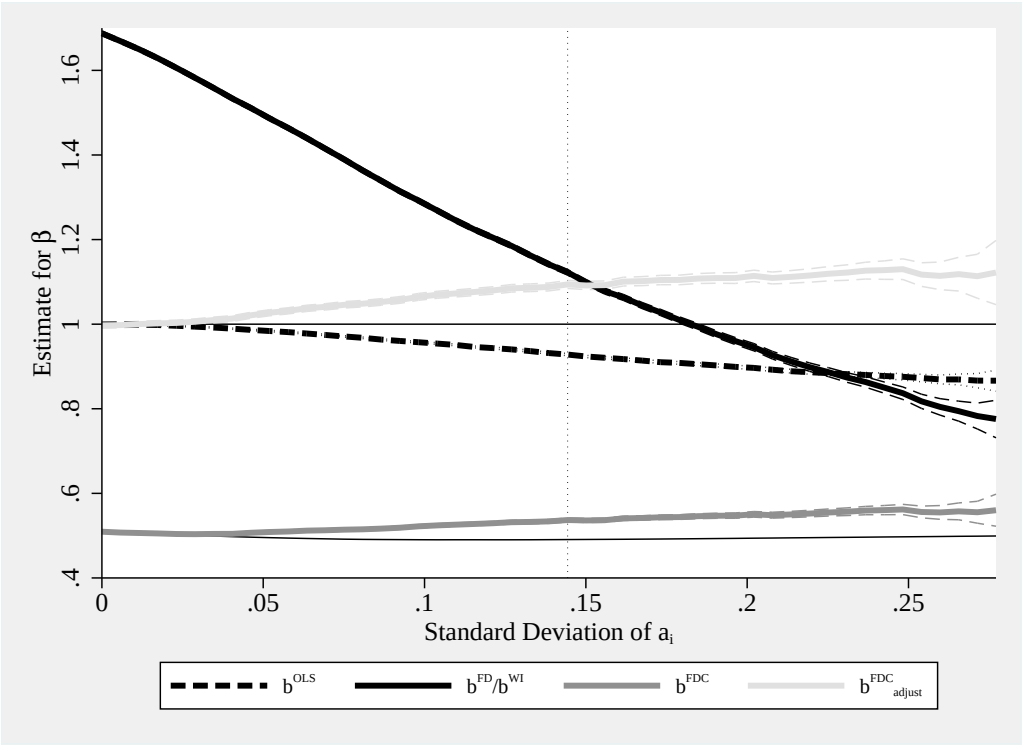

Figure 1: Estimated β coefficients as functions of $\sqrt{\text{Var}(a_i)} = q/\sqrt{12}$. DGPs of a_i and x_{it}: a_i sampled from the $U(0,q)$ distribution; $x_{it}^{STB} = ((1-q)/2)\,\mu_i + ((1-q)/2)\,\eta_{it}$ with μ_i and η_{it} independently sampled from the beta $B(6,2)$ distribution. q varied in the range between 0 and 0.96. Dashed subsidiary lines mark 95 percent confidence intervals. The thin solid subsidiary lines indicate the true coefficient value $\beta = 1$ and the β-element of $\mathbf{G}\tilde{\beta}$, respectively. The vertical dotted line indicates the corresponding results in Table 3 (lower panel, middle row). See Appendices A.5 and A.6 for simulations considering alternative DGPs for x_{it} and a_i. **Source:** Own simulations.

with the initial period excluded, i.e only two waves enter the estimation sample, b^{WI} and b^{FD} coincide. Dashed lines mark estimated 95 percent confidence intervals. They get wider with increasing values of q since x_{it} then exhibits less and less variation. The vertical dotted line indicates $\sqrt{\text{Var}(a_i)} = 0.144$ that is the parameter value for which results are reported in Table 3 (lower panel, third row). The upper thin solid line marks the true parameter value $\beta = 1$. The lower thin solid line marks the value b^{FDC} would take, if misscaling bias would be its sole source of error, i.e. the β-element of $\mathbf{G}\tilde{\beta}$. This line is very close to – but does not perfectly coincide with – the benchmark value of 0.5. The Appendices A.5 and A.6 present results for simulations that vary this design in two dimensions: (i) beta distributions other than $B(6,2)$ are considered, (ii) for a_i a Bernoulli rather an a continuous uniform distribution is assumed.[22]

[22]Simulation results in the Appendix (Fig. A1, bottom row, right; Fig. A2, bottom row) seem to suggest that b^{FDC} does not suffer from misscaling bias, if x_{it} is binary, and no unobserved heterogeneity enters the DGP. Yet this is an artifact of the simulation design which in this specific case specifies $\text{P}(y_{it} = 1 | x_{it}, \mathbf{y}_{it^-} = \mathbf{0}) = x_{it}$. Since x_{it} is either zero or one, the outcome is deterministically linked to x_{it}.

Turning back to Figure 1, for $\text{Var}(a_i) = 0$, i.e. in the absence of any unobserved heterogeneity, b^{OLS} and $b^{\text{FDC}}_{\text{adjust}}$ hit the true parameter value of 1 almost perfectly. This does not apply to b^{FD} and b^{FDC}, which are severely biased. This illustrates that the misscaling bias originates from the data transformation itself, not from its failure to remove unobserved heterogeneity. The vertical distance between the two thin solid subsidiary lines is the misscaling bias in b^{FDC} that is eliminated by $b^{\text{FDC}}_{\text{adjust}}$. Hence for little variation in a_i, misscaling is the almost sole source of bias in b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ is close to asymptotic unbiasedness. If, however, the variance of the unobserved heterogeneity increases, the survival bias kicks in. This does not only hold for b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ but also for b^{OLS}. Yet, as predicted, for the latter the bias operates in the opposite direction. With substantial survival bias in b^{FDC} – in Figure 1 this is the vertical distance between b^{FDC} and the lower thin subsidiary line – $b^{\text{FDC}}_{\text{adjust}}$ does not hit the true parameter value. Eliminating the misscaling bias rather comes to the cost of re-scaling the survival bias in b^{FDC}. Nevertheless, according to our simulations, misscaling is the dominant source of bias in b^{FDC}, even if the variance of a_i is big. Thus adjusting the first difference estimator with constant still reduces the asymptotic bias substantially. This suggests that using $b^{\text{FDC}}_{\text{adjust}}$ is advisable even in settings that are prone to survival bias which is not cured by the suggested adjustment. Moreover, the survival bias in $b^{\text{FDC}}_{\text{adjust}}$ seems to be of similar magnitude of the survival bias in b^{OLS}, yet as discussed above, this depends on the dynamic properties of the DGP of x_{it}. The behavior of b^{FD} turns out to be rather strange in the considered simulation. As indicated by Figure 1, depending on the variance of the unobserved heterogeneity b^{FD} may exhibit a substantial upward, a substantial downward bias, or no bias at all. Yet, this pattern turns out to be very sensitive even to minor changes of the DGP for x_{it}; see Appendix A.5 for simulation results using slightly altered DGPs for x_{it}. Hence using the first-differences or the within-transformation estimator without constant – or without a saturated set of wave indicators, respectively – is clearly not advisable.

3.4 Methods for Estimating Standard Errors

We now use the Monte Carlo simulation to examine different methods for estimating standard errors for $b^{\text{FDC}}_{\text{adjust}}$. The first column of Table 4 just lists the Monte Carlo simulated, small sample standard errors already reported in Table 2 (rightmost column, lower panel). They are compared to the mean estimated standard errors obtained from 10 000 Monte Carlo replications. That is, in each replication – with x_{it} and a_i kept fixed but y_{it} replaced – standard errors for $b^{\text{FDC}}_{\text{adjust}}$ and $a^{\text{FDC}}_{\text{adjust}}$ are calculated and the averages over 10 000 replications are reported in Table 4, columns 2–5. Four methods for calculating the standard errors are compared: (i) equation (22), using the true values of β and a_i; since the true parameter values are unknown in real data applications, this variant

Table 4: Monte Carlo Analysis - Estimated Standard Errors

	MC simulated	$\widehat{\text{se}}_{\text{analytic}}(b^{\text{FDC}}_{\text{adjust}})$		**H**-adjusted $\widehat{\text{se}}_{\text{robust}}(b^{\text{FDC}})$	
		true a_i and β	$\hat{a}_i$ and $\hat{\beta}$	White	cluster robust
x_{it}^{ST} **stationary**					
$\widehat{\text{se}}(b^{\text{FDC}}_{\text{adjust}})$	1.1189	1.1118	0.8734	1.1187	1.1172
$\widehat{\text{se}}(a^{\text{FDC}}_{\text{adjust}})$	0.2161	0.2149	0.1683	0.2162	0.2161
x_{it}^{RW} **follows random walk**					
$\widehat{\text{se}}(b^{\text{FDC}}_{\text{adjust}})$	0.6203	0.6180	0.4849	0.6236	0.6234
$\widehat{\text{se}}(a^{\text{FDC}}_{\text{adjust}})$	0.1209	0.1201	0.0936	0.1212	0.1218
x_{it}^{TR} **with trend and increasing variance around trend**					
$\widehat{\text{se}}(b^{\text{FDC}}_{\text{adjust}})$	0.9023	0.8963	0.6500	0.9003	0.9053
$\widehat{\text{se}}(a^{\text{FDC}}_{\text{adjust}})$	0.1855	0.1848	0.1335	0.1856	0.1867

Notes: True coefficient values: $\boldsymbol{\beta} = \mathbf{1}$, $\boldsymbol{\alpha} = \mathbf{0.1}$; $N = 400$, $T = 5$; x_{it} and a_i fixed; 10 000 replications. See Table A6 for simulation results using a different seed for the RNG.

serves as benchmark but is not an applicable method; (ii) equation (22), using the estimates $\hat{\beta}^{\text{FDC}}_{\text{adjust}}$ and $\hat{a}_i$; (iii) obtaining White (1980) heteroscedasticity robust standard errors for b^{FDC} and adjusting them by **H**; (iv) obtaining cluster robust standard errors for b^{FDC}, with clustering by i, and adjusting them by **H**.

The pattern of results is quite clear. The Monte Carlo simulated standard errors and the analytical ones that use the true parameter values almost coincide. This is what one should expect. The estimated standard errors that are based on robust variance–covariance estimation also yield very similar results on average. This means that obtaining conventional robust standard errors from a linear regression of y_{it} on $\widetilde{\Delta \mathbf{x}}_{it}$ and adjusting them appropriately, appears to be a reliable approach to calculating standard errors for $b^{\text{FDC}}_{\text{adjust}}$. Whether White (1980) or cluster robust standard errors are used does not make a significant difference. This does not come as surprise since the errors are known to be uncorrelated in the considered setting, see (20). In contrast, calculating analytical standard errors based on the estimates $\hat{\beta}^{\text{FDC}}_{\text{adjust}}$ and $\hat{a}_i$ yields results that severely underestimate the sampling variability in $b^{\text{FDC}}_{\text{adjust}}$. This is explained by the fact that $\frac{1}{T_i}\sum_{t=1}^{T_i}\left(y_{it} - \mathbf{x}_{it} b^{\text{FDC}}_{\text{adjust}}\right)$ is a poor estimator for a_i, and by numerous invalid variance estimates. The latter forces us to use $\max\left(0, (1 - \hat{a}_i - \mathbf{x}_{it}\hat{\beta}^{\text{FDC}}_{\text{adjust}})(\hat{a}_i + \mathbf{x}_{it}\hat{\beta}^{\text{FDC}}_{\text{adjust}})\right)$, rather than $(1 - \hat{a}_i - \mathbf{x}_{it}\hat{\beta}^{\text{FDC}}_{\text{adjust}})(\hat{a}_i + \mathbf{x}_{it}\hat{\beta}^{\text{FDC}}_{\text{adjust}})$, as the estimate of σ^2_{it}. This method for estimating standard errors should not be used in applied work.

The results presented in Table 4 originate from a simulation design that generates almost no survival bias in $b^{\text{FDC}}_{\text{adjust}}$, and might hence be of little relevance to settings in which survival bias is a significant issue. To address this concern, we rerun the above simulation using the design already used in section 3.3 to analyze the survival bias. As the only deviation of that design, we now consider a small sample consisting of only $N = 1\,000$ observations.[23] We consider regressions that include the initial period. Results are presented in Table 5. Though $b^{\text{FDC}}_{\text{adjust}}$ suffers

[23] Just as for the results presented in Table 3 we use $q = 0.5$, i.e. the population standard deviation of a_i is 0.144.

Table 5: MC Analysis - Estimated Standard Errors, significant **Survivor Bias**

	MC simulated	$\hat{se}_{\text{analytic}}(b^{\text{FDC}}_{\text{adjust}})$		**H**-adjusted $\hat{se}_{\text{robust}}(b^{\text{FDC}})$	
		true a_i and β	$\hat{a}_i$ and $\hat{\beta}$	White	cluster robust
x_{it}^{STT} **stationary, all variation in** x_{it} **transitory**					
$\hat{se}(b^{\text{FDC}}_{\text{adjust}})$	0.4100	0.3997	0.3444	0.4166	0.4099
$\hat{se}(a^{\text{FDC}}_{\text{adjust}})$	0.1541	0.1502	0.1271	0.1565	0.1544
x_{it}^{STB} **stationary, between-group variation in** x_{it}					
$\hat{se}(b^{\text{FDC}}_{\text{adjust}})$	0.8624	0.8474	0.7251	0.8771	0.8673
$\hat{se}(a^{\text{FDC}}_{\text{adjust}})$	0.3227	0.3169	0.2698	0.3280	0.3246
x_{it}^{RWD} **follows random walk with drift**					
$\hat{se}(b^{\text{FDC}}_{\text{adjust}})$	0.8180	0.8030	0.6803	0.8327	0.8332
$\hat{se}(a^{\text{FDC}}_{\text{adjust}})$	0.2843	0.2790	0.2356	0.2893	0.2898

Notes: True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$; $N = 1\,000$, $T = 3$; x_{it} and a_i fixed; 10 000 replications.

from non-negligible asymptotic survival bias in this setting – except for the case that x_{it} follows a random walk – our earlier results in qualitative terms still apply. Most importantly, White (1980) robust standard errors are close to their Monte Carlo simulated counterparts. As predicted by theory, the former are slightly larger. Yet, the deviation from the simulated ones is of fairly small magnitude.[24] This suggests obtaining conventional robust standard errors from simple first difference estimation and adjusting them by **H** as a reasonable and conservative method for estimating standard errors in applied work, even if $b^{\text{FDC}}_{\text{adjust}}$ suffers from survival bias.

3.5 Asymptotic Normality

Related to the inference issues discussed in section 3.4, we finally present simulation results which address the asymptotic distributions of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$. More specifically, we test whether the simulations yield normally distributed coefficient estimates as predicted by theory; see section 2.7. We rely on the same simulation designs as in section 3.4, except for the values of N. In order to preserve the spirit of simulating an asymptotic distribution, we increase N by the factor 1 000 relative to the simulations for which results are presented in the Tables 4 and 5.[25] More specifically, N is $4 \cdot 10^5$ for the variant with negligible survival bias and 10^6 for the variant with sizable survival bias.[26] Figure 2, upper panel, displays the simulated distribution of the slope coefficient for the former setting, in which survival bias is a negligible issue. The estimated kernel densities look very much 'normal like'. Moreover – presumably more important – statistical tests (D'Agostino et al., 1990; Doornik and Hansen, 2008; Bera et al., 2016; Kolmogorov-Smirnov) provide no evidence for deviations from normality. See the p-values reported in Table 6, left panel.

[24] For larger sample sizes, the deviation of simulated and White (1980) based estimated standard errors gets even smaller.

[25] Yet, even for the rather small sample sizes considered in section 3.4, the simulated distributions of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ do not exhibit a sizable deviation from normality; see Appendix A.7, Figure A3.

[26] Choosing very large values for N, like the ones considered in the sections 3.1 and 3.3, and at the same time replicating the regression analysis 10 000 times would result in excessive run times for the simulations.

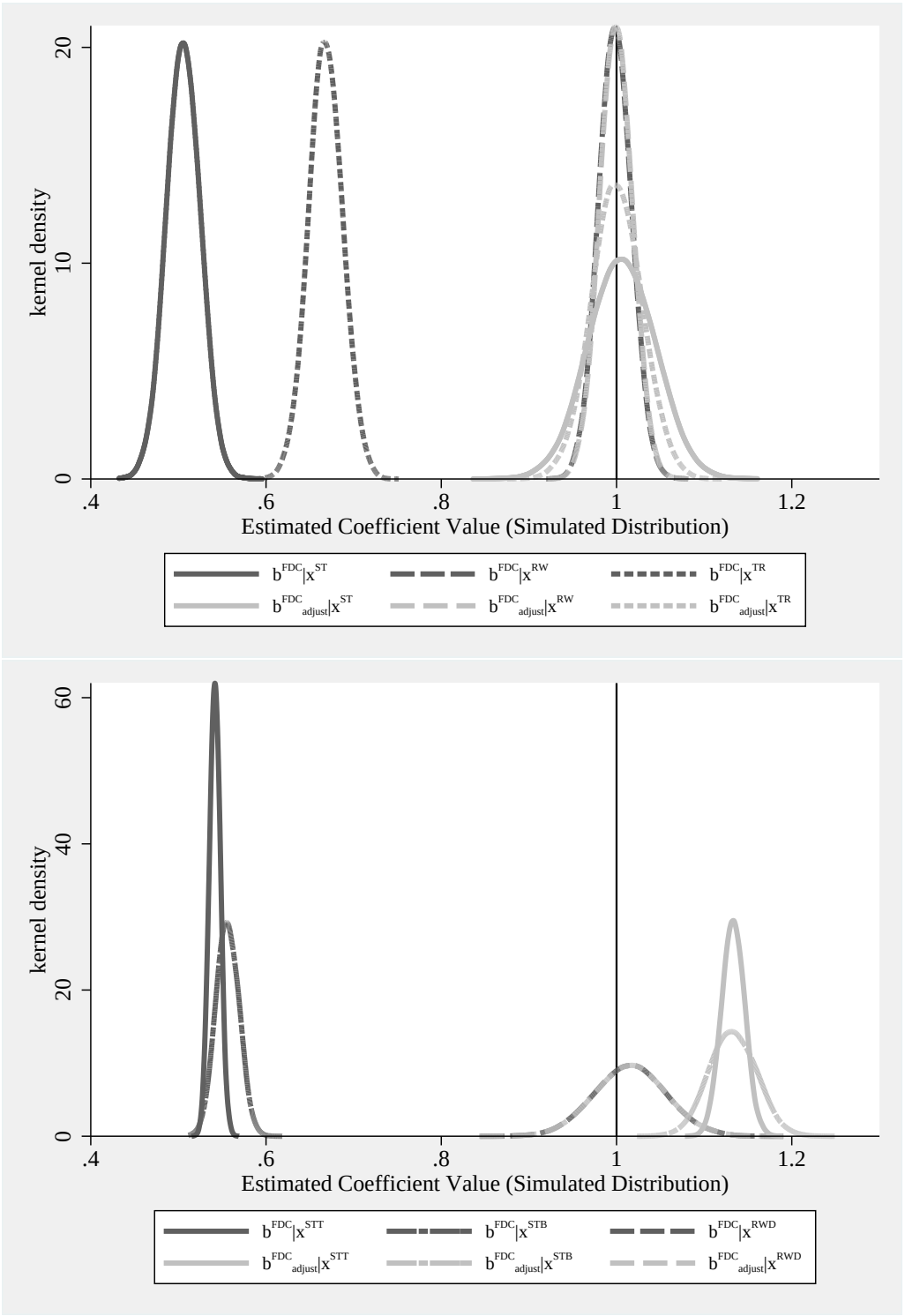

Figure 2: Monte Carlo simulated distribution of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ for different DGPs of x_{it} based on 10 000 replications; **upper panel**: same simulation design (**no** significant **survival bias**) as for the results in Table 1 (right-most columns) and Table 4 (left-most column), except for the sample size N, which is $4 \cdot 10^5$; **lower panel**: same simulation design (**significant survival bias**) as for the results in Table 3 (upper panel, right-most columns) and Table 5 (left-most column), except for the sample size N, which is 10^6. The thin vertical subsidiary lines mark the true coefficient value 1; see Appendix A.7 for corresponding Figures considering smaller samples. **Source:** Own simulations.

Table 6: MC Analysis - Tests for Normality of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ (p-values)

	no significant survival bias[†]			**significant survival bias**[‡]		
	x_{it}^{ST}	x_{it}^{RW}	x_{it}^{TR}	x_{it}^{STT}	x_{it}^{STB}	x_{it}^{RWD}
Skewness and Kurtosis test (D'Agostino et al., 1990)[♭]						
b^{FDC}	0.6996	0.2351	0.8704	0.8984	0.1968	0.8798
$b^{\text{FDC}}_{\text{adjust}}$	0.6734	0.2039	0.8781	0.8503	0.2786	0.8969
Skewness and Kurtosis test (Doornik and Hansen, 2008)[♮]						
b^{FDC}	0.6823	0.2436	0.8607	0.8969	0.1865	0.8852
$b^{\text{FDC}}_{\text{adjust}}$	0.6547	0.2128	0.8680	0.8491	0.2687	0.9031
Quantile-Mean Covariance test (Bera et al., 2016)[♯]						
b^{FDC}	0.8868	0.3524	0.1066	0.9890	0.0130	0.8102
$b^{\text{FDC}}_{\text{adjust}}$	0.9036	0.2076	0.1260	0.9911	0.0329	0.8929
Kolmogorov-Smirnov test[§]						
b^{FDC}	0.9889	0.8821	0.6704	0.9574	0.1778	0.8798
$b^{\text{FDC}}_{\text{adjust}}$	0.9796	0.9540	0.6953	0.9525	0.2559	0.9704

Notes: [†]same simulation design as for the results in Table 1 (right-most columns) and Table 4 (left-most column), yet $N = 4 \cdot 10^5$; [‡]same simulation design as for the results in Table 3 (upper panel, right-most columns) and Table 5 (left-most column), yet $N = 10^6$; DGPs for x_{it} are: x_{it}^{ST} is stationary, x_{it}^{RW} follows random walk, x_{it}^{TR} has trend and increasing variance around trend, x_{it}^{STT} is stationary, with all variation being transitory, x_{it}^{STB} is stationary, with between-group variation, x_{it}^{RWD} follows random walk with drift; 10 000 replications; [♭]omnibus test; [♮]Stata® implementation by Baum and Cox (2001) used; [♯]p-values for test based on the T_3-statistic and $\epsilon = 0.1$, Stata® implementation by Alejo et al. (2016) used; [§]known to have little power as a test for normality; **normality** is the **null** for all considered tests.

The lower panel of Figure 2 displays the simulated distributions of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ for the alternative setting, which corresponds to Table 3 (upper panel) and Table 5 and in which survival bias plays a significant role. This is graphically depicted by the distributions of $b^{\text{FDC}}_{\text{adjust}}$ not being centered at the true coefficient value of one, except for the case that x_{it} follows a random walk. Yet, the simulated distributions still look very much 'normal like'. This impression is warranted by statistical tests. See the p-values reported in Table 6, right panel.[27] The simulations, thus, confirm asymptotic normality of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$, which allows for conventional inference after using the adjusted first-differences for estimation.

4 An Application to Real Data

The empirical application presented in this section is directly based on Brown and Laschever (2012). More specifically, as the first step we replicate the results of one of their empirical models (Brown and Laschever, 2012, page 104; table 4, column 8). Subsequently, we compare these results to those we get from applying the estimators discussed in the previous sections. Thanks to the fact that the data and the code are available from the web page of the American Economic Association[28], replication of the original results is straightforward. We provide just very limited information about the analysis of Brown and Laschever (2012). Readers interested in the details of their paper, results from further empirical models in particular, are referred to the original article.

[27] The single occurrence (quantile-mean covariance test, x_{it}^{STB}) of a small p-value may well be attributed to type I error.

[28] `https://www.aeaweb.org/aej/app/data/2011-0132_data.zip`

The analysis of Brown and Laschever (2012) is concerned with the retirement behavior of school teachers in the Los Angeles Unified School District (LAUSD), in particular with the questions of how their retirement decisions are affected by the retirement behaviour of peers (teachers working at the same school). The identification rests on exogenous variation in the financial incentives for retirement that was induced by two unexpected pension reforms. Teachers were heterogeneously affected by these reforms, which allows using the reform-induced changes in financial incentives as instruments for peers' retirement behavior. In comparing different methods for estimation, we focus on one – among numerous other specifications for which results are reported in Brown and Laschever (2012) – reduced form model for simplicity.[29] In this specification, information from three panel waves is used to explain the dummy variable 'retirement', indicating that a teacher retires in the respective period, by: (i) lagged average changes in pension wealth (present value of future pension income, Brown and Laschever, 2012, p. 94) of peer teachers, which serves as instrumental variable (Table 7, first panel), (ii) teacher-specific variables capturing the changes in own financial incentives for retirement (Table 7, second and third panel), (iii) school-level controls (Table 7, fourth panel), (iv) age indicators (Table 7, fifth panel), (v) panel wave (academic year) indicators (Table 7, sixth panel), and (vi) teacher fixed effects. Since retirement is an absorbing state and teachers are no longer observed after they have retired, this empirical model fits into the considered framework very well.

Columns 1 and 2 of Table 7, denoted b^{WI}, just replicate[30] the analysis of Brown and Laschever (2012) for which the popular within-transformation estimator was used. In the original article estimated coefficients are only reported for the explanatory variables in the first and the second panel. The most important result is that the coefficient of 'lagged change in pension wealth of peers' is positive and statistically significant at the 5 percent level. This, though just marginally significant, also holds for the corresponding variable that considers a lag of two years. Hence retirement incentives to which other teachers are exposed matter, conditionally on the own incentives, for the own decision to retire. This is key to the identification of peer effects by Brown and Laschever (2012). The estimated coefficients in the second panel are consistent with theory. An increase in own pension wealth increases the probability of retirement. The coefficient attached to 'change in own peak value' (option value of postponing retirement, Brown and Laschever, 2012, p. 96), as to be expected, is negative. Yet, it lacks statistical significance. We compare the results

[29] To keep things simple, we consider the reduced form model rather than the two-stage least-squares estimation.

[30] We use a differently constructed set, yet the same number, of age indicators to parameterize the baseline hazard. This does not change the nature of the model at all. The estimated age coefficients are, however, more conveniently interpreted as they directly capture the increase in the retirement baseline hazard when a teacher gets older by one year. Moreover, we exclude from the estimation sample two teachers, whose reported ages are obviously incorrect. For one, the age increases by several years from one year to the next, for the other, the age even decreases. Excluding these six observations has virtually no impact on the estimated coefficients.

Table 7: Brown and Laschever (2012) Reduced Form Model and Alternative Empirical Models

	$b^{\mathbf{WI}}$ **(replication)**‡		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
change in pension wealth of peers ($t-1$)	0.003^{**}	0.001	0.003^{**}	0.001	-0.007	0.095
change in pension wealth of peers ($t-2$)	0.002^{*}	0.001	0.002	0.001	-0.004	0.054
change in own pension wealth	0.033^{***}	0.011	-0.003	0.009	-0.005	0.041
change in own peak value	-0.002	0.002	-0.002^{*}	0.001	-0.005^{*}	0.003
salary	0.045^{***}	0.015	0.037^{**}	0.016	0.048	0.085
years of service in LAUSD (squared)	0.002^{***}	0.000	0.002^{***}	0.000	0.000	0.000
av. age of teachers aged ≥ 55 at school	-0.002	0.003	-0.002	0.003	0.002	0.025
av. service of teachers aged ≥ 55 at school	-0.002^{*}	0.001	-0.002^{*}	0.001	-0.005	0.017
pupil to teacher ratio	0.001	0.001	0.001	0.001	0.010	0.009
share of teachers with masters or higher	-0.126^{*}	0.075	-0.154^{**}	0.072	-0.386	0.339
share of female teachers	0.148^{**}	0.067	0.146^{**}	0.060	0.263	0.600
av. rank on standardized math test	0.000	0.004	0.002	0.003	-0.003	0.008
# of teachers aged ≥ 55 at school	-0.001	0.001	-0.002^{*}	0.001	0.002	0.003
age ≥ 54 years	-0.154^{***}	0.013	-0.179^{***}	0.015		
age ≥ 55 years	-0.123^{***}	0.013	-0.163^{***}	0.015	-0.016	0.029
age ≥ 56 years	-0.140^{***}	0.012	-0.174^{***}	0.014	-0.013	0.011
age ≥ 57 years	-0.138^{***}	0.013	-0.173^{***}	0.014	0.001	0.010
age ≥ 58 years	-0.127^{***}	0.012	-0.163^{***}	0.014	0.008	0.014
age ≥ 59 years	-0.099^{***}	0.014	-0.132^{***}	0.015	0.030^{***}	0.010
age ≥ 60 years	-0.051^{***}	0.015	-0.076^{***}	0.017	0.056^{**}	0.022
age ≥ 61 years	-0.024	0.017	-0.038^{**}	0.019	0.034	0.028
age ≥ 62 years	0.027	0.020	0.023	0.021	0.060^{***}	0.020
age ≥ 63 years	-0.009	0.021	0.001	0.023	-0.022	0.031
age ≥ 64 years	-0.055^{***}	0.021	-0.054^{***}	0.021	-0.052^{*}	0.030
age ≥ 65 years	0.000	0.025	-0.009	0.026	0.037	0.046
age ≥ 66 years	-0.025	0.026	-0.024	0.026	-0.017	0.034
academic year 1999-00	-0.157^{***}	0.027				
academic year 2000-01	-0.080^{***}	0.014	-0.009	0.006	-0.018	0.051
constant	-0.451^{**}	0.219	0.110^{***}	0.017	-0.524	0.489

Notes: ‡**Replication of the results of Brown and Laschever (2012, p. 109; table 4, column 8)**, subject to a marginal modification of the estimation sample due to inconsistent age information, see fn. 30. *** p-value < 0.01; ** p-value < 0.05; * p-value < 0.1. Standard errors clustered at the school level. 21 290 observations, 8 320 teachers, and 586 school clusters for within-transformation estimation. 12 968 observations, 7 088 teachers, and 578 school clusters for first-differences estimation. Since N observations are redundant in the within-transformed model, the number of non-redundant observations for the within-transformed model does not differ from that of the first-differences models. Two further observations are missing in the first-differences estimation due to missing values in 'average rank on standardized math test' for the year 2000. While the within-transformation can still be applied to the corresponding observations for 1999 and 2001, first-differences cannot be calculated, unless one allows for unequally spaced periods. **Source:** Brown and Laschever (2012) and own estimations; variables names are – subject to minor modifications – borrowed from the additional online materials to Brown and Laschever (2012); see `https://www.aeaweb.org/articles?id=10.1257/app.4.3.90`.

from this model to the corresponding ones from alternative estimation methods, more specifically b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$. Because the model specification includes a saturated set of wave dummies, b^{FD} is fully equivalent to b^{FDC}.

The original specification of Brown and Laschever (2012) includes a set of age dummies, including one for the youngest age found in the sample, i.e., 53 years, and an 'older than' dummy for the residual age category. This renders the matrix $\mathbf{G}$ singular for two reasons. Firstly, the 'youngest age' dummy causes the problem discussed in section 2.6. Secondly, the column associated with the 'older than 65' indicator is linearly dependent on the columns associated with the age indicators. The original model specification can, hence, not one-to-one be estimated by the adjusted first-differences estimator. For this reason we excluded the dummy indicating the youngest

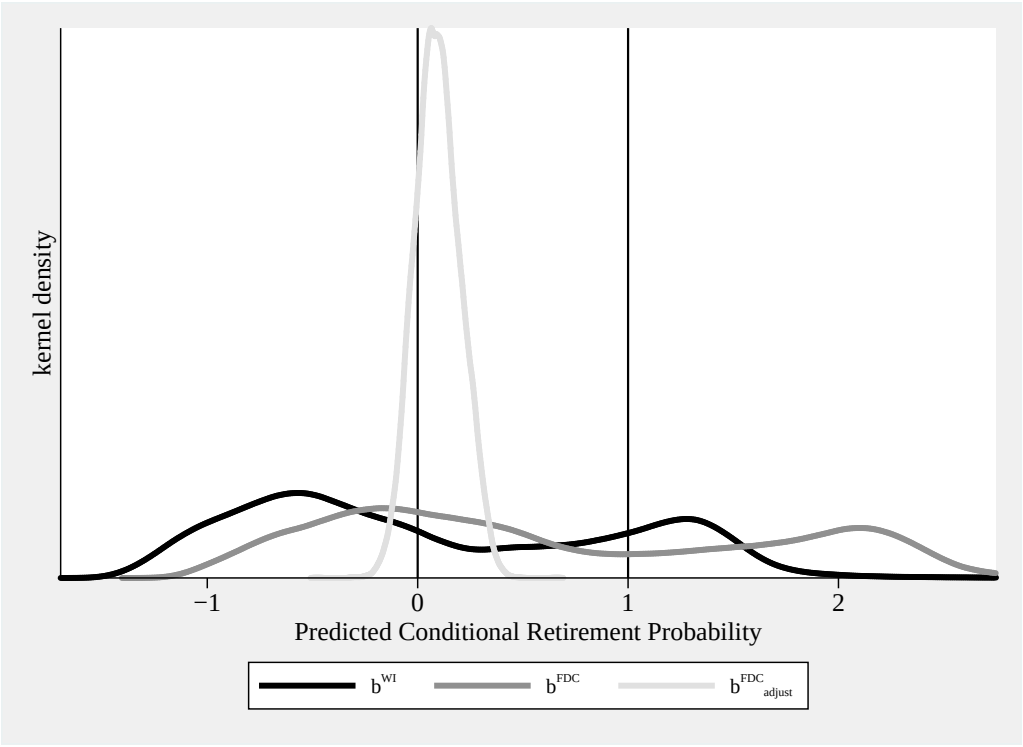

Figure 3: Sample distribution of predicted conditional retirement probabilities from b^{WI}, b^{FDC}, and $b^{\text{FDC}}_{\text{adjust}}$. Predictions from within-transformed estimator based on three waves, i.e. 21 290 obs.; predictions from first-differences estimators based on two waves, i.e. 12 968 obs. The mean outcome (rel. frequency of retirement events) is 0.085 in the three-waves sample and 0.095 in the two-waves sample. **Source:** Own calculations based on Brown and Laschever (2012).

age cohort in the sample from the adjusted first-differences estimation and re-parameterized the age indicators, see footnote 30. Naturally, one wave indicator has to be dropped if the estimation is based on first-differences.

The point estimates for the key coefficients obtained from the unadjusted first-differences estimation are very similar to the original ones. Only 'change in own pension wealth' gets substantially smaller and turns statistically insignificant. Yet, most importantly, a reduced form effect of a change in pension wealth of other teachers is still found in the first-differences estimation. Turning to the results from the adjusted first-differences estimation, this pattern changes. Very few coefficients are statistically significant. In particular, the instrumental variables (change in pension wealth of peers) turn statistically insignificant. The point estimates even turn negative. This clearly conflicts with being interpreted in terms of peer effects mattering for retirement, as suggested by Brown and Laschever (2012). Using $b^{\text{FDC}}_{\text{adjust}}$ instead of b^{WI} as the estimation method, hence, substantially changes the economic implications of the empirical analysis.

In order to shed more light on what is different about these results, we examine predicted conditional retirement probabilities.[31] Figure 3 displays the sample distribution of the fitted values the three considered estimation methods yield in the respective estimation samples.[32] As to be expected when using a linear probability model, all estimators yield some predicted probabilities outside the unit interval. Yet the extent by which this happens varies a great deal. While for b^{WI} and b^{FDC} more than 70 percent of the predictions are outside the valid range, the corresponding share for $b^{\text{FDC}}_{\text{adjust}}$ is smaller than 20 percent. With respect to b^{FDC} one reason for this result is that the predictions are incorrectly centered. In other words, the sample mean of the prediction deviates a great deal from the sample mean of the outcome variable. This does not come as a surprise, since the intercept in a model that is estimated in first differences is not $\bar{y} - \bar{\mathbf{x}}\hat{\beta}$. In contrast, the predictions from b^{WI} and $b^{\text{FDC}}_{\text{adjust}}$ are both precisely centered to the sample mean of the outcome variable. However, the former still does a very poor job in generating reasonable predictions. In fact, very little mass of the distribution of fitted values is located in the meaningful range. On the basis of the predicted probabilities, one would judge $b^{\text{FDC}}_{\text{adjust}}$ as clearly superior to b^{FDC} and b^{WI} in the present application. A possible explanation for the very different estimated distributions of retirement probabilities are the estimated age coefficients, which in absolute terms are typically much bigger for b^{WI} and b^{FDC} than for $b^{\text{FDC}}_{\text{adjust}}$. Most prominent, unlike $b^{\text{FDC}}_{\text{adjust}}$, the unadjusted estimators yield a steady and steep, and statistically significant decrease in the baseline retirement hazard for teachers in their sixth decade of life, which is in no way mirrored by the unconditional sample retirement rates.[33] Indeed, according to the results from applying the within estimator the baseline retirement hazard decreases by 83 percentage points between the age of 53 and the age of 60, which seems to make little sense. A poorly estimated baseline hazard seems to be the main reason for the poor predictions generated by the within-transformation and the simple first-differences estimator. This interpretation is corroborated by simulation results in which the with-transformation estimator yields heavily biased results for the base line hazard; see Table A2 in Appendix A.4.

[31] The predictions are calculated as $\left(\hat{\alpha}^{\text{WI}} + \mathbf{x}_{it}\hat{\beta}^{\text{WI}}\right)$, $\left(\hat{\alpha}^{\text{FDC}} + \mathbf{x}_{it}\hat{\beta}^{\text{FDC}}\right)$, and $\left(\hat{\alpha}^{\text{FDC}}_{\text{adjust}} + \mathbf{x}_{it}\hat{\beta}^{\text{FDC}}_{\text{adjust}}\right)$, respectively. They are thus unconditional on a_i.

[32] For first-differences estimators the first wave cannot be considered since only one academic year dummy is identified.

[33] The descriptive counterparts, i.e., the changes in relative retirement rates between age 54 and age 66 is: 0.002, 0.028, -0.007, 0.004, 0.018, 0.027, 0.062, 0.037, 0.026, -0.008, -0.021, 0.013, and 0.009. The strange pattern of estimated age coefficients is not an artifact of including the 'age $\geq$ 54' indicator of the specification estimated with b^{WI} and b^{FDC}. Though less pronounced, the general pattern of estimated age coefficients – and the strange distribution of the predicted probabilities – survives if 'age $\geq$ 54' is dropped.

5 Conclusions

Eliminating individual time-invariant heterogeneity by taking first-differences or applying the within-transformation to the data is a powerful tool of applied econometrics that makes the linear regression model very appealing in the analysis of panel data. However, the logic that these transformations remove individual time-invariant heterogeneity and therefore allow for consistent and unbiased estimation by least squares does not apply in a discrete-time hazard setting, in which an observation unit is only observed until that period in which the event of interest occurs. As shown above, conventional fixed-effects estimators are in fact biased and inconsistent in this case. Besides conventional survival bias, which would also affect pooled OLS even if the individual heterogeneity is uncorrelated with the explanatory variables in the population, these estimators suffer from a second source of bias that originates from the data transformation itself. It is therefore present even in the absence of any unobserved heterogeneity. This second source of bias turns out to be the dominant one in many settings, with its magnitude heavily depending on the data generating process of the explanatory variables. The conventional first-differences and the within-transformation estimators should for this reason not be applied to discrete-time hazard models.

In this paper, we suggest a novel alternative adjusted first-differences estimator for this setting that is computationally very simple and cures this second source of bias. Under the assumption that any unobserved time-invariant, individual heterogeneity is uncorrelated with the first – or alternatively higher-order – differences of the explanatory variables, it confines the bias to survival bias. It thus allows confining the bias to survival bias under alternative, supposedly weaker assumptions than pooled OLS, for which uncorrelatedness with the levels of the explanatory variables is required. Compared to conventional linear fixed-effects estimators, its crucial advantage is that it corrects for the misleading matrix-weighting of coefficients that originates from the data transformation these estimators involve. The contribution of this paper hence is twofold. Firstly, it shows why conventional linear fixed-effects estimators should not be used in a discrete-time hazard framework. Secondly, it introduces an alternative estimator that confines a possible bias to a single source. This remaining source is just a variant of conventional survival bias researchers should always be aware of when estimating a linear discrete-time hazard model.

References

Alejo, J., Bera, A., Galvao, A., Montes-Rojas, G. and Xiao, Z. (2016). Tests for normality based on the quantile-mean covariance, *Stata Journal* **16**(4): 1039–1057.

Allison, P. D. (1994). Using panel data to estimate the effects of events, *Sociological Methods & Research* **23**(2): 174–199.

Allison, P. D. (2009). *Fixed Effects Regression Models*, SAGE Publications.

Allison, P. D. and Christakis, N. A. (2006). Fixed-effects methods for the analysis of nonrepeated events, *Sociological Methodology* **36**(1): 155–172.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity, *Journal of the American Statistical Association* **90**(430): 431–442.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, 1 edn, Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of Economic Perspectives* **24**(2): 3–30.

Baum, C. F. and Cox, N. J. (2001). OMNINORM: Stata module to calculate omnibus test for univariate/multivariate normality, Statistical Software Components, Boston College Department of Economics. Revised 08 Apr 2009.

Bera, A., Galvao, A., Wang, L. and Xiao, Z. (2016). A new characterization of the normal distribution and test for normality, *Econometric Theory* **32**(5): 1216–1252.

Bogart, D. (2018). Party connections, interest groups and the slow diffusion of infrastructure: Evidence from Britain's first transport revolution, *The Economic Journal* **128**(609): 541–575.

Brown, K. M. and Laschever, R. A. (2012). When they're sixty-four: Peer effects and the timing of retirement, *American Economic Journal: Applied Economics* **4**(3): 90–115.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics*, Cambridge University Press, Cambridge.

Cantoni, D. (2012). Adopting a new religion: the case of protestantism in 16th century Germany, *The Economic Journal* **122**(560): 502–531.

Ciccone, A. (2011). Economic shocks and civil conflict: A comment, *American Economic Journal: Applied Economics* **3**(4): 215–227.

D'Agostino, R. B., Belanger, A. and D'Agostino Jr., R. B. (1990). A suggestion for using powerful and informative tests of normality, *The American Statistician* **44**(4): 316–321.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality, *Oxford Bulletin of Economics and Statistics* **70**(s1): 927–939.

Fernandes, A. M. and Paunov, C. (2015). The risks of innovation: Are innovating firms less likely to die?, *The Review of Economics and Statistics* **97**(3): 638–653.

Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects, *The Econometrics Journal* **7**(1): 98–119.

Greene, W. (2014). *Econometric Analysis*, Pearson Series in Economics, Pearson Education Limited.

Harding, R. and Stasavage, D. (2014). What democracy does (and doesn't do) for basic services: School fees, school inputs, and African elections, *The Journal of Politics* **76**(1): 229–245.

Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity, *Econometrica* **67**(5): 1001–1028.

Horowitz, J. L. and Lee, S. (2004). Semiparametric estimation of a panel data proportional hazards model with fixed effects, *Journal of Econometrics* **119**(1): 155 – 198.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model, *Economics Letters* **90**: 321–327.

Jacobson, T. and von Schedvin, E. (2015). Trade credit and the propagation of corporate failure: An empirical analysis, *Econometrica* **83**(4): 1315–1371.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models, *Oxford Bulletin of Economics and Statistics* **57**(1): 129–136.

Miguel, E., Satyanath, S. and Sergenti, E. (2004). Economic shocks and civil conflict: An instrumental variables approach, *Journal of Political Economy* **112**(4): 725–753.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics* **34**(1): 57–67.

Stammann, A., Heiß, F. and McFadden, D. (2016). Estimating fixed effects logit models with large panel data, number G01-V3 in *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016:*

Demographischer Wandel - Session: Microeconometrics, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft.

Wang, S., Greenwood, B. and Pavlou, P. A. (2017). Tempting fate: Social media posts by firms, customer purchases, and the loss of followers, *Research Paper 17-022*, Fox School of Business.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge Massachusetts.

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach*, South-Western.

A Appendix

A.1 Within-Transformation of Non-Repeated Event Data

For the within-transformation estimator the analogue to (7) reads as

$$\begin{aligned}
\mathrm{E}\left(\varepsilon_{it}^{\mathrm{WI}}|a_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{y}_{it^-}=\mathbf{0}\right) &= \mathrm{P}(y_{it}=1|a_i,\mathbf{x}_{it},\mathbf{y}_{it^-}=\mathbf{0})\left(\frac{t-1}{t}-\left(\mathbf{x}_{it}-\frac{1}{t}\sum_{s=1}^{t}\mathbf{x}_{is}\right)\beta\right)\\
&+\sum_{T_i=t+1}^{T}\left[\mathrm{P}(y_{iT_i}=1|a_i,\mathbf{x}_{iT_i},\mathbf{y}_{iT_i-}=\mathbf{0})\left(\prod_{s=t}^{T_i-1}\mathrm{P}(y_{is}=0|a_i,\mathbf{x}_{is},\mathbf{y}_{is-}=\mathbf{0})\right)\right.\\
&\left.\times\left(-\frac{1}{T_i}-\left(\mathbf{x}_{it}-\frac{1}{T_i}\sum_{s=1}^{T_i}\mathbf{x}_{is}\right)\beta\right)\right]\\
&+\left(\prod_{s=t}^{T}\mathrm{P}(y_{is}=0|a_i,\mathbf{x}_{is},\mathbf{y}_{is-}=\mathbf{0})\right)\left(-\left(\mathbf{x}_{it}-\frac{1}{T_i}\sum_{s=1}^{T}\mathbf{x}_{is}\right)\beta\right)\\
&=(a_i+\mathbf{x}_{it}\beta)\left(\frac{t-1}{t}-\left(\mathbf{x}_{it}-\frac{1}{t}\sum_{s=1}^{t}\mathbf{x}_{is}\right)\beta\right)\\
&+\sum_{T_i=t+1}^{T}\left[(a_i+\mathbf{x}_{iT_i}\beta)\left(\prod_{s=t}^{T_i-1}(1-a_i-\mathbf{x}_{is}\beta)\right)\left(-\frac{1}{T_i}-\left(\mathbf{x}_{it}-\frac{1}{T_i}\sum_{s=1}^{T_i}\mathbf{x}_{is}\right)\beta\right)\right]\\
&+\left(\prod_{s=t}^{T}(1-a_i-\mathbf{x}_{is}\beta)\right)\left(-\left(\mathbf{x}_{it}-\frac{1}{T_i}\sum_{s=1}^{T}\mathbf{x}_{is}\right)\beta\right) \qquad (24)
\end{aligned}$$

For $t=T$, that is, when the event does not occur in the waves $1,\ldots,T-1$ and in consequence T_i equals T, (24) simplifies to

$$\mathrm{E}\left(\varepsilon_{iT}^{\mathrm{WI}}|a_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},\mathbf{y}_{iT^-}=\mathbf{0}\right)=\left(\frac{T-1}{T}\right)a_i+\frac{1}{T}\left(\sum_{s=1}^{T-1}\mathbf{x}_{is}\right)\beta \qquad (25)$$

If the panel is very short and consists of only two waves, i.e., $T=2$, we get

$$\mathrm{E}\left(\varepsilon_{i2}^{\mathrm{WI}}|a_i,\mathbf{x}_{i1},\mathbf{x}_{i2},y_{i1}=0\right)=\frac{1}{2}a_i+\frac{1}{2}\mathbf{x}_{i1}\beta \qquad (26)$$

This coincides with (7) since for $T=2$, the within-transformed data are just $\frac{1}{2}$ times the first-differenced data for $t=2$ and $-\frac{1}{2}$ times the first-differenced data for $t=1$.

A.2 First-Differences Transformation of Repeated Event Data

When y_{it} is a repeated event, that is any sequence of zeros and ones can be observed, the first-differenced outcome $\Delta y_{it} \equiv y_{it} - y_{it-1}$ is

$$\Delta y_{it} = \begin{cases} 1 & \text{if} \quad y_{it} = 1 \text{ and } y_{it-1} = 0 \\ 0 & \text{if} \quad (y_{it} = 1 \text{ and } y_{it-1} = 1) \text{ or } (y_{it} = 0 \text{ and } y_{it-1} = 0) \\ -1 & \text{if} \quad y_{it} = 0 \text{ and } y_{it-1} = 1 \end{cases} \tag{27}$$

and the corresponding first-differenced disturbance $\varepsilon_{it}^{\text{FDR}} \equiv \Delta y_{it} - \Delta \mathbf{x}_{it}\beta$ reads as

$$\varepsilon_{it}^{\text{FDR}} = \begin{cases} 1 - \Delta \mathbf{x}_{it}\beta & \text{if} \quad y_{it} = 1 \text{ and } y_{it-1} = 0 \\ -\Delta \mathbf{x}_{it}\beta & \text{if} \quad (y_{it} = 1 \text{ and } y_{it-1} = 1) \text{ or } (y_{it} = 0 \text{ and } y_{it-1} = 0) \\ -1 - \Delta \mathbf{x}_{it}\beta & \text{if} \quad y_{it} = 0 \text{ and } y_{it-1} = 1 \end{cases} \tag{28}$$

For the conditional mean of the disturbance in the first-differenced linear probability model with repeated events one obtains

$$\begin{aligned} \mathrm{E}(\varepsilon_{it}^{\text{FDR}} | a_i, \mathbf{x}_{it}, \mathbf{x}_{it-1}) &= [\mathrm{P}(y_{it} = 1 | a_i, \mathbf{x}_{it}) \cdot \mathrm{P}(y_{it-1} = 0 | a_i, \mathbf{x}_{it-1})] \, (1 - \Delta \mathbf{x}_{it}\beta) \\ &\quad + [\mathrm{P}(y_{it} = 1 | a_i, \mathbf{x}_{it}) \cdot \mathrm{P}(y_{it-1} = 1 | a_i, \mathbf{x}_{it-1}) \\ &\quad + \mathrm{P}(y_{it} = 0 | a_i, \mathbf{x}_{it}) \cdot \mathrm{P}(y_{it-1} = 0 | a_i, \mathbf{x}_{it-1})] \, (-\Delta \mathbf{x}_{it}\beta) \\ &\quad + [\mathrm{P}(y_{it} = 0 | a_i, \mathbf{x}_{it}) \cdot \mathrm{P}(y_{it-1} = 1 | a_i, \mathbf{x}_{it-1})] \, (-1 - \Delta \mathbf{x}_{it}\beta) \\ &= [(a_i + \mathbf{x}_{it}\beta) \cdot (1 - a_i - \mathbf{x}_{it-1}\beta)] \, (1 - \Delta \mathbf{x}_{it}\beta) \\ &\quad + [(a_i + \mathbf{x}_{it}\beta) \cdot (a_i + \mathbf{x}_{it-1}\beta) + (1 - a_i - \mathbf{x}_{it}\beta) \cdot (1 - a_i - \mathbf{x}_{it-1}\beta)] \, (-\Delta \mathbf{x}_{it}\beta) \\ &\quad + [(1 - a_i - \mathbf{x}_{it}\beta) \cdot (a_i + \mathbf{x}_{it-1}\beta)] \, (-1 - \Delta \mathbf{x}_{it}\beta) \\ &= [(a_i + \mathbf{x}_{it}\beta) \cdot (a_i + \mathbf{x}_{it-1}\beta) + (1 - a_i - \mathbf{x}_{it}\beta) \cdot (1 - a_i - \mathbf{x}_{it-1}\beta) \\ &\quad + (a_i + \mathbf{x}_{it}\beta) \cdot (1 - a_i - \mathbf{x}_{it-1}\beta) + (1 - a_i - \mathbf{x}_{it}\beta) \cdot (a_i + \mathbf{x}_{it-1}\beta)] \, (-\Delta \mathbf{x}_{it}\beta) \\ &\quad + (a_i + \mathbf{x}_{it}\beta) \cdot (1 - a_i - \mathbf{x}_{it-1}\beta) - (1 - a_i - \mathbf{x}_{it}\beta) \cdot (a_i + \mathbf{x}_{it-1}\beta) \\ &= -\Delta \mathbf{x}_{it}\beta + \mathbf{x}_{it}\beta - \mathbf{x}_{it-1}\beta \\ &= 0 \end{aligned} \tag{29}$$

That is, for repeated events, applying the first-differences transformation eliminates unobserved, time-invariant heterogeneity and yields a transformed disturbance that is conditional mean independent of the explanatory variables, allowing unbiased and consistent estimation by least squares.

A.3 **Simulation Results for Probit Model as True DGP**

Table A1: Monte Carlo Analysis - **Probit as true DGP** (Large Sample Estimates)

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x_{it}^{ST} **stationary**: average marginal effect **0.3373** (fist wave included), and **0.3353** (fist wave excluded)										
$\hat{\beta}$	0.5667	0.0004	0.3038	0.0008	0.2383	0.0007	0.1694	0.0006	0.3363	0.0012
$\hat{\beta}$/av. marg. effect	1.6804	0.0012	0.9007	0.0025	0.7108	0.0022	0.5053	0.0019	1.0028	0.0037
x_{it}^{RW} **follows random walk**: average marginal effect **0.3359** (fist wave included), and **0.3330** (fist wave excluded)										
$\hat{\beta}$	0.4783	0.0003	0.3117	0.0006	0.3337	0.0007	0.3334	0.0006	0.3333	0.0006
$\hat{\beta}$/av. marg. effect	1.4238	0.0009	0.9280	0.0019	1.0022	0.0022	1.0011	0.0018	1.0010	0.0018
x_{it}^{TR} **with trend and incr. var. around trend**: av. marg. effect **0.3395** (fist wave incl.), and **0.3424** (fist wave excl.)										
$\hat{\beta}$	0.5389	0.0004	2.0046	0.0006	1.4954	0.0007	0.2331	0.0006	0.3484	0.0009
$\hat{\beta}$/av. marg. effect	1.5873	0.0011	5.9040	0.0018	4.3669	0.0019	0.6806	0.0018	1.0175	0.0027

Notes: True DGP: $\mathrm{P}(y_{it} = 1|a_i, x_{it}, \mathbf{y}_{it^-} = \mathbf{0}) = \Phi(-1.44 + 3(a_i + x_{it}\beta))$; true coefficient value: $\beta = 1$; $N = 4 \cdot 10^7$, $T = 5$; same DGPs for x_{it} as in the simulations discussed in section 3.1; the # of observations for x_{it}^{ST} is 72 281 765, the corresponding # of observations for x_{it}^{RW} is 72 311 334, and for x_{it}^{TR} it is 72 775 017. For b^{OLS} the #s of observations are higher by $4 \cdot 10^7$ observations, since the first wave is not eliminated by the within-transformation or the first-differences transformation. See Table 1 for corresponding simulation results assuming a DGP consistent with the linear model.

Table A1 shows results from simulations in which the linear estimators are applied to data that was generated by the process $\mathrm{P}\left(y_{it} = 1|a_i, x_{it}, \mathbf{y}_{it^-} = \mathbf{0}\right) = \Phi\left(-1.44 + 3(a_i + x_{it}\beta)\right)$, with $\beta = 1$, $\mathrm{E}(a_i) = \alpha = 0.1$, and Φ denoting the CDF of the standard normal distribution. The explanatory variable x_{it} and the unobserved heterogeneity a_i are generated by the same DGPs as considered in section 3.1. The scaling factor 3 and the constant -1.44 are introduced to generate probabilities that exhibit (almost) the same sample mean and same sample variance as the corresponding linear probabilities considered in section 3.1. Though the true slope coefficient β is still 1, in the considered probit model the quantity of interest is not β but the corresponding average of the marginal effect $\frac{\partial \mathrm{P}(y_{it}=1|a_i,x_{it},\mathbf{y}_{it^-}=\mathbf{0})}{\partial 3 x_{it}}$ that is $\frac{1}{\sum_{i=1}^{N} T_i} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \beta\phi\left(-1.44 + 3(a_i + x_{it}\beta)\right)$. Its true value is roughly $1/3$ for all considered DGPs for x_{it}. The entries in Table A1 are (i) the estimated slope coefficients – as raw estimates of the average marginal effect – and (ii) the estimated slope coefficient relative to the true mean marginal effect. The latter can directly be compared to the estimated slope coefficients in Table 1. From this comparison it becomes obvious that the pattern of biases is the same for the true DGP being linear or being of probit-type. This findig is in line with the literature (e.g. Wooldridge, 2002, 455) that states that in term of average partial effects the linear probability model does very good job in approximating the results from non-linear binary response models. In consequence, the above simulation results indicate that the advantage of the adjusted estimator over the unadjusted conventional ones carries over to settings in which the true DGP is not consistent with the linear hazard model.

A.4 **Simulation Results for Specification with Wave Indicators**

Table A2: Monte Carlo Analysis - Large Sample Estimates, **Wave Indicators** included

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}\ddagger}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
$x_{it}^{\boldsymbol{ST}}$ **stationary**										
$\hat{\beta}$	1.6648	0.0012	0.6396	0.0021	–	–	0.5008	0.0019	0.9980	0.0037
$\hat{\tau}_2$	−0.0008	0.0001	0.2947	0.0001	–	–				
$\hat{\tau}_3$	−0.0005	0.0001	0.1430	0.0001	–	–	−0.0044	0.0001	−0.0020	0.0001
$\hat{\tau}_4$	−0.0010	0.0002	0.0920	0.0001	–	–	−0.0092	0.0001	−0.0025	0.0002
$\hat{\tau}_5$	−0.0007	0.0002	0.0670	0.0002	–	–	−0.0136	0.0002	−0.0022	0.0002
$\hat{\alpha}$	−0.0331	0.0002	−0.1032	0.0004	–	–	0.2948	0.0001	0.0979	0.0007
$x_{it}^{\boldsymbol{RW}}$ **follows random walk**										
$\hat{\beta}$	1.4245	0.0010	1.2059	0.0016	–	–	1.0000	0.0018	0.9999	0.0018
$\hat{\tau}_2$	−0.0015	0.0001	0.2951	0.0001	–	–				
$\hat{\tau}_3$	−0.0007	0.0001	0.1426	0.0001	–	–	−0.0057	0.0001	−0.0022	0.0001
$\hat{\tau}_4$	−0.0004	0.0002	0.0909	0.0001	–	–	−0.0127	0.0001	−0.0023	0.0002
$\hat{\tau}_5$	0.0003	0.0002	0.0649	0.0002	–	–	−0.0205	0.0002	−0.0022	0.0002
$\hat{\alpha}$	0.0151	0.0002	−0.2143	0.0003	–	–	0.2951	0.0001	0.0975	0.0004
$x_{it}^{\boldsymbol{TR}}$ **with trend and increasing variance around trend**										
$\hat{\beta}$	1.6896	0.0012	0.9041	0.0022	–	–	0.6689	0.0019	1.0092	0.0028
$\hat{\tau}_2$	−0.0094	0.0001	0.2839	0.0001	–	–				
$\hat{\tau}_3$	−0.0091	0.0001	0.1443	0.0001	–	–	0.0079	0.0001	−0.0022	0.0001
$\hat{\tau}_4$	−0.0096	0.0002	0.0972	0.0001	–	–	0.0153	0.0002	−0.0026	0.0002
$\hat{\tau}_5$	−0.0094	0.0002	0.0740	0.0002	–	–	0.0227	0.0002	−0.0025	0.0002
$\hat{\alpha}$	−0.0293	0.0002	−0.1496	0.0004	–	–	0.2869	0.0001	0.0958	0.0006

Notes: τ_t denote coefficients of dummies indicating waves $\geq t$. True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$, $\boldsymbol{\tau_2 = \ldots = \tau_5 = 0}$; ‡ $b^{\mathbf{FD}}$ not considered since including a saturated set of waves indicators makes $b^{\mathbf{FD}}$ and $b^{\mathbf{FDC}}$ coincide; $N = 4 \cdot 10^7$, $T = 5$; the # of observations for x_{it}^{ST} is 71 748 906, the corresponding # of observations for x_{it}^{RW} is 71 823 746, and for x_{it}^{TR} it is 72 218 321. For b^{OLS} the #s of observations are higher by $4 \cdot 10^7$ observations, since the first wave is not eliminated by the within-transformation or the first-differences transformation. See Table 1 for corresponding simulation results based on specification without wave indicators.

Table A.4 displays large sample simulation results for as specification fully equivalent to the one for which results are displayed in Table 1, except for including a saturated set of time indicators. These additional dummies indicate that an observation is from waves t or a later wave. The attached true coefficients, hence, capture how the baseline hazard changes from $t-1$ to t. To isolate the effect including the time indicators has on the results, we use exactly the same simulated data that is used for generating the results shown in Table 1. This means that the true DGP does not involve time effects but exhibits a constant baseline hazard. While including these dummies has almost no effect on $\hat{\beta}$ one gets from b^{OLS}, b^{FDC}, and $b^{\text{FDC}}_{\text{adjust}}$, the within-transformation estimator b^{WI} turns out to be quite sensitive to this change of the model specification. While the extreme upward bias for an x_{it} with trend disappears and is replaced by an moderate downward bias, the upward bias for a stationary x_{it} gets more pronounced. For x_{it} following a random walk, instead of suffering from a small downward bias, b^{WI} exhibits a sizable upward bias, if time indicators are included. Moreover, b^{WI} yields estimated time effects on the baseline hazard that are completely misleading. This mirrors the counterintuitive age effects b^{WI} yields in the real data application; see section 5. The simulation results are inline with our earlier argument about $b^{\text{FDC}}_{\text{adjust}}$ being biased

Table A3: Monte Carlo Analysis - Large Samp. Est., true **Time Effects** and **Wave Indicators**

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}\ddagger}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
$x_{it}^{\boldsymbol{ST}}$ **stationary**										
$\hat{\beta}$	1.6670	0.0012	0.6394	0.0021	–	–	0.5019	0.0018	1.0001	0.0037
$\hat{\tau}_2$	−0.1008	0.0001	0.1947	0.0001	–	–				
$\hat{\tau}_3$	0.1997	0.0001	0.2937	0.0001	–	–	0.1963	0.0001	0.1983	0.0001
$\hat{\tau}_4$	−0.0512	0.0002	0.0746	0.0001	–	–	0.1405	0.0001	−0.0530	0.0002
$\hat{\tau}_5$	−0.0007	0.0002	0.0791	0.0002	–	–	0.1359	0.0002	−0.0023	0.0002
$\hat{\alpha}$	−0.0335	0.0002	−0.0918	0.0004	–	–	0.1948	0.0001	−0.0025	0.0007
$x_{it}^{\boldsymbol{RW}}$ **follows random walk**										
$\hat{\beta}$	1.4261	0.0010	1.2029	0.0017	–	–	1.0022	0.0018	1.0022	0.0018
$\hat{\tau}_2$	−0.1015	0.0001	0.1951	0.0001	–	–				
$\hat{\tau}_3$	0.1994	0.0001	0.2932	0.0001	–	–	0.1950	0.0001	0.1981	0.0001
$\hat{\tau}_4$	−0.0504	0.0002	0.0736	0.0001	–	–	0.1370	0.0001	−0.0526	0.0002
$\hat{\tau}_5$	−0.0000	0.0002	0.0765	0.0002	–	–	0.1283	0.0002	−0.0026	0.0002
$\hat{\alpha}$	0.0148	0.0002	−0.2024	0.0003	–	–	0.1951	0.0001	−0.0029	0.0004
$x_{it}^{\boldsymbol{TR}}$ **with trend and increasing variance around trend**										
$\hat{\beta}$	1.6912	0.0012	0.8988	0.0023	–	–	0.6687	0.0019	1.0046	0.0029
$\hat{\tau}_2$	−0.1094	0.0001	0.1840	0.0001	–	–				
$\hat{\tau}_3$	0.1909	0.0001	0.2949	0.0001	–	–	0.2085	0.0001	0.1981	0.0001
$\hat{\tau}_4$	−0.0596	0.0002	0.0800	0.0001	–	–	0.1650	0.0001	−0.0529	0.0002
$\hat{\tau}_5$	−0.0093	0.0002	0.0862	0.0002	–	–	0.1723	0.0002	−0.0025	0.0002
$\hat{\alpha}$	−0.0296	0.0002	−0.1369	0.0004	–	–	0.1868	0.0001	−0.0033	0.0006

Notes: τ_t denote coefficients of dummies indicating waves $\geq t$. True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$, $\boldsymbol{\tau_2 = -0.1}$, $\boldsymbol{\tau_3 = 0.2}$, $\boldsymbol{\tau_4 = -0.05}$, $\boldsymbol{\tau_5 = 0}$; ‡ $b^{\mathbf{FD}}$ not considered since including a saturated set of waves indicators makes $b^{\mathbf{FD}}$ and $b^{\mathbf{FDC}}$ coincide; $N = 4 \cdot 10^7$, $T = 5$; the # of observations for x_{it}^{ST} is 73 382 281, the corresponding # of observations for x_{it}^{RW} is 73 457 235, and for x_{it}^{TR} it is 73 847 642. For b^{OLS} the #s of observations are higher by $4 \cdot 10^7$ observations, since the first wave is not eliminated by the within-transformation or the first-differences transformation.

with regard to the baseline hazard that is α, and $\tau_2 \ldots \tau_5$. According to the estimates of $\tau_2 \ldots \tau_5$ the baseline hazard decreases over time, though the data generating process does not involve such time dependence. This is explained by the fact that $\hat{\tau}_t$ captures the decrease of $\mathrm{E}(a_i|t, \mathbf{X})$ due to selective survival.

Table A3 shows simulation result for the same model specification used to generate the results displayed in Table A2. Yet unlike the latter, here the true DGP involves time effects, i.e. the true baseline hazard is not flat. More precisely the jumps in the true baseline hazard are: $\tau_2 = -0.1$, $\tau_3 = 0.2$, $\tau_4 = -0.05$, and $\tau_5 = 0$. In qualitative terms, the results mirror what is found for a flat baseline hazard. As befor, $b^{\text{FDC}}_{\text{adjust}}$ does not estimate the baseline hazard unbiasedly. Yet, the error in the estimated baseline hazard turns out to be rather small. b^{WI} still yields poor results both in terms of the baseline hazard and in terms of the $\hat{\beta}$.

A.5 **Simulation Results for Alternative Beta Distributions**

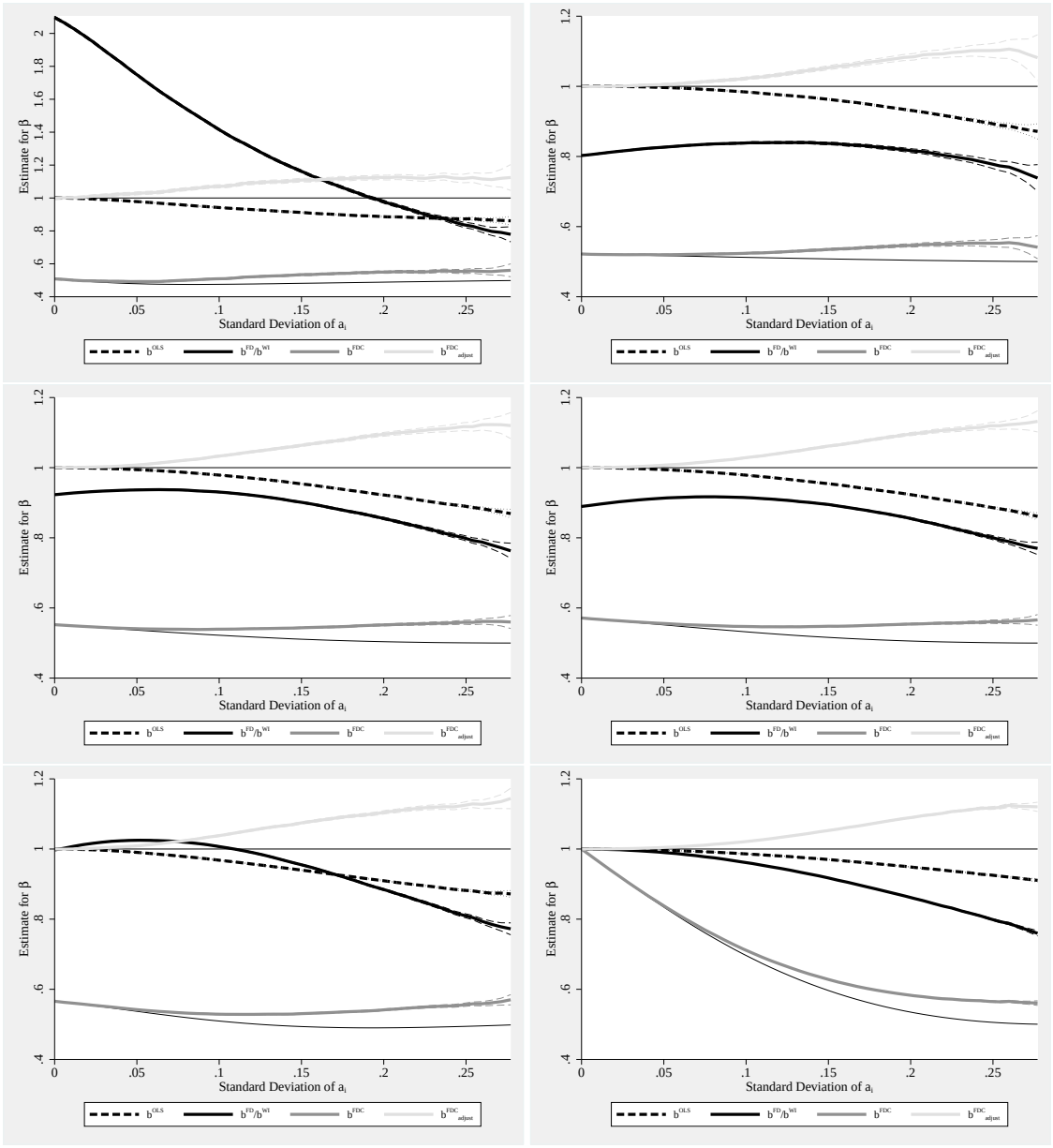

Figure A1: Estimated β coefficients as functions of $\sqrt{\text{Var}(a_i)} = q/\sqrt{12}$. DGPs of a_i and x_{it}: a_i sampled from the **continuous uniform** $U(0,q)$ distribution; $x_{it}^{STB} = \left((1-q)/2\right)\mu_i + \left((1-q)/2\right)\eta_{it}$ with μ_i and η_{it} sampled from **beta** distributions, specifically line-by-line: $\boldsymbol{B(5,1)}$, $\boldsymbol{B(3,5)}$, $\boldsymbol{B(1,1)}$, $\boldsymbol{B(0.5,0.5)}$, $\boldsymbol{B(0.4,0.2)}$, and $\boldsymbol{B(o,o)}$, with $o \to 0$. The latter (bottom row, right) is a rather special case for which the beta distribution coincides with the **Bernoulli** $b(0.5)$ distribution. To emphasize the idea of x_{it} being Bernoulli distributed, in this case we adjust the DGP for x_{it} as follows: $x_{it}^{STB} = (1-q)\rho_{it}$ with $\rho_{it} \sim b(\psi_i)$ and $\psi_i \sim U(0,1)$. q varied in the range between 0 and 0.96. Dashed subsidiary lines mark 95 percent confidence intervals. Thin solid subsidiary lines indicate the true coefficient value $\beta = 1$ and the β-element of $\mathbf{G}\tilde{\beta}$, respectively. See section 3.3 for a detailed description of the Monte Carlo experiment. **Source:** Own simulations.

A.6 **Simulation Results for Bernoulli distributed a_i**

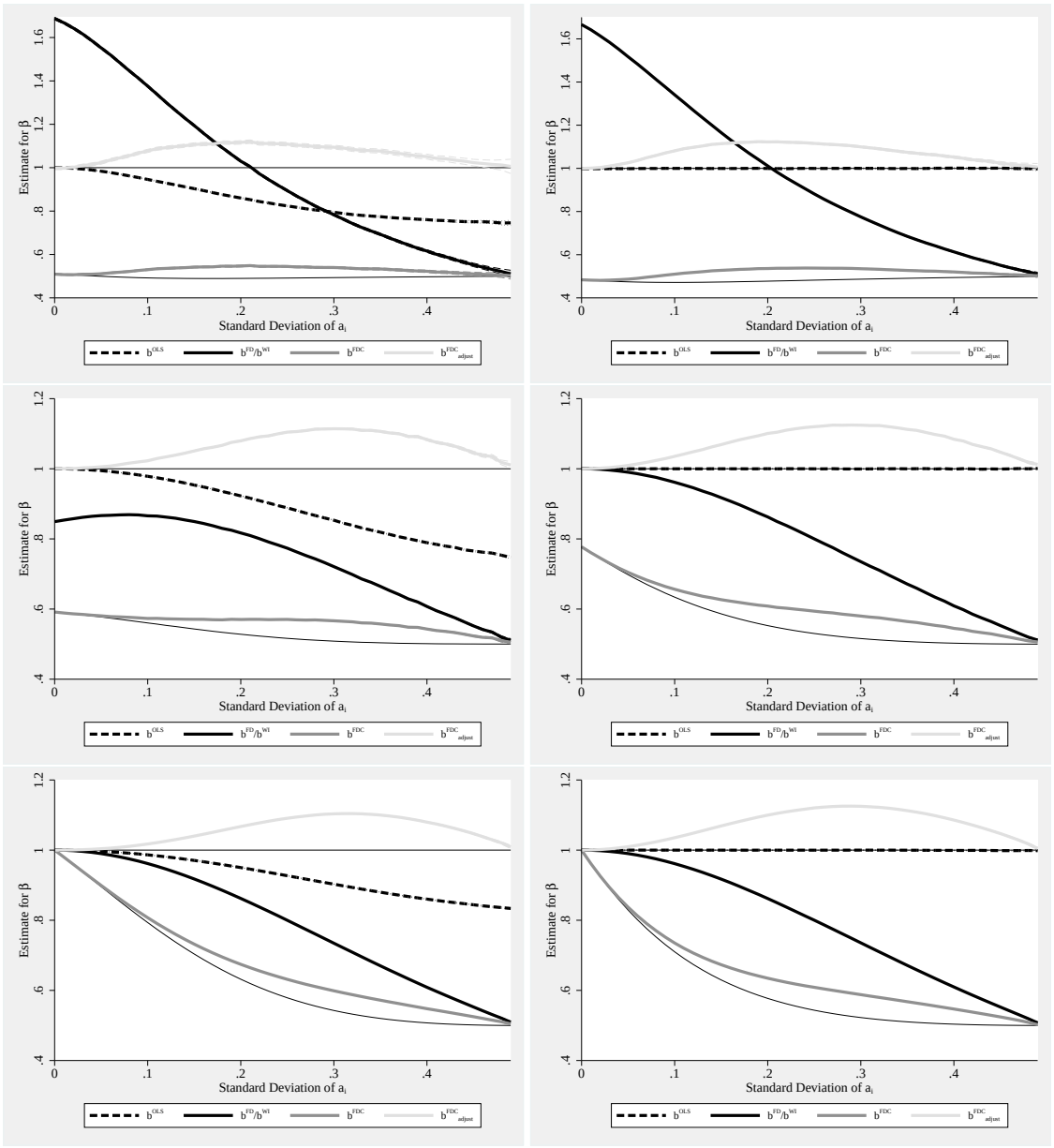

Figure A2: Estimated β coefficients as functions of $\sqrt{\text{Var}(a_i)} = q/2$. DGPs of a_i and x_{it}: $a_i = q\iota_i$, with ι_i sampled from the **Bernoulli** $b(0.5)$ distribution; $x_{it}^{STB} = \left((1-q)/2\right)\mu_i + \left((1-q)/2\right)\eta_{it}$ (left column) and $x_{it}^{STT} = (1-q)\eta_{it}$ (right column), with μ_i and η_{it} sampled from **beta** distributions, specifically: $\boldsymbol{B(6,2)}$ (first row) and $\boldsymbol{B(0.2,0.2)}$ (second row), in the bottom row we use the **Bernoulli** $b(\psi)$ distribution, instead of the beta, to generate x_{it}, specifically: $x_{it}^{STB} = (1-q)\rho_{it}$ with $\rho_{it} \sim b(\psi_i)$ and $\psi_i \sim U(0,1)$ (left), and $x_{it}^{STT} = (1-q)\varrho_{it}$ with $\varrho_{it} \sim b(0.5)$ (right). q varied in the range between 0 and 0.98. Dashed subsidiary lines mark 95 percent confidence intervals. Thin solid subsidiary lines indicate the true coefficient value $\beta = 1$ and the β-element of $\mathbf{G}\tilde{\beta}$, respectively. See section 3.3 for a detailed description of the Monte Carlo experiment. **Source:** Own simulations.

A.7 Simulated Distribution of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ (small sample)

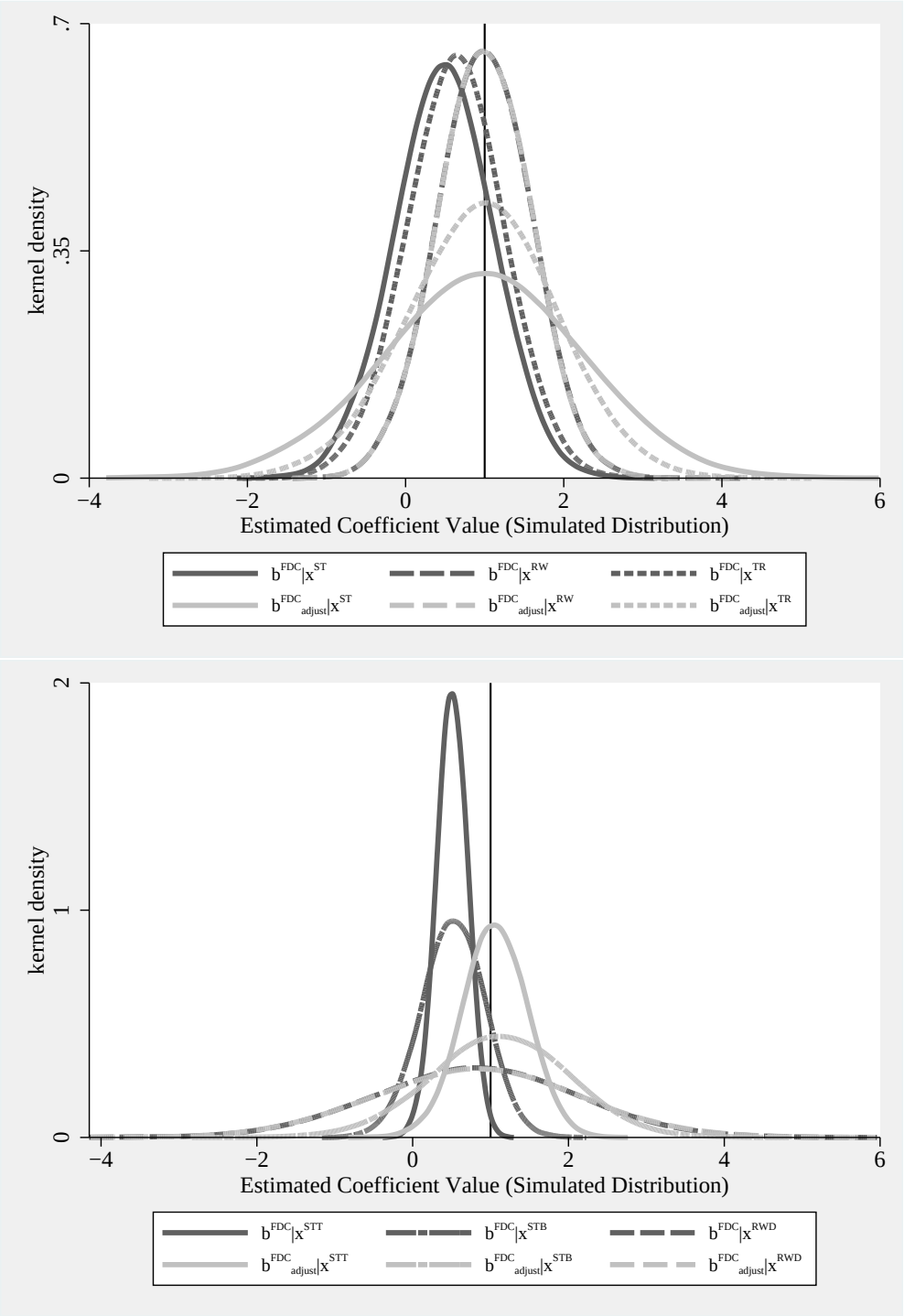

Figure A3: Monte Carlo simulated distribution of b^{FDC} and $b^{\text{FDC}}_{\text{adjust}}$ for different DGPs of x_{it} based on 10 000 replications; **upper panel**: same simulation design (**no** significant **survival bias**) as for the results in Table 2 (lower panel, right-most columns) and Table 4 (left-most column), small sample $N = \mathbf{4 \cdot 10^2}$; **lower panel**: same simulation design (**significant survival bias**) as for the results in Table 5 (left-most column), small sample $N = \mathbf{10^3}$. The thin vertical subsidiary lines mark the true coefficient value 1. **Source:** Own simulations.

A.8 **Simulation Results for Alternative RNG Seeds**

Table A4: Monte Carlo Analysis - Large Sample Estimates (alternative RNG seed)

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x_{it}^{ST} **stationary**										
$\hat{\beta}$	1.6712	0.0012	0.9041	0.0025	0.7113	0.0022	0.5048	0.0019	1.0060	0.0037
$\hat{\alpha}$	−0.0353	0.0002	0.1157	0.0005			0.2899	0.0001	0.0939	0.0007
x_{it}^{RW} **follows random walk**										
$\hat{\beta}$	1.4262	0.0009	0.9445	0.0019	0.9999	0.0022	1.0003	0.0018	1.0005	0.0018
$\hat{\alpha}$	0.0135	0.0002	0.1077	0.0004			0.2882	0.0001	0.0950	0.0004
x_{it}^{TR} **with trend and increasing variance around trend**										
$\hat{\beta}$	1.5715	0.0012	6.0331	0.0019	4.4948	0.0020	0.6677	0.0019	1.0004	0.0028
$\hat{\alpha}$	−0.0180	0.0002	−0.9148	0.0004			0.2950	0.0001	0.0951	0.0006

Notes: True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$; $N = 4 \cdot 10^7$, $T = 5$; the # of observations for x_{it}^{ST} is 71 728 549, the corresponding # of observations for x_{it}^{RW} is 71 820 407, and for x_{it}^{TR} it is 72 225 012. For b^{OLS} the #s of observations are higher by $4 \cdot 10^7$ observations, since the first wave is not eliminated by the within-transformation or the first-differences transformation. See Table 1 for simulation results using a different seed for the RNG.

Table A5: Monte Carlo Analysis - Small Sample Estimates (alternative RNG seed)

	$b^{\mathbf{OLS}}$		$b^{\mathbf{WI}}$		$b^{\mathbf{FD}}$		$b^{\mathbf{FDC}}$		$b^{\mathbf{FDC}}_{\text{adjust}}$	
	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†	Mean	S.D.†
					x_{it} **and** a_i **random**					
x_{it}^{ST} **stationary**										
$\hat{\beta}$	1.6712	0.3813	0.9021	0.8081	0.7068	0.7169	0.5000	0.5967	0.9907	1.1873
$\hat{\alpha}$	−0.0345	0.0749	0.1167	0.1589			0.2906	0.0172	0.0977	0.2316
x_{it}^{RW} **follows random walk**										
$\hat{\beta}$	1.4301	0.2941	0.9472	0.6091	1.0063	0.6885	1.0063	0.5803	1.0070	0.5799
$\hat{\alpha}$	0.0134	0.0573	0.1071	0.1197			0.2887	0.0172	0.0947	0.1125
x_{it}^{TR} **with trend and increasing variance around trend**										
$\hat{\beta}$	1.5757	0.3670	6.0381	0.5989	4.5047	0.6652	0.6624	0.5984	0.9855	0.8892
$\hat{\alpha}$	−0.0183	0.0736	−0.9153	0.1151			0.2954	0.0184	0.0986	0.1854
					x_{it} **and** a_i **fixed**					
x_{it}^{ST} **stationary**										
$\hat{\beta}$	1.6721	0.3750	1.1176	0.7340	0.8101	0.6786	0.5115	0.5852	0.9987	1.1434
$\hat{\alpha}$	−0.0355	0.0734	0.0732	0.1439			0.2892	0.0167	0.0950	0.2230
x_{it}^{RW} **follows random walk**										
$\hat{\beta}$	1.4301	0.2922	0.3939	0.5121	0.5579	0.6429	1.0232	0.5863	1.0158	0.5813
$\hat{\alpha}$	0.0138	0.0561	0.2130	0.0999			0.2851	0.0170	0.0921	0.1100
x_{it}^{TR} **with trend and increasing variance around trend**										
$\hat{\beta}$	1.5864	0.3757	6.0438	0.5885	4.4193	0.6429	0.6837	0.5937	1.0111	0.8761
$\hat{\alpha}$	−0.0207	0.0750	−0.9135	0.1116			0.2939	0.0180	0.0928	0.1819

Notes: True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$; $N = 400$, $T = 5$; 10 000 replications. †S.D. denotes the empirical standard deviation of the coefficient in the simulated sample. In order to interpret these values in terms of standard errors for the respective mean-estimator, one has to multiply the value of the S.D. by $10\,000^{-0.5} = 0.001$. See Table 2 for simulation results using a different seed for the RNG.

Table A6: Monte Carlo Analysis - Estimated Standard Errors (alternative RNG seed)

	MC simulated	$\hat{se}_{analytic}(b^{FDC}_{adjust})$		**H**-adjusted $\hat{se}_{robust}(b^{FDC})$	
		true a_i and β	$\hat{a}_i$ and $\hat{\beta}$	White	cluster robust
x^{ST}_{it} **stationary**					
$\hat{se}(b^{FDC}_{adjust})$	1.1434	1.1550	0.9040	1.1631	1.1601
$\hat{se}(a^{FDC}_{adjust})$	0.2230	0.2253	0.1759	0.2269	0.2264
x^{RW}_{it} **follows random walk**					
$\hat{se}(b^{FDC}_{adjust})$	0.5813	0.5767	0.4517	0.5828	0.5830
$\hat{se}(a^{FDC}_{adjust})$	0.1100	0.1092	0.0849	0.1104	0.1112
x^{TR}_{it} **with trend and increasing variance around trend**					
$\hat{se}(b^{FDC}_{adjust})$	0.8761	0.8746	0.6357	0.8800	0.8844
$\hat{se}(a^{FDC}_{adjust})$	0.1819	0.1818	0.1315	0.1829	0.1839

Notes: True coefficient values: $\boldsymbol{\beta = 1}$, $\boldsymbol{\alpha = 0.1}$; $N = 400$, $T = 5$; x_{it} and a_i fixed; 10 000 replications. See Table 4 for simulation results using a different seed for the RNG.