Discussion Papers in Economics

No. 03/2021

Linear Fixed-Effects Estimation with Non-Repeated Outcomes

Helmut Farbmacher
Max Planck Society, MEA

Tauchmann, Harald
FAU Erlangen-Nürnberg

ISSN 1867-6707

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

Linear Fixed-Effects Estimation with Non-Repeated Outcomes*

Helmut Farbmacher

Max Planck Society, MEA

Harald Tauchmann

Universität Erlangen-Nürnberg

RWI – Leibniz Institut für Wirtschaftsforschung

CINCH – Health Economics Research Center

May 2021

Abstract

This paper demonstrates that popular linear fixed-effects panel-data estimators are biased and inconsistent when applied in a discrete-time hazard setting – that is, one in which the outcome variable is a binary dummy indicating an absorbing state, even if the data-generating process is fully consistent with the linear discrete-time hazard model. In addition to conventional survival bias, these estimators suffer from another source of – frequently severe – bias that originates from the data transformation itself and, unlike survival bias, is present even in the absence of any unobserved heterogeneity. We suggest an alternative estimation strategy, which is instrumental variables estimation using first-differences of the exogenous variables as instruments for their levels. Monte Carlo simulations and an empirical application substantiate our theoretical results.

JEL Codes: C23, C25, C41.

Keywords: linear probability model, individual fixed effects, discrete-time hazard, absorbing state, survival bias, instrumental variables estimation.

*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. We 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, also known as absorbing states. Death, retirement, firm bankruptcy, plant closure, technology adoption, and smoking initiation are just a few of many examples. Hazard models, also referred to as duration, failure-time, survival, time-to-event, and event history analysis, are commonly used to empirically analyze such outcomes. If an analysis is based on panel data in which the outcome is not observed continuously but only at a limited number of points in time[1], discrete-time hazard models are often regarded as the estimation method of choice. These models are simply stacked binary outcome models (Jenkins, 1995; Tutz and Schmid, 2016), such as probit, logit, or cloglog (Prentice and Gloeckler, 1978). The discrete-time hazard binary-outcome model considers the process that leads to the absorbing state to be a finite series of binary choices and is therefore simple and intuitive.

Following the general trend in applied econometrics towards using linear models, which are often not meant to specify the data-generating process correctly but rather to identify 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, 2001; Angrist and Pischke, 2009). One virtue of the linear probability model is that it allows unobserved individual heterogeneity to be removed as a possible source of bias in a straightforward fashion 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)

In fact, linear probability models with fixed effects at the level of observation units have recently been applied not only to repeated events, but also frequently to non-repeated event data. Examples are firm death (Frazer, 2005; Jacobson and von Schedvin, 2015; Fernandes and Paunov, 2015), retirement (Brown and Laschever, 2012), technology and confession adoption (Cantoni, 2012; Bogart, 2018), smoking onset (Do and Finkelstein, 2012), health insurance transition (Grunow and Nuscheler, 2014), school fees abolition (Harding and Stasavage, 2014), death (Bloemen et al., 2017), and unfollowing social media posts (Wang et al., 2020). There seems to be little awareness that the favorable properties of linear fixed-effects estimators do not apply in the same way to non-repeated events as they do to other kinds of dependent variables. Indeed, we are not aware of any article that explicitly establishes the properties of linear fixed-effects estimators in a discrete-time hazard setting.[2]

[1] This includes cases in which the time structure is intrinsically discrete (e.g. termination of a rolling fixed-period contract; see, for example, the application in section 5, where school teachers can retire by the end of an academic year) 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 (Cameron and Trivedi, 2005, p. 578).

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

In this paper, we demonstrate that conventional linear fixed-effects estimators (first-differences, within-transformation) exhibit severe shortcomings when applied to non-repeated event data. By failing to remove unobserved time-invariant individual heterogeneity and, additionally, rendering the conditional mean of the disturbance a function of the explanatory variables, they are biased and inconsistent in this setting. The bias originates from two sources: One is selective survival, which renders the unobserved heterogeneity correlated with the explanatory variables in the estimation sample. It is worth noting that 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 because it originates from taking first differences or using the within transformation, inducing the exogenous variables to enter the conditional mean of the disturbance. For this reason, this second source of bias – unlike the survival 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 these findings, we suggest an instrumental variables (IV) estimator, which instruments the exogenous variables by their own first-differences. This estimator, which can also be interpreted as an adjusted conventional first-differences estimator, eliminates the second source of bias.

The contribution of this paper is twofold. First, we elucidate why conventional fixed-effects estimators should not be used in a discrete-time hazard framework. Second, we suggest an alternative IV estimator that – though not consistent – usually suffers from a much smaller asymptotic bias than the familiar estimators and confines it to the survival bias. This is a source of bias researchers should already be aware of in our setting, even if the unobserved heterogeneity is uncorrelated with the explanatory variables in the population.

The remainder of this paper is organized as follows. Section 2 introduces the model framework. In section 3, we establish the biasedness of the conventional fixed-effects estimators and develop a simple instrumental variables estimator that cures the asymptotic bias that is driven by data transformation. In section 4, we use Monte Carlo simulations to compare the different estimators. Section 5 presents an empirical application that is based on the analysis of peer effects in the timing of retirement by Brown and Laschever (2012). Section 6 concludes.

non-repeated events as explanatory variables rather than outcome variables. Horowitz and Lee (2004) and Lee (2008) suggest fixed-effects estimators for continuous-time proportional hazard models with multiple spells. Horowitz (1999) proposes a random-effects estimator for a similar setting with single-spell data. Occasionally, the applied literature (e.g. McGarry, 2004; Wettstein, 2020; Finkelstein et al., 2019) touch upon the idea that linear fixed-effects estimation may not be advisable when the outcome is an absorbing state, but they do not dig deeper into this issue.

2 Model

In order to illustrate our argument straightforwardly, we analyze the considered estimators in a setting that is fully consistent with the linear hazard assumption. We therefore consider a linear probability model in a panel data setting, which we assume correctly captures the true data-generating process. 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, fixed and, compared to the number of cross-sectional units, small. Any argument regarding asymptotic properties is thus in terms of $N \to \infty$. y_{it} denotes a binary outcome variable. $\mathbf{x}_{it}$ is a row vector of exogenous explanatory variables observed for unit i in period t. The scalar a_i denotes unobserved, time-invariant individual heterogeneity. If $\mathbf{x}_{it}$ includes a constant term, we can assume $\mathrm{E}(a_i) = 0$ with no loss of generality. $\boldsymbol{\beta}$ is a column vector of coefficients subject to estimation. We assume $a_i + \mathbf{x}_{it}\boldsymbol{\beta} \in [0, 1]$ 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 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.[3] In other words, after observing $y_{it} = 1$ for the first time, any subsequent observations of i do not contain additional information about the data-generating process of interest because $\mathrm{P}(y_{it+s} = 1 | y_{it} = 1) = 1$ always holds for $s \geq 1$, irrespective of $\mathbf{x}_{it+s}$. In many applications, one might not even observe $\mathbf{x}_{it+s}$.[4] The number of periods $T_i \leq T$ for which unit i is (effectively) observed is therefore 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. The data-generating process (DGP) of y_{it} reads as

$$y_{it} = a_i + \mathbf{x}_{it}\boldsymbol{\beta} + \varepsilon_{it} \quad \text{with} \quad t \leq T_i \tag{1}$$

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

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

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

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

because y_{it} equals one for the terminal observation of a noncensored unit and is otherwise zero. Assuming zero conditional mean of the disturbance

$$\mathrm{E}\left(\varepsilon_{it}|a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, t \leq T_i\right) = 0 \tag{3}$$

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

$$\mathrm{P}\left(y_{it} = 1|a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, t \leq T_i\right) = a_i + \mathbf{x}_{it}\boldsymbol{\beta} \tag{4}$$

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

3 Estimation

It is well-known that pooled ordinary least squares (OLS) estimation cannot take into account the heterogeneity in a_i, which renders pooled OLS biased and inconsistent if $\mathrm{Cov}(a_i, \mathbf{x}_{it}) \neq \mathbf{0}$. One might therefore think of applying first-differences or the within transformation to the data in order to eliminate a_i and allow unbiased and consistent estimation by least squares. However, we show that these well-established approaches do not succeed in the setting we consider here , i.e. with the outcome variable being a binary dummy indicating an absorbing state. We propose an instrumental variable strategy that corrects the first-differences transformation. We could not find an analogous correction for the regular within transformation.

3.1 First differences with non-repeated outcomes

First, we examine first-differences estimation $\boldsymbol{b}^{\mathrm{FD}}$ of the linear probability model outlined above.

$$\boldsymbol{b}^{\mathrm{FD}} = \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\Delta y_{it}\right) = \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'y_{it}\right) \tag{5}$$

with $\Delta\mathbf{x}_{it} \equiv \mathbf{x}_{it} - \mathbf{x}_{it-1}$ denoting the vector of the first-differenced right-hand-side variables, and with $\Delta y_{it} \equiv y_{it} - y_{it-1} = y_{it}$ because $y_{it-1} = 0$ follows from the fact that the outcome is a non-repeated event. Therefore, the outcome remains unaffected by the first-differences transformation, implying that the disturbance needs to compensate fully for the transformation that is applied to the right-hand side. The disturbance in this regression model is $\varepsilon_{it}^{\mathrm{FD}} \equiv y_{it} - \Delta\mathbf{x}_{it}\boldsymbol{\beta}$ and its conditional mean reads as

$$\mathrm{E}\left(\varepsilon_{it}^{\mathrm{FD}}|a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, t \leq T_i\right) = a_i + \mathbf{x}_{it-1}\boldsymbol{\beta} \tag{6}$$

see Appendix A.1 for details. In our setting, taking first-differences fails to remove unobserved individual heterogeneity and also fails to generate a transformed disturbance that is conditional mean independent of the exogenous variables, violating the conditions required for unbiasedness. The impossibility of transforming the outcome variable is also reflected in the probability limit,

$$\begin{aligned}\operatorname{plim}(\boldsymbol{b}^{\mathrm{FD}}) = &\operatorname{plim}\left(\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\mathbf{x}_{it}\right)\right)\boldsymbol{\beta} \\ &+ \operatorname{plim}\left(\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'a_i\right)\right) \qquad (7)\end{aligned}$$

see Appendix A.2 for details. Equation (7) reveals that $\boldsymbol{\beta}$ enters $\operatorname{plim}(\boldsymbol{b}^{\mathrm{FD}})$ erroneously scaled by the matrix $\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\mathbf{x}_{it}\right)$, which we further denote as $\mathbf{G}$. This misscaling bias is present even if a_i is uncorrelated with the regressors in the population or even in the absence of any unobserved, time-invariant individual heterogeneity. It therefore does not originate from a failure to remove individual heterogeneity but from the nature of y, which does not allow us to transform the outcome variable. Because (6) is non-negative and thus systematically deviates from zero, it is obviously important to include a constant in $\Delta\mathbf{x}_{it}$. Note, however, that doing so does not alter the fact that the conditional mean is function of $\mathbf{x}_{it-1}$ and, consequently, that $\boldsymbol{\beta}$ enters (7) erroneously scaled.[5]

Our just-identified instrumental variables estimator with $\Delta\mathbf{x}_{it}$ serving as instruments[6] for $\mathbf{x}_{it}$ can solve this issue

$$\boldsymbol{b}^{\mathrm{IV}} = \left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'y_{it}\right) \qquad (8)$$

with

$$\operatorname{plim}\left(\boldsymbol{b}^{\mathrm{IV}}\right) = \boldsymbol{\beta} + \operatorname{plim}\left(\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\mathbf{x}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'a_i\right)\right) \qquad (9)$$

The suggested IV estimator also gives further insights into the shortcomings of $\boldsymbol{b}^{\mathrm{FD}}$ in our setting. From a two-stage least-squares perspective, first differences would only estimate the reduced-form coefficients. The corresponding first-stage estimates are collected in $\mathbf{G}$. To obtain informative estimates of the coefficients of interest, we have to rescale $\boldsymbol{b}^{\mathrm{FD}}$ by the inverse of the first-stage estimates, giving us $\mathbf{G}^{-1}\boldsymbol{b}^{\mathrm{FD}} = \boldsymbol{b}^{\mathrm{IV}}$. The shape of the rescaling matrix $\mathbf{G}^{-1}$ depends

[5] In the vast majority of applications, $\Delta\mathbf{x}_{it}$ effectively includes a constant anyway, because discrete-time hazard models usually allow for duration dependence of the baseline hazard by including a trend, a polynomial of t, or – more typically – a saturated set of wave indicators.

[6] Considering a constant term in the first stage of the suggested IV is essential, but to keep the notion simple, we continue denoting the vector of instruments as $\Delta\mathbf{x}_{it}$.

strongly on the data-generating process of the variables in $\mathbf{x}_{it}$. If $\mathbf{x}_{it}$ follows a random walk, asymptotically the first stage yields that each variable in $\mathbf{x}_{it}$ is best predicted by its own change and $\mathbf{G}^{-1}$ converges in probability to the identity matrix $\boldsymbol{I}$. In this special case, $\boldsymbol{b}^{\mathrm{IV}}$ coincides with the reduced-form estimator $\boldsymbol{b}^{\mathrm{FD}}$. Another interesting special case is that $\mathbf{x}_{it}$ is covariance stationary. The population first stage then asymptotically yields $\mathbf{G} = \frac{1}{2}\boldsymbol{I}$. This makes the rescaling of the reduced-form coefficients by a factor of two an important benchmark for settings in which the process we consider here exhibits little selectivity.

An important related implication is that $\boldsymbol{b}^{\mathrm{IV}}$ exists only if the first stage does not degenerate and $\mathbf{G}$ is nonsingular. In other words, $\Delta\mathbf{x}_{it}$ need to be informative as instruments for each and every element of $\mathbf{x}_{it}$. An extreme example of a violation of this condition 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 (cf. the application in section 5), age^{min} equals zero for all observations that enter the first-stage regressions, that is for $t > 1$, making $\mathbf{G}$ contain a column of zeros. More generally, the data-generating process of $\mathbf{x}_{it}$ is decisive for whether $\Delta\mathbf{x}_{it}$ is a promising instrument for $\mathbf{x}_{it}$. Weak instruments may therefore be an issue for $\boldsymbol{b}^{\mathrm{IV}}$ even in settings that are less extreme than that in the example above. The fact that our IV estimator cannot estimate some empirical models that can be estimated using the conventional first-differences estimator seems, at first glance, to be a major shortcoming of $\boldsymbol{b}^{\mathrm{IV}}$. However, the non-existence of $\boldsymbol{b}^{\mathrm{IV}}$ reveals that one cannot obtain information about some model parameters of interest, even if the corresponding coefficients are seemingly identified by $\boldsymbol{b}^{\mathrm{FD}}$. From (7) we see – ignoring for a moment the second term on the right-hand-side – that $\boldsymbol{b}^{\mathrm{FD}}$ converges in probability to a matrix-weighted sum of the true model parameters $\boldsymbol{\beta}$. Yet, β_l receives no weight in this sum if the lth column of $\mathbf{G}$ is $\mathbf{0}$ and, consequently, there is no way to retrieve any information about β_l from $\boldsymbol{b}^{\mathrm{FD}}$.

The conventional within-transformation estimator, which is frequently regarded as 'the fixed-effects estimator', is also biased in our setting. In fact, the situation is even worse. Unlike for $\boldsymbol{b}^{\mathrm{FD}}$, it is not the vector of observed lagged values $\mathbf{x}_{it-1}$ that enters the conditional mean of the disturbance, but a conditional expectation of $\bar{\mathbf{x}}_i$ that involves (i) future values of $\mathbf{x}_{it}$, which may not be observed for $t > T_i$, (ii) the unobserved individual heterogeneity, and (iii) the unknown coefficients of interest. For this reason, unlike the first-differences estimator, the conventional within-transformation provides no basis for an instrumental variables approach to eliminate the asymptotic misscaling bias. We discuss this issue in more detail in Appendix A.3.

3.2 Survival bias

Survival bias plays an important role when we analyze the remainder term of $\boldsymbol{b}^{\mathrm{IV}}$ in (9), which involves

$$\operatorname{plim}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}'a_i\right)=\operatorname{plim}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{x}_{it}'a_i\right)-\operatorname{plim}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\mathbf{x}_{it-1}'a_i\right) \tag{10}$$

This difference may deviate from $\mathbf{0}$ for two reasons: Firstly, (10) will obviously not vanish in the limit if we have $\operatorname{Cov}(a_i,\mathbf{x}_{it})\neq\operatorname{Cov}(a_i,\mathbf{x}_{it-1})$ in the population, i.e., the individual heterogeneity is correlated with *changes* in the explanatory variables, which would clearly conflict with using $\Delta\mathbf{x}_{it}$ as instruments. For the population, one may rule this out by assumption, and we focus on the cases for which this assumption is satisfied.

Yet, secondly, even assuming that a_i is uncorrelated with $\Delta\mathbf{x}_{it}$ in the population does not render (10) zero. The reason for this is survival bias, also referred to as 'weeding out' or the 'sorting effect' (Nicoletti and Rondinelli, 2010), in the sense that conditioning on $t\leq T_i$ affects the covariance of a_i and $\mathbf{x}_{it}$. This is most obvious for $\operatorname{Cov}(a_i,\mathbf{x}_{it-1}|t\leq T_i)$. Conditioning on $t\leq T_i$ means that $\mathbf{x}_{it-1}$ enters the conditional covariance only if $y_{it-1}=0$ holds. This implies – for a nonnegative $\boldsymbol{\beta}$ – 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 $\boldsymbol{\beta}=\mathbf{0}$. This does not apply one to one to $\operatorname{Cov}(a_i,\mathbf{x}_{it}|t\leq T_i)$ because 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 $\mathbf{x}_{it}$. Consequently, survival does not bias the estimates if $\mathbf{x}_{it}$ follows a random walk unless it exhibits a drift that is related to a_i. The smaller the persistence of $\mathbf{x}_{it}$, however, the more $\operatorname{Cov}(a_i,\mathbf{x}_{it}|t\leq T_i)$ may deviate from $\operatorname{Cov}(a_i,\mathbf{x}_{it-1}|t\leq T_i)$ due to selective survival, rendering a_i and $\Delta\mathbf{x}_{it}$ positively correlated for a positive $\boldsymbol{\beta}$. In addition to the dynamic properties of $\mathbf{x}_{it}$, the variance of a_i plays an important role in determining the size 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 (10) close to zero and, in turn, renders survival bias a minor issue.

It is important to note that survival bias is not specific to $\boldsymbol{b}^{\mathrm{IV}}$ or the first-differences estimator. Pooled OLS, for instance, also 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 $\operatorname{Cov}(a_i,\mathbf{x}_{it}|t\leq T_i)$ that matter. Thus, for OLS the survival bias acts in the opposite direction

and increases, rather than decreases, in the degree of persistence $\mathbf{x}_{it}$ exhibits. footnoteThis argument applies first of all to an ordinary least squares regression that includes a saturated set of period indicators. In this case, only the covariance between a_i and $\mathbf{x}_{it}$ conditional on t contributes to the bias. If no period indicators are included, another source of survival bias kicks in. More specifically, units with a small a_i and – given that $\boldsymbol{\beta}$ is positive – small $\mathbf{x}_{it}$ have a higher chance of surviving and contributing many observations to the estimation sample. This may generate, unconditionally on t, a positive correlation of a_i and $\mathbf{x}_{it}$. In other words, if time dummies are not included, the between-period correlation of a_i and $\mathbf{x}_{it}$ also contributes to the survival bias, which acts in the opposite direction to that in the within-period correlation. Moreover, because a conditional covariance rather than a difference in conditional covariances generates the survival bias, between-group heterogeneity (that is, differences in the level of $\mathbf{x}_{it}$ across the units i) contribute to the bias.

The result that $\boldsymbol{b}^{\mathrm{IV}}$ suffers only from survival bias critically hinges on having $\mathrm{Cov}(a_i, \Delta\mathbf{x}_{it}) = \mathbf{0}$ in the population. Contingent on the specific application, this non-testable assumption might be neither valid nor plausible. However, assuming $\mathrm{Cov}(a_i, \Delta^j\mathbf{x}_{it}) = \mathbf{0}$ instead, with the integer j greater than unity, may possibly be more plausible. In such settings, an estimator $\boldsymbol{b}^{\mathrm{IV,\,j}}$ based on higher-order differences $\Delta^j\mathbf{x}_{it}$ can – analogously to $\boldsymbol{b}^{\mathrm{IV}}$ – be constructed in a straightforward manner. In other words, $\boldsymbol{b}^{\mathrm{IV,\,j}}$ is a just-identified IV with $\Delta^j\mathbf{x}_{it}$ serving as instruments for $\mathbf{x}_{it}$. 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. Following the same line of argument as above, $\boldsymbol{b}^{\mathrm{IV,\,j}}$ suffers from survival bias but no other source of asymptotic bias, as long as $\mathrm{Cov}(a_i, \Delta^j\mathbf{x}_{it}) = \mathbf{0}$ holds in the population. Naturally, $\boldsymbol{b}^{\mathrm{IV,\,j}}$ coincides with $\boldsymbol{b}^{\mathrm{IV}}$ for $j = 1$, and with pooled OLS for $j = 0$. Evidently, taking higher-order differences removes much variation from the variables used as instruments and may result in weak instruments.

3.3 The Asymptotic Distribution of the suggested IV Estimator

Though the asymptotic properties of the just-identified IV estimator are, in general, well known, some special features of $\boldsymbol{b}^{\mathrm{IV}}$ must be taken into account when establishing its asymptotic distribution and developing approaches to estimate asymptotic standard errors. From (2) to (4) we obtain for the disturbance variance

$$\mathrm{Var}(\varepsilon_{it}|a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, t \leq T_i) \quad = \quad (a_i + \mathbf{x}_{it}\boldsymbol{\beta})\,(1 - a_i - \mathbf{x}_{it}\boldsymbol{\beta}) \tag{11}$$

This is the error variance of a standard linear probability model (cf. Aldrich and Nelson, 1984, p. 13), except for the constant being individual specific. For the disturbance covariance we obtain

$$\text{Cov}\left(\varepsilon_{it}, \varepsilon_{it-s} | a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, t \leq T_i\right) = 0 \quad \text{for} \quad s \geq 1 \tag{12}$$

because y_{it} is observed conditionally only on $y_{it-s} = 0$, and in consequence, conditionally on ε_{it-s} taking one specific value. The disturbances are therefore subject to the familiar form of heteroscedasticity that applies to the linear probability model, albeit without exhibiting within-group correlation. Building on general results for the properties of the linear instrumental variables estimator and using (9), (11), and (12), the asymptotic distribution of $\boldsymbol{b}^{\text{IV}}$ reads as

$$\boldsymbol{b}^{\text{IV}} \overset{a}{\sim} \textit{Normal}\left(\boldsymbol{\beta} + \mathbf{Q}_{\Delta\mathbf{x}\mathbf{x}}^{-1}\mathbf{Q}_{\Delta\mathbf{x}a}\,,\, \frac{1}{M}\mathbf{Q}_{\Delta\mathbf{x}\mathbf{x}}^{-1}\,\mathbf{Q}_{\sigma\Delta\mathbf{x}\Delta\mathbf{x}}\,\mathbf{Q}_{\Delta\mathbf{x}\mathbf{x}}^{\prime-1}\right) \tag{13}$$

where $\mathbf{Q}_{\Delta\mathbf{x}\mathbf{x}} \equiv \text{plim}\left(\frac{1}{M}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'\mathbf{x}_{it}\right)$, $\mathbf{Q}_{\Delta\mathbf{x}a} \equiv \text{plim}\left(\frac{1}{M}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\Delta\mathbf{x}_{it}{}'a_i\right)$, $\mathbf{Q}_{\sigma\Delta\mathbf{x}\Delta\mathbf{x}} \equiv \text{plim}\left(\frac{1}{M}\sum_{i=1}^{N}\sum_{t=2}^{T_i}\sigma_{it}^2\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}\right)$, $M \equiv \sum_{i=1}^{N}(T_i - 1)$, and $\sigma_{it}^2 \equiv (a_i + \mathbf{x}_{it}\boldsymbol{\beta})\,(1 - a_i - \mathbf{x}_{it}\boldsymbol{\beta})$.

Estimating the asymptotic covariance matrix is not straightforward, however, because estimating σ_{it}^2 is not trivial. Firstly, consistent estimators are not available for a_i and, due to survival bias, also not for $\boldsymbol{\beta}$ except in very special cases. Secondly, the prediction $\left(\hat{a}_i + \mathbf{x}_{it}\hat{\boldsymbol{\beta}}\right)$ may well be negative or exceed unity, leading to invalid estimates of σ_{it}^2. A natural alternative to the parametric estimation approach is to use the heteroscedasticity-robust White (1980) estimator. More specifically, this is estimating $\mathbf{Q}_{\sigma\Delta\mathbf{x}\Delta\mathbf{x}}$ as $\frac{1}{M}\sum_{i=1}^{N}\sum_{t=2}^{T_i}e_{it}^2\Delta\mathbf{x}_{it}{}'\Delta\mathbf{x}_{it}$, where e_{it} denotes the residuals from first-differences-based IV estimation. However, due to the survival bias of $\boldsymbol{b}^{\text{IV}}$, this estimator may not be consistent for $\mathbf{Q}_{\sigma\Delta\mathbf{x}\Delta\mathbf{x}}$ – although simulations suggest that the familiar robust estimator still approximates the true variances fairly accurately. Moreover, the heteroscedasticity-robust estimator is probably conservative because it typically overestimates the variance by using residuals that capture bias.

4 Monte Carlo Analysis

In this section, we present results of our Monte Carlo (MC) simulations. For y_{it} we consider the data-generation process described in section 2, with $\mathbf{x}_{it}$ consisting of just one variable x_{it}.[7] The

[7] 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 restrictions on the DGPs of $\mathbf{x}_{it}$ and a_i to guarantee $\text{P}\left(y_{it} = 1 | a_i, x_{it}, t \leq T_i\right) \in [0, 1]$. For this reason, applied researchers might be interested primarily in the performance – in terms of estimating average marginal effects – of the linear estimators when applied to data that are generated by a process consistent with classical nonlinear binary outcome models such as probit or logit. The simulation results presented in the Appendix A.4 consider this case.

slope coefficient is $\beta = 1$[8] and the constant is $\alpha = 0.1$. We specify a_i to be iid. continuously uniformly $U(-0.05, 0.05)$ distributed. We consider a short panel with $T = 5$. We examine the properties of the different estimators for three data-generating processes for x_{it}:

(i) $x_{it}^{ST} = a_i + 0.165 + 0.07 \cdot \zeta_{it}$, with $\zeta_{it} \sim$ iid. $\mathcal{B}(0.2, 0.2)$,
i.e., x_{it}^{ST} is stationary

(ii) $x_{it}^{RW} = x_{it-1}^{RW} + \big(0.1 \cdot \nu_{it} - 0.05\big)$, with $x_{i1}^{RW} = a_i + 0.2$, and $\nu_{it} \sim$ iid. $\mathcal{B}(0.2, 0.2)$,
i.e., x_{it}^{RW} follows a random walk without drift

(iii) $x_{it}^{TR} = a_i + 0.175 + 0.025 \cdot t \cdot \eta_{it}$, with $\eta_{it} \sim$ iid. $\mathcal{B}(0.2, 0.2)$,
i.e., x_{it}^{TR} exhibits a trend and increasing variance around the trend

$\mathcal{B}$ denotes the beta distribution. For all three data-generating processes, a_i is positively correlated with x_{it} but uncorrelated with Δx_{it} in the population.[9] In addition to $\boldsymbol{b}^{\text{IV}}$ and $\boldsymbol{b}^{\text{FD}}$, we consider pooled ordinary least squares $\boldsymbol{b}^{\text{OLS}}$ as reference and the within-transformation estimator $\boldsymbol{b}^{\text{WI}}$, which appears to be the most popular fixed-effects estimator in applied work. With regard to the first-differences estimator, we focus on the version with a constant term because its reduced-form interpretation only holds with a constant included.[10]

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 these are not generated non-parametrically by replicating the analysis, but are calculated following the procedure suggested in section 3.3. They are therefore not meant for assessing the sampling variability of the different estimation methods by means of an MC simulation, but rather are reported only to provide some intuition on 'how distant from infinite size' the artificial sample is because the standard errors would collapse to zero in this case.

To study the estimators' properties in a sample of moderate size, we choose $N = 4\,000$. Here we replicate the regressions 10 000 times. The reported coefficients are averages over the replications, and the reported standard deviations are calculated non-parametrically from the simulated

[8] This choice was made simply to make the simulation results more easily 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 therefore yield a β-coefficient whose magnitude would be better in line with what one would consider a marginal effect in a binary outcome model.

[9] 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 $a_i + \alpha + x_{it}\beta \in [0, 1]$ is satisfied for any i and any $t = 1, \ldots, 5$. For the correlations with the unobserved heterogeneity, we have $\mathrm{Cor}(a_i, x_{it}^{ST}) = 0.70$, $\mathrm{Cor}(a_i, x_{it}^{RW}) = 0.44$, and $\mathrm{Cor}(a_i, x_{it}^{TR}) = 0.59$. The beta $\mathcal{B}(0.2, 0.2)$ distribution is chosen to have – e.g. compared to using $U(0, 1)$ – much variation in x_{it}, albeit using a continuous distribution with bounded support.

[10] Results for $\boldsymbol{b}^{\text{FD}}$ without a constant, which except for the case of x_{it} following a random walk parallel the results for $\boldsymbol{b}^{\text{WI}}$, are available upon request.

distribution. Thus, they illustrate the degree to which the different estimators suffer from sampling error in our setting. We evaluate the estimators' small-sample properties conditional on a_i and x_{it}. Hence, we keep a_i and x_{it} fixed and only resample y_{it} in each replication. See Table 2 for the simulation results.[11]

4.1 Large-Sample Properties

The large-sample results are presented in Table 1. In line with x_{it} being positively correlated with the unobserved individual heterogeneity a_i, the estimated β-coefficient from $\boldsymbol{b}^{\text{OLS}}$ exhibits substantial upward bias. The results for $\boldsymbol{b}^{\text{IV}}$ are also in line with the theoretical large-sample properties derived above. First, $\boldsymbol{b}^{\text{IV}}$ hits the true value of β almost exactly. The simulations therefore point to survival bias being of little importance in our setting. For x_{it}^{ST} and x_{it}^{TR}, the estimate of β is marginally bigger than the true parameter, which is in line with the direction of the survival bias predicted by theory. Yet, the deviation from unity is small enough that it could also be attributed to sampling error, though the standard errors are tiny. The fact that $\boldsymbol{b}^{\text{IV}}$ yields an estimate $\hat{\beta}$ closest to the true coefficient value for x_{it}^{RW} is also in line with theory because no survival bias occurs in this case. A selection effect is, however, captured by the estimates of the baseline hazard $\hat{\alpha}$, which are somewhat smaller than 0.1. This deviation from the true parameter value captures that the averages of a_i in the estimation samples are smaller than zero by about 0.005 due to selective survival.

The simulation results for $\boldsymbol{b}^{\text{FD}}$ also confirm the theoretical results. If x_{it} follows a random walk, the estimated slope coefficient almost coincides with its counterpart from the IV estimation, reflecting that, in this case, the reduced form estimator and IV asymptotically coincide. For the stationary right-hand-side variable x_{it}^{ST}, $\boldsymbol{b}^{\text{FD}}$ yields a slope coefficient of almost exactly $\beta/2$. This is approximately the value one should expect because of the misscaling bias suffered by $\boldsymbol{b}^{\text{FD}}$. If the mean and variance of x_{it} are functions of time, the slope coefficient of $\boldsymbol{b}^{\text{FD}}$ is erroneously scaled by a factor between one-half and one. The estimated constants strongly deviate from the true α and also from the estimates yielded by $\boldsymbol{b}^{\text{IV}}$. They do not represent meaningful estimates of the baseline hazard but capture the fact that the first-differences model has non-zero mean disturbances as shown in (6).

Finally we turn to the results of the within estimator $\boldsymbol{b}^{\text{WI}}$. For the stationary regressor x_{it}^{ST}, the estimated slope coefficient exhibits a substantial bias towards zero. The simulation also yields a sizable downward bias if x_{it} follows a random walk. When x_{it} has a trend, $\boldsymbol{b}^{\text{WI}}$ exhibits an upward

[11] Also resampling a_i and x_{it} makes little difference in the considered settings, the results get even closer to their large-sample counterparts. In very small samples, however, the behavior of $\boldsymbol{b}^{\text{WI}}$ becomes sensitive to whether a_i and x_{it} are resampled in each replication.

Table 1: Monte Carlo Analysis - Large Sample Estimates

	$\boldsymbol{b}^{\text{OLS}}$		$\boldsymbol{b}^{\text{WI}}$		$\boldsymbol{b}^{\text{FD}}$		$\boldsymbol{b}^{\text{IV}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x_{it}^{ST} **stationary**								
$\hat{\beta}$	1.4866	0.0010	0.9023	0.0017	0.5043	0.0013	1.0045	0.0025
$\hat{\alpha}$	0.0010	0.0002	0.1161	0.0003	0.2896	0.0001	0.0942	0.0005
x_{it}^{RW} **follows random walk**								
$\hat{\beta}$	1.2574	0.0007	0.9447	0.0013	0.9991	0.0013	0.9992	0.0012
$\hat{\alpha}$	0.0468	0.0001	0.1075	0.0003	0.2859	0.0001	0.0952	0.0002
x_{it}^{TR} **with trend and increasing variance around trend**								
$\hat{\beta}$	1.4350	0.0010	3.9783	0.0014	0.6685	0.0013	1.0015	0.0019
$\hat{\alpha}$	0.0095	0.0002	−0.5017	0.0003	0.2947	0.0001	0.0949	0.0004

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 71 732 683, the corresponding # of observations for x_{it}^{RW} is 71 929 363, and for x_{it}^{TR} it is 72 211 807. For $\boldsymbol{b}^{\text{OLS}}$ the #s of observations are higher by $4 \cdot 10^7$ observations because no wave is eliminated by the within transformation or the first-differences transformation.

bias of bizarre magnitude. This simulation result is puzzling at first glance but can be easily explained. If x_{it} has a trend, the within-transformed regressor $\ddot{x}_{it}$ is strongly determined by the individual survival time T_i. For the trend being positive, units that survive longer mechanically exhibit lager values of the group mean $\bar{x}_i^{TR}$ and, in turn, exhibit, conditionally on t, smaller values of $\ddot{x}_{it}^{TR}$. This generates a strong spurious correlation between $\ddot{x}_{it}^{TR}$ and y_{it}; see Appendix A.3 for a more detailed discussion. For this reason, $\boldsymbol{b}^{\text{WI}}$ may even yield a large negative estimate for β if the trend in x_{it}^{TR} is negative. [12] The extreme bias that $\boldsymbol{b}^{\text{WI}}$ exhibits when using a regressor with a trend is moderated if a set of wave indicators is included as additional right-hand-side variables. However, the time effects themselves are then severely biased. Moreover, including period indicators may exacerbate the bias of $\boldsymbol{b}^{\text{WI}}$ for other DGPs; see Appendix A.5 for a more detailed discussion of how including time indicators affects the results of the different estimators.

4.2 Small-Sample Properties

Turning to the simulations that consider a sample of moderate size, Table 2 indicates that – in terms of biases – the small-sample results are very close the their large-sample counterparts. That is, the simulations do not point to a sizable small-sample bias of $\boldsymbol{b}^{\text{IV}}$, while the other three estimators turn out to be biased. $\boldsymbol{b}^{\text{WI}}$ appears to be the most sensitive to the change in the simulation design as it exhibits a substantially bigger bias for the case of x_{it} following a random walk compared to large-sample simulation. The results in Table 2 therefore suggest that the asymptotic properties of the estimators, in particular those derived for $\boldsymbol{b}^{\text{IV}}$, matter in samples of a size familiar to applied researchers, at least in settings similar to those considered in the present simulation. This also applies to the standard errors estimated for $\boldsymbol{b}^{\text{IV}}$. The non-parametric, simulation-based estimated standard errors match the averages of their analytically estimated counterparts (re-

[12] For instance, $\hat{\beta}^{\text{WI}} = -2.0734$ if DGP (iii) is changed to $x_{it}^{TR} = a_i + 0.225 - 0.025 \cdot t \cdot \eta_{it}$.

Table 2: Monte Carlo Analysis - Small Sample Estimates

	b^{OLS}		b^{WI}		b^{FD}		b^{IV}	
	Mean	S.D.† (S.E.‡)	Mean	S.D.† (S.E.‡)	Mean	S.D.† (S.E.‡)	Mean	S.D.† (S.E.‡)
x_{it}^{ST} **stationary**								
$\hat{\beta}$	1.4896	0.1031	0.9134	0.1630	0.5065	0.1306	1.0079	0.2589
		(0.1025)		(0.1746)		(0.1294)		(0.2559)
$\hat{\alpha}$	0.0007	0.0203	0.1142	0.0321	0.2901	0.0053	0.0938	0.0506
		(0.0201)		(0.0344)		(0.0054)		(0.0500)
x_{it}^{RW} **follows random walk**								
$\hat{\beta}$	1.2569	0.0701	0.8739	0.1091	1.0142	0.1242	1.0023	0.1219
		(0.0703)		(0.1277)		(0.1257)		(0.1228)
$\hat{\alpha}$	0.0474	0.0135	0.1215	0.0214	0.2857	0.0052	0.0950	0.0232
		(0.0136)		(0.0248)		(0.0053)		(0.0234)
x_{it}^{TR} **with trend and increasing variance around trend**								
$\hat{\beta}$	1.4401	0.1014	3.9917	0.1443	0.6718	0.1307	1.0052	0.1936
		(0.1023)		(0.1446)		(0.1293)		(0.1922)
$\hat{\alpha}$	0.0088	0.0204	−0.5042	0.0277	0.2951	0.0055	0.0944	0.0405
		(0.0205)		(0.0283)		(0.0056)		(0.0402)

Notes: True coefficient values: $\boldsymbol{\beta} = \mathbf{1}$, $\boldsymbol{\alpha} = \mathbf{0.1}$; $N = 4\,000$, $T = 5$; 10 000 replications. †S.D. denotes the empirical standard deviation of the estimated coefficient in the simulation. ‡S.E. denotes the mean of the (heteroscedasticity-robust) estimated standard errors calculated in each replication.

ported in parentheses in Table 2) well. Moreover, they are almost exactly 100 times bigger than their analytically derived counterparts reported in Table 1. This factor mirrors the square root of the relative sample size. These results suggest that the method of White (1980) does a good job estimating standard errors for $\boldsymbol{b}^{\text{IV}}$, at least in settings comparable to those considered in the simulation. Not surprisingly, $\boldsymbol{b}^{\text{OLS}}$ has the smallest variance because it uses all variation in x_{it}. As an IV, which by construction is picky in terms of the variation in x_{it} that is used, $\boldsymbol{b}^{\text{IV}}$ exhibits a relatively large variance. The variance of $\boldsymbol{b}^{\text{IV}}$ is smallest if x_{it} follows a random walk. This finding makes sense because, in this case, all variation in x_{it}^{RW}, except for the variation in the initial values, is explained by the instrument Δx_{it}^{RW}.

4.3 Analyzing the Survival Bias

The simulation results discussed above provide little evidence that survival bias is a substantial issue for the considered estimation methods, for $\boldsymbol{b}^{\text{IV}}$ in particular. This judgment, however, might just be an artifact of the choice of model parameters, and survival bias might be a more important issue in different settings. In order to account for this possibility and allow for settings more prone to survival bias, we adjust the simulation design in several ways: (i) We analyze the behavior of the estimators as a function of the variance of the unobserved heterogeneity. More precisely, we sample a_i from the $U\left(\frac{-q}{2}, \frac{q}{2}\right)$ distribution and vary q between 0 and 0.96 and thus consider values for $\sqrt{\text{Var}(a_i)} = \frac{q}{\sqrt{12}}$ in the range from 0 to 0.277. This allows for standard deviations that substantially exceed the value considered in the simulations discussed so far. (ii) To guarantee valid hazard rates within the unit interval, we have to make the constant a function of q; more

specifically we specify $\alpha = q/2$. (iii) Considering larger values of α decreases the survival rate in the artificial sample. For this reason, we adjust the number of units to $N = 10^8$ and the length of the panel to $T = 3$. In this section we therefore analyze the properties of the estimators only in a large sample. (iv) We exclude the initial wave from the estimation sample to bring left-truncation into the simulation, which is a common feature of data used in duration analyses (e.g. Kalbfleisch and Prentice, 2002). (v) We consider a DGP for which a_i and x_{it} are uncorrelated in the population. This makes survival bias the only source of bias for $\boldsymbol{b}^{\text{OLS}}$, allows this type of bias to be compared between $\boldsymbol{b}^{\text{OLS}}$ and $\boldsymbol{b}^{\text{IV}}$. (vi) Because survival bias in $\boldsymbol{b}^{\text{OLS}}$ originates from between-group variation in x_{it} whereas within-group variation generates the survival bias in $\boldsymbol{b}^{\text{IV}}$, we consider a stationary DGP for x_{it} that involves both sources of variation. More specifically, we consider $x_{it} = \frac{1-q}{2}(\mu_i + \omega_{it})$, with $\mu_i \sim$ iid. $\mathcal{B}(0.2, 0.2)$ and $\omega_{it} \sim$ iid. $\mathcal{B}(0.2, 0.2)$. As above, q needs to enter the DGP to bound the hazard rate to the unit interval. To consider an alternative DGP, we replace $\mathcal{B}(0.2, 0.2)$ with $\mathcal{B}(6, 2)$. Otherwise, the DGP for y_{it} is the same as above, with unity still being the true value of β.

Figure 1 depicts the slope coefficients estimated by $\boldsymbol{b}^{\text{OLS}}$, $\boldsymbol{b}^{\text{WI}}$, $\boldsymbol{b}^{\text{FD}}$, and $\boldsymbol{b}^{\text{IV}}$ as functions of $\sqrt{\text{Var}(a_i)}$.[13] The upper panel refers to the design for which $\mathcal{B}(0.2, 0.2)$ enters the DGP of x_{it}, and the lower panel refers to the design that involves $\mathcal{B}(6, 2)$. Dashed lines represent estimated 95 percent confidence intervals, which get wider with increasing values of q because x_{it} then exhibits less and less variation. The upper thin solid line marks the true parameter value $\beta = 1$. The lower thin solid line signifies the value $\boldsymbol{b}^{\text{FD}}$ would take if misscaling bias would be its sole source of error, i.e. the β-element of $\mathbf{G}\boldsymbol{\beta}$.[14]

For $\text{Var}(a_i) = 0$, i.e. in the absence of any unobserved heterogeneity, $\boldsymbol{b}^{\text{OLS}}$ and $\boldsymbol{b}^{\text{IV}}$ hit the true parameter value of 1 almost perfectly. This does not apply to $\boldsymbol{b}^{\text{FD}}$ and $\boldsymbol{b}^{\text{WI}}$, which are severely biased. This illustrates that the misscaling bias in the latter two estimators does not originate from the failure to remove unobserved heterogeneity, but from the within- and the first-differences transformation itself. The vertical distance between the two thin solid subsidiary lines is the misscaling bias in $\boldsymbol{b}^{\text{FD}}$ that is eliminated by $\boldsymbol{b}^{\text{IV}}$. Hence, if a_i exhibits little variation, misscaling is almost the only source of bias in $\boldsymbol{b}^{\text{FD}}$, implying that $\boldsymbol{b}^{\text{IV}}$ is close to asymptotic unbiasedness. If, however, the variance of the unobserved heterogeneity increases, the survival bias kicks in. This applies not only to $\boldsymbol{b}^{\text{FD}}$ and $\boldsymbol{b}^{\text{IV}}$ but also to $\boldsymbol{b}^{\text{OLS}}$. Yet, as predicted, for the latter the bias operates in the opposite direction. With substantial survival bias in $\boldsymbol{b}^{\text{FD}}$ – in Figure 1 this is the vertical distance between $\boldsymbol{b}^{\text{FD}}$ and the lower thin subsidiary line – $\boldsymbol{b}^{\text{IV}}$, which rescales the reduced form

[13] Because we consider regressions with the initial period excluded, i.e. only two waves enter the estimation sample, $\boldsymbol{b}^{\text{WI}}$ coincides with $\boldsymbol{b}^{\text{FD}}$ without constant. Moreover, $\boldsymbol{b}^{\text{FD}}$ coincides with the within-transformation estimator, which includes a wave indicator.

[14] Because of left-truncation, this line does not exactly hit the benchmark value of 0.5 for $\text{Var}(a_i) = 0$.

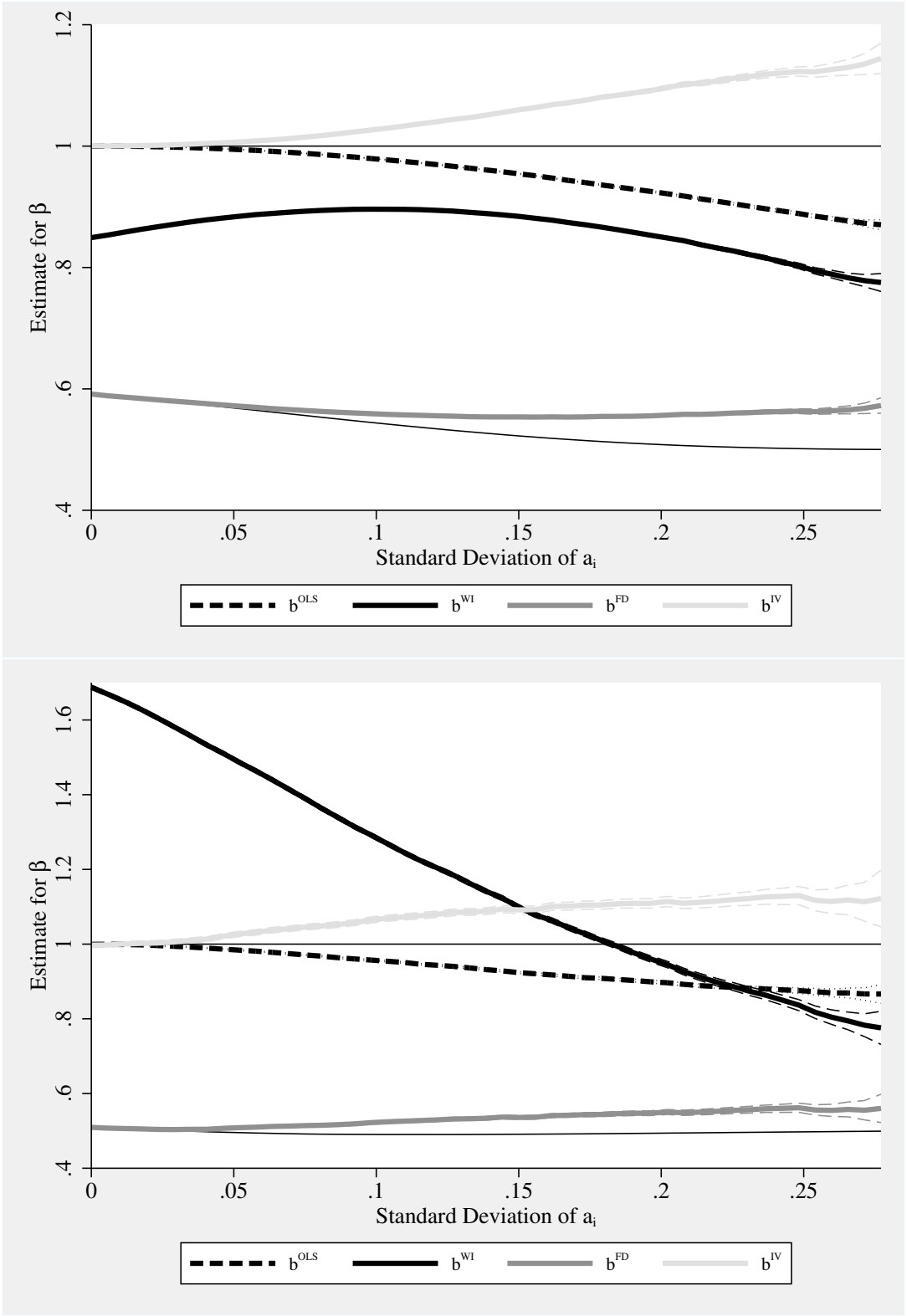

Figure 1: Estimated β coefficients as functions of $\sqrt{\text{Var}(a_i)} = q/\sqrt{12}$. DGPs of a_i and x_{it}: a_i sampled from the $U(-q/2, q/2)$ distribution; $x_{it} = {}^{(1-q)}/_{2}\,(\mu_i + \omega_{it})$ with μ_i and ω_{it} independently sampled from the beta $\mathcal{B}(\mathbf{0.2, 0.2})$ (upper panel) and the beta $\mathcal{B}(\mathbf{6, 2})$ (lower panel) distribution. q varies in the range between 0 and 0.96. Dashed subsidiary lines represent 95 percent confidence intervals. The thin solid subsidiary lines indicate the true coefficient value $\beta = 1$ and the β-element of $\mathbf{G}\boldsymbol{\beta}$, respectively. **Source:** Authors' own simulations.

estimator $\boldsymbol{b}^{\mathrm{FD}}$, does not hit the true parameter value. Eliminating the misscaling bias comes at the cost of rescaling the survival bias in $\boldsymbol{b}^{\mathrm{FD}}$. Nevertheless, according to our simulations, misscaling is the dominant source of bias in $\boldsymbol{b}^{\mathrm{FD}}$, even if the variance of a_i is very large. Thus, using $\boldsymbol{b}^{\mathrm{IV}}$ instead of the reduced form estimator $\boldsymbol{b}^{\mathrm{FD}}$ still reduces the asymptotic bias substantially. This suggests that using $\boldsymbol{b}^{\mathrm{IV}}$ is advisable even in settings that are prone to survival bias. Moreover, the survival bias in $\boldsymbol{b}^{\mathrm{IV}}$ seems to be of similar magnitude to that in $\boldsymbol{b}^{\mathrm{OLS}}$, yet as discussed above, this crucially depends on the properties of the DGP of x_{it}.

The behavior of $\boldsymbol{b}^{\mathrm{WI}}$ turns out to be rather strange in our simulation. While considering different beta distributions in the DGP of x_{it} has little effect on the behaviors of $\boldsymbol{b}^{\mathrm{OLS}}$, $\boldsymbol{b}^{\mathrm{FD}}$, and $\boldsymbol{b}^{\mathrm{IV}}$, the bias of $\boldsymbol{b}^{\mathrm{WI}}$ is very sensitive to this choice. With $\mathcal{B}(0.2, 0.2)$, the within estimator is biased towards zero throughout, yet the size of the bias is not monotonic in $\sqrt{\mathrm{Var}(a_i)}$. If, however, $\mathcal{B}(6, 2)$ enters the DGP of x_{it}, $\boldsymbol{b}^{\mathrm{WI}}$ may – depending on the variance of the unobserved heterogeneity – exhibit a substantial upward bias, a substantial downward bias, or no bias at all. This finding corroborates our earlier result that $\boldsymbol{b}^{\mathrm{WI}}$ is very sensitive to how the right-hand-side variables are generated and may exhibit a severe bias in any direction.

5 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 2, column 7). Subsequently, we compare these results to those we obtain from applying the estimators discussed in the previous sections. Thanks to the fact that the data and the code are published in Brown and Laschever (2012/2019), replication of the original results is straightforward. We provide here only very limited information about the analysis of Brown and Laschever (2012). Readers interested in the details of their paper, including in particular results from further empirical models, are referred to the original article.

The analysis of Brown and Laschever (2012) is concerned with the retirement behavior of school teachers in the Los Angeles Unified School District (LAUSD). Although their article focuses on how retirement decisions are affected by the retirement behavior of peer teachers, we concentrate on a relatively small and simple model specification from Brown and Laschever (2012) that

does not look at peer effects, but addresses the upstream[15] question of whether financial incentives matter for the timing of retirement.

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) individual 'pension wealth' (present value of future pension income, Brown and Laschever, 2012, p. 99) and a dummy for a 'positive peak value' (indicating that postponing retirement increases pension wealth, Brown and Laschever, 2012, p. 99), which capture the financial incentives for retiring in the current period (Table 3, first panel); (ii) teacher-level controls (Table 3, second panel), (iii) school-level controls (Table 3, third panel), (iv) age indicators, with an age of 55 years – the most represented age in the sample – serving as reference (Table 3, fourth panel), (v) panel wave (academic year) indicators (Table 3, fifth panel), and (vi) teacher fixed effects. Because teachers are no longer observed in the data after they have retired, retirement acts as an absorbing state. This empirical model therefore fits very well into our framework.

Columns 1 and 2 of Table 3, denoted $\boldsymbol{b}^{\mathrm{WI}}$, simply replicate the analysis of Brown and Laschever (2012), for which the popular within-transformation estimator was used. 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, and for the other, the age decreases. Excluding these six observations has virtually no impact on the estimated coefficients. 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 'pension wealth' is positive and statistically highly significant while the coefficient of 'positive peak value' is negative and statistically significant as well. This confirms that teachers respond to financial incentives in timing their retirement, which is crucial for the further analysis of Brown and Laschever (2012). We compare the results from this model to the corresponding ones from alternative estimation methods, more specifically $\boldsymbol{b}^{\mathrm{FD}}$ and $\boldsymbol{b}^{\mathrm{IV}}$.

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. This renders the matrix $\mathbf{G}$ singular for the reason discussed in section 3. Therefore, the original model specification cannot be estimated one to one by $\boldsymbol{b}^{\mathrm{IV}}$. For this reason, we exclude the dummy indicating the youngest age cohort in the sample from the IV estimation. Naturally, also one wave indicator must be dropped if the estimation is based on first-differences.

[15] In the key regressions of Brown and Laschever (2012), identification rests on exogenous variation in the financial incentives for retirement induced by two unexpected pension reforms. The effects of these reforms on teachers were heterogeneous, allowing the reform-induced changes in financial incentives to be used as instruments for peers' retirement behavior. Establishing that financial incentives affect retirement decisions is thus a crucial precondition for identifying peer effects.

Table 3: Brown and Laschever (2012) Simple Retirement Model and Alternative Empirical Models

	$\boldsymbol{b}^{\text{WI}}$ ‡		$\boldsymbol{b}^{\text{FD}}$		$\boldsymbol{b}^{\text{IV}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
pension wealth ($100 000)	0.038***	0.008	0.014*	0.008	0.021	0.016
positive peak value	−0.088***	0.020	−0.066***	0.020	−0.100***	0.028
salary ($10 000)	0.023	0.016	0.020	0.017	0.054	0.043
years of service in LAUSD squared	0.001***	0.000	0.002***	0.000	−0.000	0.000
av. age of teachers aged ≥ 55 at school	−0.001	0.003	−0.002	0.003	−0.000	0.007
av. service of teachers aged ≥ 55 at school	−0.002	0.001	−0.002	0.001	−0.004*	0.002
pupil to teacher ratio	0.001	0.001	0.001	0.001	0.010**	0.005
share of teachers with masters or higher	−0.115	0.074	−0.139*	0.072	−0.399**	0.173
share of female teachers	0.142**	0.067	0.145**	0.061	0.310*	0.178
av. rank on standardized math test	0.000	0.003	0.002	0.003	−0.004	0.009
# of teachers aged ≥ 55 at school	−0.001	0.001	−0.002	0.001	0.002	0.001
age = 53 years	0.306***	0.025	0.350***	0.029		
age = 54 years	0.141***	0.013	0.168***	0.015	−0.000	0.009
age = 56 years	−0.156***	0.013	−0.178***	0.015	−0.025***	0.009
age = 57 years	−0.310***	0.025	−0.356***	0.029	−0.026**	0.011
age = 58 years	−0.452***	0.036	−0.522***	0.043	−0.020	0.013
age = 59 years	−0.567***	0.048	−0.658***	0.057	0.008	0.016
age = 60 years	−0.631***	0.060	−0.738***	0.071	0.060***	0.019
age = 61 years	−0.694***	0.072	−0.799***	0.086	0.066***	0.022
age = 62 years	−0.686***	0.086	−0.785***	0.101	0.119***	0.027
age = 63 years	−0.727***	0.097	−0.804***	0.114	0.077**	0.032
age = 64 years	−0.788***	0.109	−0.860***	0.127	0.025	0.031
age = 65 years	−0.794***	0.119	−0.871***	0.141	0.066**	0.033
age ≥ 66 years	−0.825***	0.131	−0.898***	0.154	0.046	0.036
academic year 2000-01	0.092***	0.016				
academic year 2001-02	0.197***	0.029	0.030***	0.010	0.015*	0.009
constant	−0.412*	0.224	0.103***	0.020	−0.437	0.472

Notes: ‡ **Replication of the results of Brown and Laschever (2012, p. 104; table 2, column 7)**, subject to a marginal modification of the estimation sample due to inconsistent age information. *** 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. Because N observations are redundant in the within-transformed model, the number of non-redundant observations does not deviate between the within-transformed and the first-differences model. 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. Although 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 authors' own estimations; variable names are – subject to minor modifications – borrowed from the online appendix to Brown and Laschever (2012); see `https://www.aeaweb.org/aej/app/app/2011-0132_app.pdf`.

Except for the key coefficients that capture the effects of financial incentives on retirement, the estimates obtained from $\boldsymbol{b}^{\text{FD}}$ are very close to the original ones. Yet, the coefficients of 'positive peak value' and in particular 'pension wealth' are smaller in magnitude. However, they stay – at least marginally – statistically significant. Thus, in qualitative terms, the results from first-differences estimation does not challenge the main results of the original within-transformation based regression.

Turning to the results from IV estimation, we see that this pattern changes. The coefficients of the control variables are substantially different both in terms of magnitude and in terms of statistical significance. School-characteristics seem, for instance, to be of greater importance for retirement if one considers the results from $\boldsymbol{b}^{\text{IV}}$. The divergence of the estimation results cannot be attributed to $\Delta\mathbf{x}_{it}$ being only weak instruments for $\mathbf{x}_{it}$ because the Kleibergen and Paap (2006) test

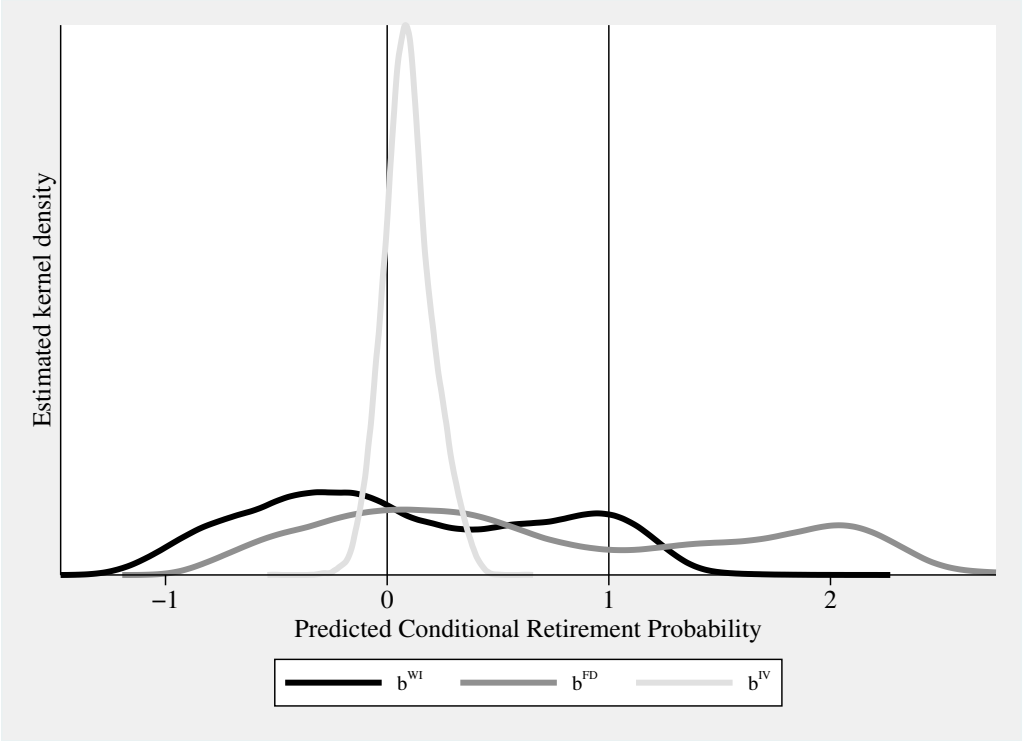

Figure 2: Sample distribution of predicted conditional retirement probabilities from $\boldsymbol{b}^{\text{WI}}$, $\boldsymbol{b}^{\text{FD}}$, and $\boldsymbol{b}^{\text{IV}}$. Predictions from within-transformed estimator based on three waves, i.e. 21 290 obs.; predictions from first-differences estimators based on two waves, i.e. 12 968 obs. The mean outcome (rel. frequency of retirement events) is 0.085 in the three-wave sample and 0.095 in the two-wave sample. **Source:** Authors' own calculations based on Brown and Laschever (2012/2019).

clearly rejects[16] the null of general underidentification, and the Sanderson and Windmeijer (2016) test rejects the null for each individual regressor. With respect to the coefficients of prime importance , the results from $\boldsymbol{b}^{\text{IV}}$ are also not in line with the original ones because 'pension wealth' – which is of prime importance – loses statistical significance. One may, however, argue that the confidence intervals of the incentive coefficients overlap for all three estimation procedures, implying that their results differ merely in economic terms.

To shed more light on what is different about these results, we examine predicted conditional retirement probabilities.[17] Figure 2 displays the sample distribution of the fitted values yielded by the three estimation methods in the respective estimation samples. 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: Whereas for $\boldsymbol{b}^{\text{WI}}$ and $\boldsymbol{b}^{\text{FD}}$ more than 60 percent of the predictions are outside the valid range, the corresponding share for

[16] *LM*-$\chi^2(1)$-statistic: 45.06, p-value: 0.0000; stata® implementation `underid` by Schaffer and Windmeijer (2020) used. In general weak instruments may, however, be an issue for $\boldsymbol{b}^{\text{IV}}$; cf. section 3. The underidentification tests are, for instance, far from rejecting the null if the estimator is applied to a richer, reduced-form model specification (Brown and Laschever, 2012, p. 109; table 4, column 8).

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

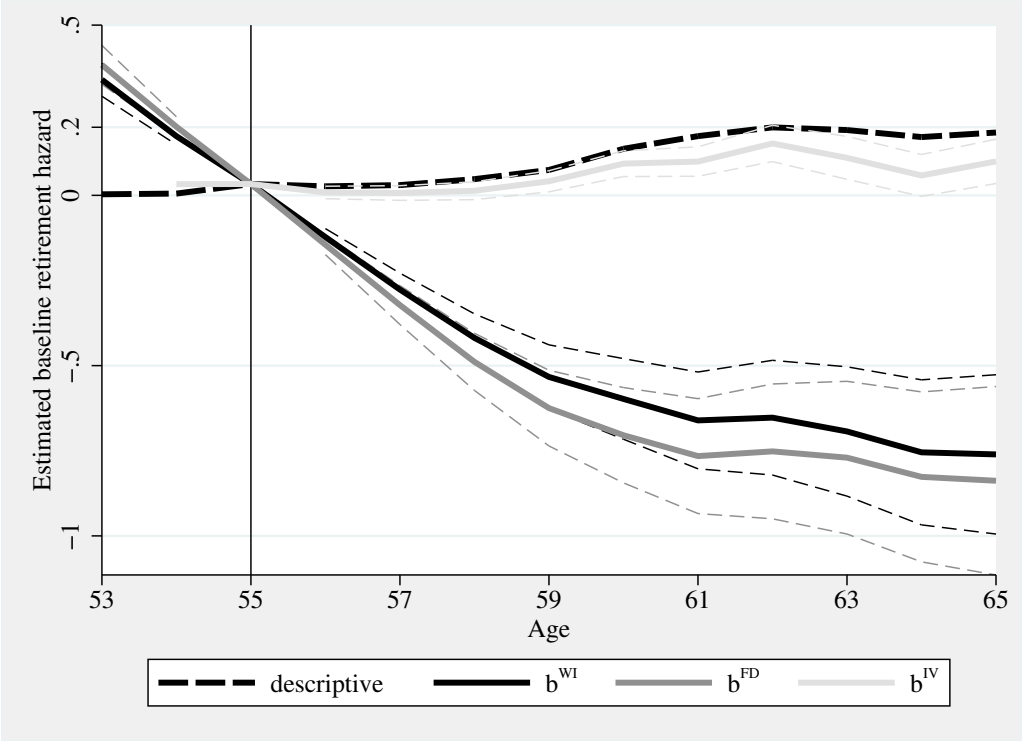

Figure 3: Estimated baseline-hazards; levels normalized to match the descriptive sample hazard ($^{83}/_{2491} = 0.033$) at the reference age (55 years); thin dashed lines mark 95 percent confidence intervals. **Source:** Authors' own calculations based on Brown and Laschever (2012/2019).

$\boldsymbol{b}^{\text{IV}}$ is smaller than 20 percent. Thus, the first two estimators do a very poor job in generating reasonable predictions. In fact, little mass of the distribution of fitted values is located in the meaningful range. On the basis of the predicted probabilities, one would judge $\boldsymbol{b}^{\text{IV}}$ to be clearly superior to $\boldsymbol{b}^{\text{FD}}$ and $\boldsymbol{b}^{\text{WI}}$ in the present application.

One possible explanation for the very different estimated distributions of retirement probabilities is the estimated age coefficients, which in absolute terms are typically much bigger for $\boldsymbol{b}^{\text{WI}}$ and $\boldsymbol{b}^{\text{FD}}$ than for $\boldsymbol{b}^{\text{IV}}$. Figure 3 depicts the baseline hazards[18] that are estimated by the age coefficients and compares them to their descriptive counterpart, i.e. to the age-specific relative retirement frequency. Descriptively, the retirement hazard is close to zero for teachers younger than 55 and then steadily increases until the age of 62, where it reaches roughly 20 percent. The baseline hazard estimated by $\boldsymbol{b}^{\text{IV}}$ roughly follows this pattern but exhibits a smaller age gradient. The latter finding makes sense because the empirical hazard – unlike the estimated baseline hazard for which financial incentives have been controlled for – not only captures the genuine age-specific inclination to retire, but also pension rules that financially disincentivize early retire-

[18]The level of the baseline hazard is individual-specific and only its shape is estimated by the age coefficients. In Figure 3 we normalize the level such that estimated baseline hazards coincide with their descriptive counterpart for the reference age-category of 55 years.

ment; see Brown and Laschever (2012, p. 94). In sharp contrast, $\boldsymbol{b}^{\text{WI}}$ and $\boldsymbol{b}^{\text{FD}}$ yield a steady and steep decrease in the baseline retirement hazard for teachers over virtually the entire considered age range, a decrease that is in no way mirrored by the unconditional sample retirement rates. Indeed, according to the results from the within estimator, the baseline retirement hazard decreases by 110 percentage points between the ages of 53 and 65, a result that makes little sense. A poorly estimated baseline hazard would appear 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 baseline hazard; see Table A3 in Appendix A.5.

6 Conclusions

Taking first-differences or applying the within-transformation estimator to eliminate individual time-invariant heterogeneity are powerful tools in applied econometrics that make the linear regression model very appealing when analyzing 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 observed only until that period in which the event of interest occurs. Indeed, as shown above, conventional fixed-effects estimators are biased and inconsistent in this case. In addition to conventional survival bias, which would also affect pooled OLS even if the individual heterogeneity were uncorrelated with the explanatory variables in the population, these estimators suffer from a second source of bias that originates from the data transformation itself and is therefore present even in the absence of any unobserved heterogeneity. Examining the classical linear fixed-effects estimators from an instrumental-variables perspective makes the nature of this bias more obvious. For the first-differences estimator, this bias is simply the failure to rescale the coefficient estimates of the reduced-form model. For the within-transformation, this bias originates from the fact that the endogeneity of the survival time invalidates group-mean deviations as instruments. This second source of bias turns out to be the dominant one in many settings, with its magnitude depending heavily on the data-generating process for 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 an alternative instrumental variables estimator that uses first-differences as instrument for the levels. It addresses the misscaling bias inherent to first-differences estimation by appropriately rescaling the estimated coefficients. 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 – and it does so under alternative, weaker assumptions than pooled OLS, for which uncorrelatedness with the levels of the explanatory variables is required. The contribution of this paper is thus twofold. First, it shows why conventional linear fixed-effects estimators should not be used in a discrete-time hazard framework. Second, it introduces an alternative estimator that confines possible bias to a single source. This remaining source is simply a variant of the conventional survival bias that researchers should always be aware of when estimating a linear discrete-time hazard model.

References

Aldrich, J. and Nelson, F. (1984). *Linear Probability, Logit and Probit Model*, Quantitative Applications in the Social Sciences, SAGE Publications, London.

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, London.

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

Amemiya, T. and MaCurdy, T. E. (1986). Instrumental-variable estimation of an error-components model, *Econometrica* **54**(4): 869–880.

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors, *Journal of Business & Economic Statistics* **19**(1): 2–28.

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, Princeton.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models, *Journal of Econometrics* **68**(1): 29–51.

Bloemen, H., Hochguertel, S. and Zweerink, J. (2017). The causal effect of retirement on mortality: Evidence from targeted incentives to retire early, *Health Economics* **26**(12): e204–e218.

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.

Brown, K. M. and Laschever, R. A. (2012/2019). Replication data for: When they're sixty-four: Peer effects and the timing of retirement, Nashville, TN: *American Economic Association* [publisher], 2012; Ann Arbor, MI: *Inter-university Consortium for Political and Social Research* [distributor], 2019-10-12. https://doi.org/10.3886/E113831V1.

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.

Do, Y. K. and Finkelstein, E. A. (2012). Youth employment, income, and smoking initiation: Results from Korean panel data, *Journal of Adolescent Health* **51**(3): 226–232.

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.

Finkelstein, A., Gentzkow, M. and Williams, H. L. (2019). Place-based drivers of mortality: Evidence from migration, *Working Paper 25975*, National Bureau of Economic Research.

Frazer, G. (2005). Which firms die? A look at manufacturing firm exit in Ghana, *Economic Development and Cultural Change* **53**(3): 585–617.

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.

Grunow, M. and Nuscheler, R. (2014). Public and private health insurance in Germany: The ignored risk selection problem, *Health Economics* **23**(6): 670–687.

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.

Hausman, J. A. and Taylor, W. E. (1981). Panel data and unobservable individual effects, *Econometrica* **49**(6): 1377–1398.

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.

Im, K. S., Ahn, S. C., Schmidt, P. and Wooldridge, J. M. (1999). Efficient estimation of panel data models with strictly exogenous explanatory variable, *Journal of Econometrics* **93**(1): 177–201.

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.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, Wiley.

Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition, *Journal of Econometrics* **133**(1): 97–126.

Lee, S. (2008). Estimating panel data duration models with censored data, *Econometric Theory* **24**(5): 1254–1276.

McGarry, K. (2004). Health and retirement: Do changes in health affect retirement expectations?, *The Journal of Human Resources* **39**(3): 624–648.

Nicoletti, C. and Rondinelli, C. (2010). The (mis)specification of discrete duration models with unobserved heterogeneity: A Monte Carlo study, *Journal of Econometrics* **159**(1): 1–13.

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.

Sanderson, E. and Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables, *Journal of Econometrics* **190**(2): 212–221.

Schaffer, M. E. and Windmeijer, F. (2020). UNDERID: Stata module producing postestimation tests of under- and over-identification after linear IV estimation, *Statistical Software Components*, Boston College Department of Economics.

Tutz, G. and Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*, Springer.

Wang, S. A., Greenwood, B. N. and Pavlou, P. A. (2020). Tempting fate: Social media posts, unfollowing, and long-term sales, *MIS Quarterly* **44**(4): 1521–1571.

Wettstein, G. (2020). Retirement lock and prescription drug insurance: Evidence from Medicare Part D, *American Economic Journal: Economic Policy* **12**(1): 389–417.

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 Conditional Mean of Disturbances of FD Estimation

The disturbance in the first-differences model is $\varepsilon_{it}^{\mathrm{FD}} \equiv y_{it} - \Delta \mathbf{x}_{it}\boldsymbol{\beta}$. For its conditional mean follows

$$\begin{aligned}
\mathrm{E}(\varepsilon_{it}^{\mathrm{FD}}|a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, t \leq T_i) &= \mathrm{E}(\varepsilon_{it}^{\mathrm{FD}}|a_i, \mathbf{x}_{it}, \mathbf{x}_{it-1}, t \leq T_i) \\
&= \mathrm{P}(y_{it} = 1|a_i, \mathbf{x}_{it}, t \leq T_i)\,(1 - \Delta \mathbf{x}_{it}\boldsymbol{\beta}) \\
&\quad + \mathrm{P}(y_{it} = 0|a_i, \mathbf{x}_{it}, t \leq T_i)\,(-\Delta \mathbf{x}_{it}\boldsymbol{\beta}) \\
&= (a_i + \mathbf{x}_{it}\boldsymbol{\beta})\,(1 - \Delta \mathbf{x}_{it}\boldsymbol{\beta}) + (1 - a_i - \mathbf{x}_{it}\boldsymbol{\beta})\,(-\Delta \mathbf{x}_{it}\boldsymbol{\beta}) \\
&= a_i + \mathbf{x}_{it-1}\boldsymbol{\beta} \qquad (14)
\end{aligned}$$

A.2 Probability Limit of the FD Estimator

We rewrite the first-difference estimator (5) as

$$\boldsymbol{b}^{\mathrm{FD}} = \boldsymbol{\beta} + \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \Delta \mathbf{x}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \varepsilon_{it}^{\mathrm{FD}} \right) \qquad (15)$$

Based on (6) 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

$$\operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \varepsilon_{it}^{\mathrm{FD}} \right) = \operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' a_i \right) + \operatorname{plim} \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \mathbf{x}_{it-1} \right) \boldsymbol{\beta} \qquad (16)$$

Using the identity $\left(\boldsymbol{I} + \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \Delta \mathbf{x}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \mathbf{x}_{it-1} \right) \right) \equiv \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \Delta \mathbf{x}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=2}^{T_i} \Delta \mathbf{x}_{it}{}' \mathbf{x}_{it} \right)$, which follows from $\mathbf{x}_{it-1} = \mathbf{x}_{it} - \Delta \mathbf{x}_{it}$, from (15) and (16) we get (7).

A.3 Estimation by the Within-Transformation Estimator

In this Appendix we examine the classical within-transformation estimator, denoted $\boldsymbol{b}^{\mathrm{WI}}$. To align the way of representing the estimator with section 3, we think of the within-estimator as applying the within-transformation only to the right-hand-side variables, i.e. $\ddot{\mathbf{x}}_{it} \equiv (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$, with

$\bar{\mathbf{x}}_i \equiv \frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{x}_{it}$, serve as explanatory variables. It is important to note that in terms of the estimated slope coefficients, this way of formulating the estimator is fully equivalent to applying the within-transformation also to y_{it}, which is presumably the most popular way of thinking about $\boldsymbol{b}^{\text{WI}}$.[19] In this model the disturbance term $\varepsilon_{it}^{\text{WI}}$ reads as $y_{it} - \ddot{\mathbf{x}}_{it}\boldsymbol{\beta}$ and for its conditional mean we obtain

$$
\begin{aligned}
\mathrm{E}\left(\varepsilon_{it}^{\text{WI}}|a_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},t\leq T_i\right) &= \mathrm{P}(y_{it}=1|a_i,\mathbf{x}_{it},t\leq T_i)\left(1-\left(\mathbf{x}_{it}-\frac{1}{t}\sum_{s=1}^{t}\mathbf{x}_{is}\right)\boldsymbol{\beta}\right)\\
&\quad+\sum_{r=t+1}^{T}\left[\mathrm{P}(y_{ir}=1|a_i,\mathbf{x}_{ir},r\leq T_i)\left(\prod_{s=t}^{r-1}\mathrm{P}(y_{is}=0|a_i,\mathbf{x}_{is},s\leq T_i)\right)\right.\\
&\qquad\left.\times\left(-\left(\mathbf{x}_{it}-\frac{1}{r}\sum_{s=1}^{r}\mathbf{x}_{is}\right)\boldsymbol{\beta}\right)\right]\\
&\quad+\left(\prod_{s=t}^{T}\mathrm{P}(y_{is}=0|a_i,\mathbf{x}_{is},s\leq T_i)\right)\left(-\left(\mathbf{x}_{it}-\frac{1}{T}\sum_{s=1}^{T}\mathbf{x}_{is}\right)\boldsymbol{\beta}\right)\\
&= (a_i+\mathbf{x}_{it}\boldsymbol{\beta})\left(1-\left(\mathbf{x}_{it}-\frac{1}{t}\sum_{s=1}^{t}\mathbf{x}_{is}\right)\boldsymbol{\beta}\right)\\
&\quad+\sum_{r=t+1}^{T}\left[(a_i+\mathbf{x}_{ir}\boldsymbol{\beta})\left(\prod_{s=t}^{r-1}(1-a_i-\mathbf{x}_{is}\boldsymbol{\beta})\right)\left(-\left(\mathbf{x}_{it}-\frac{1}{r}\sum_{s=1}^{r}\mathbf{x}_{is}\right)\boldsymbol{\beta}\right)\right]\\
&\quad+\left(\prod_{s=t}^{T}(1-a_i-\mathbf{x}_{is}\boldsymbol{\beta})\right)\left(-\left(\mathbf{x}_{it}-\frac{1}{T}\sum_{s=1}^{T}\mathbf{x}_{is}\right)\boldsymbol{\beta}\right)\\
&= (a_i+\mathbf{x}_{it}\boldsymbol{\beta})-\mathbf{x}_{it}\boldsymbol{\beta}+\sum_{r=t}^{T}\left(\mathrm{P}\left(T_i=r|a_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},T_i\geq t\right)\frac{1}{r}\sum_{s=1}^{r}\mathbf{x}_{is}\right)\boldsymbol{\beta}\\
&= a_i+\mathrm{E}(\bar{\mathbf{x}}_i)_t\boldsymbol{\beta} \qquad (17)
\end{aligned}
$$

with $\mathrm{E}(\bar{\mathbf{x}}_i)_t \equiv \mathrm{E}\left(\bar{\mathbf{x}}_i|a_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},t\leq T_i\right) = \sum_{r=t}^{T}\left(\mathrm{P}\left(T_i=r|a_i,\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},T_i\geq t\right)\frac{1}{r}\sum_{s=1}^{r}\mathbf{x}_{is}\right)$. That is $\mathrm{E}(\bar{\mathbf{x}}_i)_t$ denotes the expected value of $\bar{\mathbf{x}}_i$ conditional on unit i having survived at least until period t. For $t=T$, $\mathrm{E}(\bar{\mathbf{x}}_i)_t$ simplifies to $\frac{1}{T}\sum_{s=1}^{T}\mathbf{x}_{is}$. Equation (17) reveals that applying the within-transformation – just as taking first-difference – does neither remove the unobserved heterogeneity nor does it yield a disturbance that is conditional mean independent of the explanatory variables. The necessary conditions for unbiasedness are, thus, also violated for the classical within-estimator, even in the absence of unobserved time-invariant heterogeneity.

Unlike for the first-differences estimator, the explanatory variables enter the conditional mean of disturbance not in terms of observed lagged values but in terms of unknown conditional

[19] The equivalence become obvious by thinking of the within-transformation as 'partialling out' a saturated set of group indicators, which does not require transforming the left-hand-side variable (e.g. Wooldridge, 2009). This holds because, for any product of two data matrices, it makes no difference if either both or just one of them is transformed into group-mean deviations, since the 'residual marker' is symmetric and idempotent (Greene, 2014).

means. This is why the within-transformation estimator does not provide a basis for an instrumental variables estimator that mirrors $\boldsymbol{b}^{\mathrm{IV}}$. In this respect, it is telling that $\boldsymbol{b}^{\mathrm{WI}}$ is already an IV that uses $\ddot{\mathbf{x}}_{it}$ as instruments for $\mathbf{x}_{it}$, as shown in Arellano and Bover (1995) and with more rigor also in Im et al. (1999).[20] However, in a non-repeated event setting $\ddot{\mathbf{x}}_{it}$ is endogenous and ill-suited as instrument, since it is an immediate function of the ultimate outcome T_i. It is worth mentioning that Im et al. (1999) stress that the classical panel data setting – i.e. one without an absorbing state at the left-hand-side – allows for numerous instruments that can be used instead of $\ddot{\mathbf{x}}_{it}$, with $\Delta\mathbf{x}_{it}$ being among them.

[20] In applied econometrics the close link between fixed effects and instrumental variables estimation seems to have attracted little attention. Notable exceptions are Hausman and Taylor (1981) and Amemiya and MaCurdy (1986).

A.4 **Simulation Results for Probit Model as True DGP**

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

	b^{OLS}		b^{WI}		b^{FD}		b^{IV}	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x_{it}^{ST}: **true av. marg. effect** 1.0104 **(first wave incl.), and** 1.0046 **(first wave excl.)**								
$\hat{\beta}$	1.5118	0.0010	0.9110	0.0017	0.5113	0.0013	1.0143	0.0025
x_{it}^{RW}: **true av. marg. effect** 1.0018 **(first wave incl.), and** 0.9898 **(first wave excl.)**								
$\hat{\beta}$	1.2465	0.0007	0.9190	0.0013	0.9915	0.0013	0.9916	0.0012
x_{it}^{TR}: **true av. marg. effect** 1.0172 **(first wave incl.), and** 1.0252 **(first wave excl.)**								
$\hat{\beta}$	1.4778	0.0010	3.9868	0.0014	0.6949	0.0013	1.0389	0.0019

Notes: True DGP: $\mathrm{P}\left(y_{it}=1|a_i,x_{it},t\leq T_i\right)=\Phi\big(-1.44+3(a_i+\alpha+x_{it}\beta)\big)$; true coefficient value: $\beta=1$; $N=4\cdot 10^7$, $T=5$; same DGPs for x_{it} as in the simulations discussed in section 4; the # of observations for x_{it}^{ST} is 72 180 570, the corresponding # of observations for x_{it}^{RW} is 72 281 575, and for x_{it}^{TR} it is 72 722 785. For $\boldsymbol{b}^{\text{OLS}}$ the #s of observations are higher by $4\cdot 10^7$ observations, since no wave is 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 $\mathrm{P}\left(y_{it}=1|a_i,x_{it},t\leq T_i\right)=\Phi\big(-1.44+3(a_i+\alpha+x_{it}\beta)\big)$, with $\beta=1$ 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 4.1. The scaling factor 3 and the location parameter -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 4.1. 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 marginal effect $3\beta\frac{1}{M+N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\phi\big(-1.44+3(a_i+\alpha+x_{it}\beta)\big)$. For all considered DGPs, its value almost coincides with β, with and without the first wave being included. From comparing the coefficient estimate to the true average marginal effects 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, p. 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.

One may object that the above simulation considers a setting in which the linear and the probit model generate similar average marginal effects, making linear estimators mechanically perform well even if the true DGP is non-linear. To address this concern we consider an alternative DGP that generates marginal effects that more strongly deviate from the what the linear model yields. More specifically we consider $\mathrm{P}\left(y_{it}=1|a_i,x_{it},t\leq T_i\right)=\Phi\big(-1+\frac{3}{2}(a_i+\alpha+x_{it}\beta)\big)$, with the $Normal(0.5,4)$ distribution, instead of the $\mathcal{B}(0.2,0.2)$ distribution, entering the DGPs for x_{it}. Yet, this does not change the pattern of results in qualitative terms; see Table A2. In consequence, the

Table A2: Monte Carlo Analysis - **Probit as true DGP** (Large Sample Estimates; x_{it} normal)

	$\boldsymbol{b}^{\text{OLS}}$		$\boldsymbol{b}^{\text{WI}}$		$\boldsymbol{b}^{\text{FD}}$		$\boldsymbol{b}^{\text{IV}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
$x_{it}^{\boldsymbol{ST}}$: **true av. marg. effect** 0.5041 **(first wave incl.), and** 0.5034 **(first wave excl.)**								
$\hat{\beta}$	0.5247	0.0003	0.4554	0.0004	0.2553	0.0003	0.5043	0.0005
$x_{it}^{\boldsymbol{RW}}$: **true av. marg. effect** 0.4840 **(first wave incl.), and** 0.4679 **(first wave excl.)**								
$\hat{\beta}$	0.4544	0.0002	0.4153	0.0003	0.4682	0.0003	0.4682	0.0002
$x_{it}^{\boldsymbol{TR}}$: **true av. marg. effect** 0.5062 **(first wave incl.), and** 0.5059 **(first wave excl.)**								
$\hat{\beta}$	0.5284	0.0003	0.6836	0.0004	0.3381	0.0003	0.5053	0.0004

Notes: True DGP: $\mathrm{P}\left(y_{it}=1|a_i,x_{it},t\leq T_i\right)=\Phi\big(-1+\frac{3}{2}(a_i+\alpha+x_{it}\beta)\big)$; true coefficient value: $\beta=1$; $N=4\cdot 10^7$, $T=5$; deviating from section 4, $Normal(0.5,4)$ replaces $\mathcal{B}(0.2,0.2)$ in DGPs for x_{it}; the # of observations for x_{it}^{ST} is 71 990 728, the corresponding # of observations for x_{it}^{RW} is 72 528 280, and for x_{it}^{TR} it is 72 654 020. For $\boldsymbol{b}^{\text{OLS}}$ the #s of observations are higher by $4\cdot 10^7$ observations, since no wave is 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.

simulation results indicate that the advantage of $\boldsymbol{b}^{\text{IV}}$ over conventional fixed-effects estimators carries over to settings, in which the true DGP is not fully consistent with the linear hazard model.

A.5 **Simulation Results for Specification with Wave Indicators**

Table A3: Monte Carlo Analysis - Large Sample Estimates, **Wave Indicators** included

	$\boldsymbol{b}^{\text{OLS}}$		$\boldsymbol{b}^{\text{WI}}$		$\boldsymbol{b}^{\text{FD}}$		$\boldsymbol{b}^{\text{IV}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x_{it}^{ST} **stationary**								
$\hat{\beta}$	1.4837	0.0010	0.6411	0.0015	0.5043	0.0013	1.0045	0.0025
$\hat{\tau}_2$	−0.0011	0.0001	0.2945	0.0001				
$\hat{\tau}_3$	−0.0025	0.0001	0.4369	0.0001	−0.0049	0.0001	−0.0025	0.0001
$\hat{\tau}_4$	−0.0037	0.0001	0.5289	0.0001	−0.0143	0.0002	−0.0048	0.0001
$\hat{\tau}_5$	−0.0049	0.0002	0.5956	0.0002	−0.0283	0.0003	−0.0071	0.0002
$\hat{\alpha}$	0.0032	0.0002	−0.1031	0.0003	0.2947	0.0001	0.0968	0.0005
x_{it}^{RW} **follows random walk**								
$\hat{\beta}$	1.2550	0.0007	1.2017	0.0011	0.9990	0.0013	0.9992	0.0012
$\hat{\tau}_2$	−0.0017	0.0001	0.2953	0.0001				
$\hat{\tau}_3$	−0.0028	0.0001	0.4371	0.0001	−0.0073	0.0001	−0.0024	0.0001
$\hat{\tau}_4$	−0.0033	0.0001	0.5265	0.0001	−0.0242	0.0002	−0.0048	0.0001
$\hat{\tau}_5$	−0.0031	0.0002	0.5888	0.0002	−0.0529	0.0003	−0.0070	0.0002
$\hat{\alpha}$	0.0489	0.0002	−0.2127	0.0002	0.2953	0.0001	0.0978	0.0003
x_{it}^{TR} **with trend and increasing variance around trend**								
$\hat{\beta}$	1.5074	0.0011	0.8991	0.0016	0.6650	0.0013	1.0032	0.0019
$\hat{\tau}_2$	−0.0076	0.0001	0.2840	0.0001				
$\hat{\tau}_3$	−0.0151	0.0001	0.4279	0.0001	0.0075	0.0001	−0.0024	0.0001
$\hat{\tau}_4$	−0.0225	0.0001	0.5251	0.0002	0.0223	0.0002	−0.0047	0.0002
$\hat{\tau}_5$	−0.0300	0.0002	0.5989	0.0002	0.0442	0.0003	−0.0071	0.0002
$\hat{\alpha}$	0.0050	0.0002	−0.1484	0.0003	0.2869	0.0001	0.0970	0.0004

Notes: τ_t denote coefficients of wave indicators. True coefficient values: $\beta = 1$, $\alpha = 0.1$, $\tau_2 = \ldots = \tau_5 = 0$; $N = 4 \cdot 10^7$, $T = 5$; the # of observations for x_{it}^{ST} is 71 732 683, the corresponding # of observations for x_{it}^{RW} is 71 929 363, and for x_{it}^{TR} it is 72 211 807. For $\boldsymbol{b}^{\text{OLS}}$ the #s of observations are higher by $4 \cdot 10^7$ observations, since no wave is 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.5 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. The attached true coefficients, hence, capture how the baseline hazard evolves over time. 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 $\boldsymbol{b}^{\text{OLS}}$, $\boldsymbol{b}^{\text{FD}}$, and $\boldsymbol{b}^{\text{IV}}$, the within-transformation estimator $\boldsymbol{b}^{\text{WI}}$ 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 downward bias for a stationary x_{it} gets more pronounced. For x_{it} following a random walk, instead of suffering from a small downward bias, $\boldsymbol{b}^{\text{WI}}$ exhibits a sizable upward bias, if time indicators are included. Moreover, $\boldsymbol{b}^{\text{WI}}$ yields estimated time effects on the baseline hazard that are completely misleading. This mirrors the counterintuitive age effects $\boldsymbol{b}^{\text{WI}}$ yields in the real data application; see section 6. The simulation results are inline with our earlier argument about $\boldsymbol{b}^{\text{IV}}$ being biased 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

Table A4: Monte Carlo Analysis - Large Samp. Est., true **Time Effects** and **Wave Indicators**

	$\boldsymbol{b}^{\text{OLS}}$		$\boldsymbol{b}^{\text{WI}}$		$\boldsymbol{b}^{\text{FD}}$		$\boldsymbol{b}^{\text{IV}}$	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
x_{it}^{ST} **stationary**								
$\hat{\beta}$	1.4838	0.0011	0.6275	0.0017	0.5045	0.0014	1.0041	0.0028
$\hat{\tau}_2$	−0.0011	0.0001	0.2945	0.0001				
$\hat{\tau}_3$	0.1975	0.0001	0.6369	0.0001	0.1951	0.0001	0.1975	0.0001
$\hat{\tau}_4$	−0.0041	0.0002	0.5937	0.0002	0.1836	0.0002	−0.0057	0.0002
$\hat{\tau}_5$	0.1946	0.0002	0.8600	0.0002	0.3679	0.0004	0.1920	0.0002
$\hat{\alpha}$	0.0032	0.0002	−0.0924	0.0003	0.2947	0.0001	0.0969	0.0006
x_{it}^{RW} **follows random walk**								
$\hat{\beta}$	1.2770	0.0008	1.1867	0.0013	1.0012	0.0014	1.0013	0.0014
$\hat{\tau}_2$	−0.0016	0.0001	0.2953	0.0001				
$\hat{\tau}_3$	0.1975	0.0001	0.6373	0.0001	0.1929	0.0001	0.1978	0.0001
$\hat{\tau}_4$	−0.0032	0.0002	0.5905	0.0002	0.1721	0.0002	−0.0057	0.0002
$\hat{\tau}_5$	0.1976	0.0002	0.8523	0.0002	0.3402	0.0004	0.1925	0.0002
$\hat{\alpha}$	0.0445	0.0002	−0.2021	0.0003	0.2953	0.0001	0.0974	0.0003
x_{it}^{TR} **with trend and increasing variance around trend**								
$\hat{\beta}$	1.5380	0.0012	0.8951	0.0018	0.6745	0.0015	1.0039	0.0022
$\hat{\tau}_2$	−0.0079	0.0001	0.2840	0.0001				
$\hat{\tau}_3$	0.1843	0.0001	0.6280	0.0001	0.2075	0.0001	0.1976	0.0001
$\hat{\tau}_4$	−0.0238	0.0002	0.5900	0.0002	0.2201	0.0003	−0.0055	0.0002
$\hat{\tau}_5$	0.1684	0.0002	0.8634	0.0002	0.4403	0.0004	0.1920	0.0002
$\hat{\alpha}$	−0.0008	0.0002	−0.1394	0.0003	0.2868	0.0001	0.0969	0.0004

Notes: τ_t denote coefficients of wave indicators. True coefficient values: $\beta = 1$, $\alpha = 0.1$, $\tau_2 = 0$, $\tau_3 = 0.2$, $\tau_4 = 0$, $\tau_5 = 0.2$; $N = 4 \cdot 10^7$, $T = 5$; the # of observations for x_{it}^{ST} is 64 986 815, the corresponding # of observations for x_{it}^{RW} is 65 167 537, and for x_{it}^{TR} it is 65 445 856. For $\boldsymbol{b}^{\text{OLS}}$ the #s of observations are higher by $4 \cdot 10^7$ observations, since no wave is eliminated by the within-transformation or the first-differences transformation. See Table 1 for corresponding simulation results based on specification without wave indicators.

involve such time dependence. This is explained by the fact that the $\hat{\tau}_t$ capture the decrease of $\mathrm{E}(a_i|t, \mathbf{X})$ due to selective survival.

Table A4 shows simulation result for the same model specification used to generate the results displayed in Table A3. Yet unlike the latter, here the true DGP involves time effects, i.e. the true baseline hazard is not flat. More precisely the true baseline hazard is inflated by 0.2 in the periods three and five, that is $\tau_2 = 0$, $\tau_3 = 0.2$, $\tau_4 = 0$, and $\tau_5 = 0.2$. In qualitative terms, the results mirror what is found for a flat baseline hazard. As befor, $\boldsymbol{b}^{\text{IV}}$ does not estimate the baseline hazard unbiasedly. Yet, the error in the estimated baseline hazard turns out to be rather small. $\boldsymbol{b}^{\text{WI}}$ still yields poor results both in terms of the baseline hazard and in terms of the $\hat{\beta}$.